Czech Technical University in Prague
Faculty of Electrical Engineering
Department of Computer Science

NETWORK REPRESENTATION OF LATENT FEATURES EXTRACTED FROM TEXT DOCUMENTS

by

Ondřej Háva

A thesis submitted to
the Faculty of Electrical Engineering, Czech Technical University in Prague,
in partial fulfilment of the requirements for the degree of Doctor.

PhD programme: Electrical Engineering and Information Technology
Specialization: Information Science and Computer Engineering

April 2015

Thesis Supervisor:

Ing. Miroslav Skrbek, Ph.D.

Department of Digital Design

Faculty of Information Technology

Czech Technical University in Prague

Thákurova 7

166 29 Praha 6

Czech Republic

Thesis Supervisor Specialist:

Ing. Pavel Kordík, Ph.D.

Department of Theoretical Computer Science

Faculty of Information Technology

Czech Technical University in Prague

Thákurova 7

166 29 Praha 6

Czech Republic

Abstract

The accuracy of predictive models is strongly influenced by the quality of their input attributes. A model cannot exploit any information that is not encoded in its inputs. An important piece of information can be lost through the data preparation phase when the attributes are extracted from various sources of raw data and are added to a modeling data matrix.

A critical step of the selection of the input attributes is magnified in text mining tasks when the documents written in a natural language are classified, clustered or retrieved. A written text contains a huge amount of information that is scattered over different linguistic levels. The number of possible attributes that can be derived from each document is extremely high, which is caused by the richness of natural languages.

The basic morphological level offers usually tens of thousands of attributes in the form of different words or even word forms that have to be thoroughly selected or transformed into a manageable number of informative attributes. However, other useful information is hidden in the higher structure levels. Such information can be revealed from the context. Indeed, the contextual ties play a critical role in the text comprehension and it is worth attempting to extract same contextual attributes from text documents to improve their classification and other text mining tasks.

New structures called the context networks are proposed in the thesis. They encode contextual ties among words, terms, topics or other building blocks of text documents. The contextual ties are influenced by the neighborhood of the linguistic entities in a text. Inspired by standard n-gram language models, a fixed length context window is defined for each entity and all contextual ties from all context windows within a document are aggregated and encoded into the document context network.

The structure of the document context network can be rather complicated; the network itself is not an appropriate document representation, yet. We propose to reduce it using centralities of its nodes that represent selected linguistic entities. The method is preferably combined with an extraction of higher comprehensive features like topics to further reduce dimensionality of final vectors that represent the documents.

An exact formula that estimates the reduction of contextual diversity of original documents when they are encoded using the proposed representation is hard to express, hence the experimental results are provided in the thesis. Both simulations and real collection experiments confirm that the proposed representations successfully mix information about the document content and its context. However, it was experimentally proved that an encoding of within-document contextual ties does not generally improve the quality of standard text mining models with a few exceptions. These exceptions include tasks where documents cannot be distinguished by their context. Hence it is not generally worth investing computational resources to derive contextual attributes of documents; representations that rely on adjusted frequencies of linguistic entities offer a faster performance of text mining models with the comparable quality.

Keywords

document representation, dimensionality reduction, contextual attributes, context window, context network, network centralities

Acknowledgements

First of all, I would like to express my gratitude to my thesis supervisor, Miroslav Skrbek. He has been a constant source of encouragement and insight during my research. He, together with Pavel Kordík, provided me with numerous opportunities of professional advancement. Their continued support is gratefully acknowledged. Their efforts as thesis supervisors contributed substantially to the quality and completeness of the thesis.

Many other people influenced my work. I wish to thank to Miroslav Šnorek. He deserves the credit for many suggestions during the course of this work and my PhD study. His encouragement has always been a valuable source of support and I have learned a great deal from him.

I would like to express my special appreciation to the director of ACREA CR (former SPSS CR), Jan Řehák. He enabled me to combine my job with the studies and supported my effort to complete the thesis.

Finally, my greatest thanks belong to my wife Vladimíra and my children Klára and Martin. Their support was of great importance during my whole studies.

Contents

Foreword

People have developed the written text to the contemporary state for ages. Writing systems have been changed throughout historymany times and people of different cultures still use different symbol sets and vocabularies in their natural languages. The first texts appeared approximately 5000 years ago. Ancient manuscripts utilized picture-writing, cuneiform-writing of hieroglyphs. Current writing systems use three kinds of symbols: alphabets, syllabaries, or logographies. Any particular system can even have attributes of more than one category. In the alphabetic category, there is a standard set of letters that are classified as consonants or vowels. A syllabary is a set of written symbols that stand for syllables. Logographic writing systems use a single symbol for the whole word. For example, most Chinese characters are classified as logograms and many syllabaries are present in Japanese or Korean.

The diversity and complexity of natural languages is even reasonably higher than the diversity of writing symbols. The number of active spoken languages varies from 6000 to 7000[1] based on the definition of a language. They can be grouped into hundreds of language families. A human being learns to speak, read and write his/her mother tongue for many years and cannot understand to the overwhelming majority of foreign languages. Main causes of complexity of natural languages include the size of vocabularies, linguistic rule irregularities and ambiguity of words or even phrases. Hence the level of language awareness can be evaluated using different criteria such as the size of active vocabulary, the knowledge of morphology and syntax or the ability to recognize semantics. For example, The Dictionary of the Czech Language includes approximately 192 thousands of entries[2]. In addition, there are tens or hundreds of morphological paradigms in Czech based on the distinction detail. Finally, it is apparently impossible to count the true number of potentially correct phrases together with their meanings.

If we focus on a single natural language only, we are not able to efficiently and precisely encode the full range of linguistic rules to implement them into a computer program. If such linguistic system was implemented, it would help us to recognize the correctness of any written text, to uncover hidden syntactic structures or to establish semantic relations among entities. Unfortunately, we are able to develop simplified statistical models only that can for example assign the probability of correctness of a text or they can extract the most probable syntax parse of a sentence. Nevertheless, the tuning of any text processing statistical model is computationally expensive. From a statistical point of view the texts are samples of categorical data of huge dimensionality. Hence one needs an enormous volume of training text data to submit enough examples of correct words and phrases to the models. With respect to the size of vocabulary it is evident that no currently available collection of text documents is large enough to supply the reasonable number of examples. For example, if we consider a simple model that evaluates each word triplet (three adjacent words or 3-grams) and we restrict the vocabulary size to 50 thousand words, we can construct 125 trillion correct or incorrect 3-grams. If only 0.1% of them are correct, we need approximately 2.5 million of books of unique 3-grams to see them all just once. Regarding the high variability of the frequency of distinct 3-grams in natural texts, we would need even much more books. The whole National Library of the Czech Republic offers 6.2 million printed books[3][4]. If we

[1] http://www.linguisticsociety.org/content/how-many-languages-are-there-world
[2] http://lexiko.ujc.cas.cz/index.php?page=3
[3] The bibliographic questions and answers on http://www.ptejteseknihovny.cz/dotazy/pocet-knih-v-nk-cr-1
[4] In 2010 Google estimated that there were 1.3 billion books all over the world (http://www.cnews.cz/google-na-svete-je-presne-129-864-880-knih).

consider the electronic text only, the large text corpuses are built to statistically analyze natural languages, but their sizes are also too small to cover all possible contexts. For example, balanced version of the Czech National Corpus includes 100 million words[5] which is equivalent approximately to 2000 books. To make the situation even worse, we have to rely on the positive examples only, no corpus of incorrect texts is available. On the other hand, the volume of unstructured texts on the Internet is rapidly growing. Beside newscasts there is the huge number of blogs and messages on social networks that offer a massive source of text data.

Contemporary common computers are not able to process all available text data efficiently to store such a complex model. But the human brain with 100 billion neurons is able to learn well at least one natural language in several years without presenting such a large number of books. The questions arise: What are the basic building blocks of a natural language? What kind of rules and vocabularies should be stored to get a useful language model? Can be reading comprehension delegated to computers?

If we need a computer to efficiently manipulate text documents, we must extract well-defined structured attributes that properly describe the presented text. Apparently documents are not only containers for letters, syllables or words. The order of these basic components within a document is always important because it reflects ideas or topics hidden behind the explicit text and the relations among them. If we are able to extract such topics and relations, then the structured representation of documents should be similar to our perception of a text. A human being usually does not remember a presented text exactly, but he/she is able to exploit efficiently the information from the text. To enable computers to manipulate text documents we may utilize our linguistic knowledge about building blocks of natural languages. Or we can propose artificial compression of text streams into structured vectors and investigate the usefulness of these vectors for the tasks that we perform with text documents. Both approaches have been intensively studied for many years. Due to this research the basic problems in the field of language technology are acceptably solved. For example, they include the spam detection or the extraction of named entities from a text. Other problems that have not been solved yet make promising progress. The examples include the machine translation or the sentiment recognition. Nevertheless, many challenging and tough problems like dialog systems still wait for a satisfactory solution. The basic research in the field of document representation may help to find the reasonable and efficient solution for many problems associated with the text comprehension.

[5] The balanced corpus SYN2010 described on http://wiki.korpus.cz/doku.php/cnk:syn2010

1 Problem statement and goals

The goal of the thesis is to propose and test a new vector representation of text documents that takes into account the adjacency of any linguistic entities within a document.

The quality of a text mining or data mining solution can be increased if appropriate information is added to input attributes. Any modeling algorithm is not able to provide satisfactory predictions using improper inputs. In the text mining field co-occurrences or associations among linguistic entities play a critical role in comprehension of the text and unfortunately they are often neglected when documents are transformed to structured vectors that serve as model inputs.

In text mining applications, the input attributes are derived from free texts. They should describe a complex structure of the written documents. The hidden meanings of the words and phrases together with the order of ideas presented in a text are the essential properties of the documents and they contain valuable pieces of information that is worth encoding into a structured representation of documents.

The proposed procedure for feature extraction should be reasonably fast to enable an efficient extraction of features from large collections of documents. The features that constitute vectors of the low number of dimensions are preferable because they can serve better as the input data for convenient machine learning algorithms without the need for the further preprocessing. The input features should capture as much as possible of document diversity to ensure quality of the models. These two requirements (the small number of features and the capture of document diversity) are contradictory and the reasonable trade-off needs to be selected.

With the above mentioned restrictions in mind the goals of the thesis can summarized as:

- Propose a document representation that enables to improve quality of standard text mining solutions. A low-dimensional vector representation of text documents that encode both their contents and contexts is preferable. The dimensions may represent any linguistic entities extracted from the text together with their relations. The proposed representation should include the reduction of the natural variability of the text where the same topics can be expressed in several different ways.
- Develop the procedure for extraction of the fixed number of features that does not rely on huge linguistic resources and is not dependent on a language. This requirement ensures the faster processing of documents and it guarantees that the extraction procedure is applicable for collections written in different natural languages. The extraction method that does not produce missing values is preferable.
- Prove theoretically that the proposed representation encodes well the order of selected linguistic entities. Compare the reduction of diversity of documents within a collection using the proposed representation and a commonly used representation. Compare also the reduction of diversity for different variants of the proposed representation.
- Test the appropriateness of the proposed representation for common text mining tasks on simulated documents. The simulated documents fulfill all the assumptions of the proposed procedures, hence the direct impact to the diversity reduction and to the model performance can be observed. Compare the results with the ones based on the common document representation.

- Test the appropriateness of the proposed representation for common text mining tasks on different collections of real documents. The real documents may violate some assumptions or to include some important relations that are not taken into account, hence the proposed representation may interact uniquely with the text mining models. Compare again the proposed representation with the common representation.

2 Contribution of thesis

A new approach to extraction input attributes for text mining models in order to improve their performance is presented in the thesis; the thesis introduces a new vector representation of text documents. On the contrary to other document representations, the proposed representation comprises also the information about the order of selected linguistic entities within a document. The proposed representation is applicable to any common linguistic entities, hence the entity identification within a document is a mandatory step performed in advance. Entity frequency scores are enhanced by contextual scores in the proposed procedure and several different methods of combining the scores are tested. The proposed representation can be based on low-level entities such as words or stems or on complex entities such as terms or concepts. Even abstract or latent entities such as topics can be used if their sequences are identifiable within documents.

The main contribution of the thesis is the analysis of the usefulness of the proposed contextual enhancement of document vectors in the common text mining tasks. The reduction of document separability that is influenced by the vector representation of unstructured texts is studied theoretically and also in experiments. The thesis describes why the proposed representations capture better the document contextual diversity than the standard approaches. It was experimentally shown that the proposed enhancements of document vectors seldom improve document retrieval, classification or clustering. We suppose that while the contextual information is critical for specialized language processing procedures such as the machine translation, the standard text mining tasks such as the document classification are principally sensitive to the document content. Hence we recommend using simple standard frequency scores of extracted linguistic entities to represent a document by a numeric vector for text mining models; the enhancements that reflect entity adjacencies can be skipped to make the document processing faster.

2.1 Organization of thesis

The thesis is organized as follows: The summary of the approaches to representation of text documents can be found in chapter 3. This summary also focuses on the methods of dimensionality reduction that are used in text mining and can be applied in the process of obtaining the proposed representation. Chapter 4 describes the proposed contextual vector representation of text documents. The procedure is illustrated by simple examples. The theoretical aspects of the new document representation are analyzed in chapter 5; we try to assess how the document diversity is reduced by the different representations using the assumptions about the document genesis borrowed from n-gram language models.

Chapter 6 offers description of experiments. The experiments are performed on simulated documents and on different collections of real documents as well. The same chapter also summarizes the results of these experiments. The theoretical assessments together with the practical experiments imply the final conclusions about the usefulness of the proposed representation. The conclusions are presented in chapter 7.

3 Related work

Text mining is an emerging area of computer science which exhibits strong relations with natural language processing, data mining, machine learning and knowledge management. Text mining seeks to extract useful information from unstructured textual data through the identification and exploration of interesting patterns (Feldman & Sanger, 2007).

The history of text mining started deeply in the last century. Document indexing was extensively studied already in the 1950s and the 1960s (Luhn, 1958). In the late 1950s the first automatic text retrieval system was suggested (Luhn, 1957). It was based on a comparison of content identifiers attached both to stored texts and to the users' information queries. Measurement of document similarity and document clustering are also rather old (Jardine & van Rijsbergen, 1971). Probably the most frequent text mining task classification was solved already in the late 1980s when machine learning algorithms started to be widely used (Hayes & Weinstein, 1990). A classification applied to a text is sometimes referred as text categorization (TC) (Sebastiani & Delle Ricerche, 2002). In TC a set of documents is automatically sorted into predefined categories. TC is employed in text filtering, categorization of web pages or in sentiment analysis. While TC is supervised task, document clustering (DC) deals with discovery of groups of documents that minimize inner-cluster similarity and maximize inter-cluster similarity. DC has been intensively studied since the 1990s (Anick & Vaithyanathan, 1997) and it is still the emerging text mining area (Aggarwal & Zhai, 2012).

There are many applications of basic text mining procedures in order to solve more specific tasks over unstructured texts. They include spam detection, market intelligence, detection of plagiarism or enhancement of search engines. The examples of currently emerging tasks are text summarization (Mani, 2001) or sentiment analysis (Dey & Haque, 2008).

Figure 1: The general steps of text processing.

Regardless of the text mining task the unstructured text from documents must be somehow transformed to a structured representation. Preferably each document is represented by a vector of constant length; the document collection then constitues a matrix of row vectors. In the basic bag-of-words (BOW) representation (Salton et al., 1975) a document is modeled as a container of vocabulary tokens regardless of the order of tokens in a document. Vocabulary tokens serve as features and frequencies of tokens are used as weights. In the simplest case the weights are 0/1 indicators and denote the presence or absence of particular tokens in a document. The more sophisticated approaches for weights derivation were tested in the 1970s and the 1980s (Salton & Buckley, 1988) and they are still widely used. The proposed weighting schemas utilize the two-component multiplicative approach. The first component reflects the token frequency in a particular document while the second component adjusts the token importance by its global frequency in the whole document collection. From many proposed weighting schemas the term frequency / inverse document frequency (TF-IDF) approach is the most often used one.

Apart from BOW approaches there are different methods that enable to construct feature vectors from the documents that are considered as strings of characters or streams of words.

The main difference between the stream representation and BOW representation is that the former retains ordering information. The simple, but commonly used representations are referred as n-gram representations. If a document is considered as a sequence of characters, n-grams are subsequences of these characters of the length n (Cavnar & Trenkle, 1994). Such a representation is useful for example for multi-language collections where universal vocabulary is not available. Similarly to character n-grams, if a document is viewed as a sequence of vocabulary words, n-grams are then subsequences of n words. The word n-grams are essential for language models (Chomsky, 1956) where the theory of Markov chains (Markov, 1913) is exploited to estimate the probability of a token conditionally on its context. n-gram approaches take the order of text units into account, so they make the additional information hidden in a text available for mining tasks. On the other hand, the context encoding further increases the dimensionality of extracted vectors. Therefore the dimensionality reduction techniques must be taken into account. n-gram language models can be substituted by different approaches that take the context of words into account as well (Schwenk, 2007) (Mikolov et al., 2011). For example, neural network language models often outperform n-gram models (Bengio et al., 2003), but their usefulness for an efficient document vector representation is disputable.

The need for improvement of text documents representation caused that the computational linguistics started to be important in the data preparation phase of text mining projects (Feldman & Sanger, 2007). Some essential procedures are usually borrowed from the natural language processing (NLP). They include tokenization (Grefenstette & Tapanainen, 1994), stemming (Porter, 1980) (Xu & Croft, 1998), lemmatization (Liu et al., 2012), part-of-speech tagging (Brill, 1992) (Ratnaparkhi, 1996), word sense disambiguation (Ide & Véronis, 1198) or even shallow parsing (Earley, 1970) (Tomita, 1986) (Charniak, 1997). The software kids that offer transformation of text documents to feature vectors often enable to develop the pipeline of these NLP procedures to comfortably tag, index and structurally represent the text documents. Instead of tokenization the above mentioned procedures are not necessary to propose any BOW representation, but they are useful for the partial dimensionality reduction. Since they exploit a natural approach to the linguistic text analysis, they are language dependent and often resource intensive.

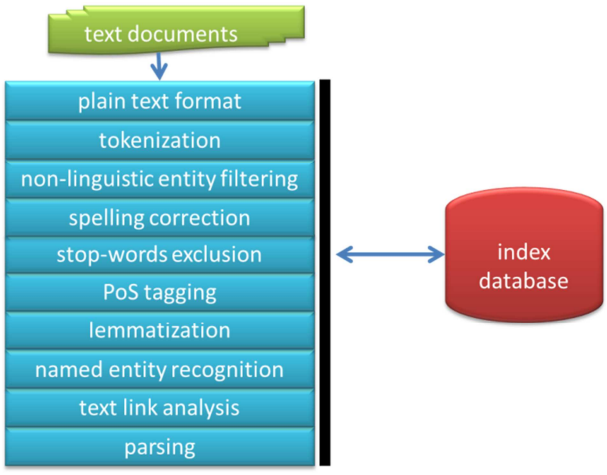


Figure 2: An example of NLP pipeline.

Even though the tagging and the indexing of documents achieved by NLP procedures results in an informative document representation in the semi-structured format, it is still a too rich representation to serve as an input data for further machine learning algorithms. And because NLP procedures often rely on extensive language dependent resources, text miners frequently use more universal approaches for the reduction of the dimensionality of the basic BOW

representation that reveal and utilize hidden relationships among the words. The methods provide either a small number of extracted features or they filter out some input features (Lewis, 1992). The comparison of feature selection approaches for the purpose of the document categorization is made in (Yang & Pedersen, 1997). The feature selection methods can be either unsupervised, such as the document frequency thresholding, or supervised. The supervised methods rely on a statistical evidence of an association among tokens and target categories such as the chi-square statistics or the mutual information.

The feature extraction can be performed by the clustering of extracted tokens. The weights of the new extracted features are computed as sums of original token weights. In (Verbeek, 2000) the authors search for appropriate words to create the supervised clusters that provide a reasonable predictive power for consequent classifiers. They also propose the estimate of the optimal number of clusters.

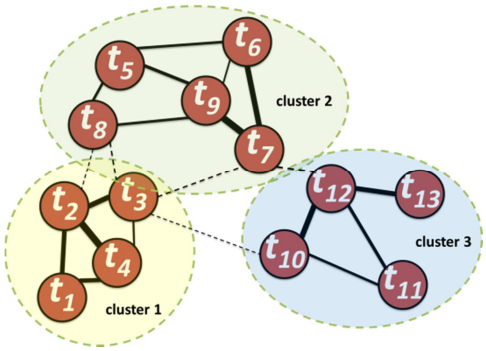


Figure 3: An example of unsupervised token clustering in a token network. The thickness of connections is proportional to an association between token pairs. The association can be for example derived from token co-occurrences within documents.

The widely used feature extraction methods utilize the linking of original features. The new feature weights are produced by different combinations of the original ones; the extraction process can be often described as a projection of a document representation to a new low dimensional space. The common example of the feature extraction technique is Latent Semantic Analysis (LSA) (Deerwester et al., 1990). LSA uses a linear projection to a low-dimensional space of latent features that maintains as much variability of the original features as possible. The new dimensions are determined by the singular value decomposition (SVD) of the original document-tem matrix (Golub & Van Loan, 1996). The additional enhancement of LSA was introduced in the probabilistic Latent Semantic Indexing (pLSI) model (Hofmann, 1999), also known as the Aspect Model. Even though pLSI exploits the same SVD approach as LSA, the theory behind it is based on a generative model where each token in a document is regarded as a result of a sampling from a mixture model. The mixture components are multinomial random variables that can be viewed as representations of latent topics. If one wish to consider an exchangeable representation for documents and tokens a more complex mixture models should be considered. This line of thinking leads to the model of Latent Dirichlet Allocation (LDA) (Blei et al., 2003). The exchangeability in LDA implies the model with a conditionally independent and identically distributed mixture of topics with respect to an underlining latent parameter of a probability distribution. LDA is then the complex generative model which describes the genesis of observed documents and enables to assign latent topics to individual tokens in a document. Instead of using the matrix algebra the parameters of the LDA process are estimated by the expectation–maximization algorithm (ME) (Dempster et al., 1977). Note that all here mentioned standard methods of the feature extraction are based on the bag-of-words assumption which means that the order of tokens in

a document can be neglected. In the probability theory this assumption is referred as the exchangeability of words in a document (Aldous, 1983).

The alternative way to represent documents by topics that are hidden in a text is Explicit Semantic Analysis (ESA). ESA exploits some standard vocabulary of the topics that are known and described in advance. In (Gabrilovich & Markovitch, 2007) the authors utilize Wikipedia. They compare a document with Wikipedia articles. The Wikipedia articles serve as features and the similarity scores as weights.

All features extracted from unstructured texts can be enhanced by features extracted from a semi-structured representation like XML or some other structured document data or metadata such as the document source, the date or the author. For example DBpedia data set (Bizer et al., 2009) that enables to semantically query the content of Wikipedia was extracted both from the structured and the unstructured parts of Wikipedia articles. Generally features extracted from plain documents in the data preparation phase can be merged with any hard coded database structured data to provide a complex view on explored units such as patients, customers or products (SPSS Inc., 2008). Text mining and data mining share many machine learning methods and they are considered as related domains.

3.1 Dimensionality reduction in text mining

The quality and efficiency of any data mining task such as classification, clustering or regression is dependent on the information hidden in the features that are used as predictors. On the contrary the noisiness of input features can reduce the model quality. For example commonly used tokens such as "the" may not be very useful in improving the quality of text mining classifier. Therefore it is critical to select an appropriate set of features so that the noisy ones are removed and the informative ones are retained before the model is built.

Feature selection methods can be organized into three categories depending on how they combine the feature selection search with the construction of a text mining model: filter techniques, wrappers and embedded methods (Saayes et al., 2007). The filter techniques evaluate the relevance of a feature by looking only at intrinsic properties of input data; they do not interact with the model. The wrapper methods search for a relevant subset of features using evaluation measures of the subsequent model. The embedded methods are integral parts of models; they are closely related to the modeling algorithm. Most of the feature selection methods can perform the feature ranking when each input feature receives its rank or score based on its individual predictive power. The examples of the feature selection methods include the document frequency selection, the entropy-based ranking or the term contribution.

In addition to the feature selection and feature ranking methods, the number of usually standalone feature extraction approaches is available. The standard feature extraction methods that are used in text mining to improve the quality of a document representation or to compress a sparse document vectors include Latent Semantic Indexing (LSI), Probabilistic Latent Semantic Indexing (pLSI), Latent Dirichlet Allocation (LDA) or Non-negative Matrix Factorization (NMF). In these techniques correlations among tokens that occur in the same documents are exploited in order to construct some new features that correspond to hidden topics or principal components in a document collection.

3.1.1 Feature selection methods

Feature selection and ranking methods are common and easy to apply in supervised problems such as the document classification (Yang & Pedersen, 1997) where target document categories are available for training documents. Generally, the ranking can be performed by measuring a correlation, building single variable classifiers or by exploiting information

theoretic criteria (Guyon & Elisseeff, 2003). Beside common measures that evaluate the strength of a relation between a particular input and a target variable like chi-square or correlation, the special methods used in text mining include BM25 (Robertson & Zaragoza, 2009), Relevance Propagation (Qin et al., 2005) or PageRank. However, a number of simple and efficient unsupervised methods can be used in text mining as well. They often exploit the document similarity measures (Grefenstette & Pulman, 2010) that are applied in document clustering (Sruthi & Reddy, 2013).

group	distance examples
counts/indicator vectors	character counts, word counts, cosine similarity, dice similarity, Euclidian distance, city bock distance, Mahalanobis Distance, Pearson correlation, MASI distance, Jaccard similarity, Sørensen-Dice coefficient, Tversky index, Tanimoto distance, overlap coefficient
stringology	Hamming distance, Levenshtein distance, Damerau-Levenshtein distance, Jaro-Winkler distance, ratio similarity, Lee distance
information theory, probability	Kullback-Leibler divergence, Kendall tau distance, cross-entropy, mutual information
machine translation	BLEU,NIST, WER, ROUGE, METEOR,
ontology-based	path similarity, Wu-Palmer similarity, Lin similarity, Leacock-Chodorow similarity, Mao similarity, Resnik similarity, Jiang similarity, Knappe similarity,

Table 1: Examples of document similarity measures

3.1.1.1 Document frequency selection

Probably the simplest and also often used method for the feature selection is the exploitation of the document frequency to filter out the useless features. The filtering of very frequent words reduces their noise effect. The tokens which are too frequent in the collection should be removed because they are typically the commonly used words such as "the" or "of" in English. These non-discriminative tokens are referred as stop words (Rijsbergen, 1975). Stop word lists are usually available for common natural languages. They can be directly applied to tokenized documents to remove the listed words. The typical stop word list includes several hundreds items. Note that popular TF-IDF weighting method (Salton & Buckley, 1988) can also partially filter out very frequent words in a soft way, but the standard list of stop words provide a universal set of words to prune independently on the collection.

In addition, the words that occur extremely infrequently should be removed from documents as well. They do not exhibit any significant relational pattern that can contribute to the model building. Such words often include misspellings or typographical errors. Especially the document collections downloaded from blogs or social networks likely contain these words with the mistakes.

Similarly non-linguistic entities do not occur frequently in a collection. The non-linguistic entities include identifiers such as URLs, phone numbers, e-mails, sometimes also dates or numbers. They are recognized by special algorithms after the tokenization. As the other infrequent tokens they do not create useful patterns, but if the specific tokens are transformed to vaguer ones, they can become useful features. For example, exact phone numbers are transformed to common tokens referred as "phone_number".

3.1.1.2 Term strength

The term strength (Wilbur & Sirotkin, 1992) measures how a word is informative for identifying a relation between a pair of documents. Firstly, we have to define when two documents are related. It is easy in the supervised situations in which the predefined target categories of documents are available. Because it is not practical to create manually document categories in large unsupervised collections, it is desirable to define the purely unsupervised

concept where two documents are related. It is possible to use the cosine similarity (Salton, 1983) to measure the relatedness of a document pair.

Two documents are related if their cosine similarity is above the threshold. Then the strength *z(w)* of a term *w* is usually being defined over a random sample of the related documents as the ratio of the number of pairs in which *w* occurs in both documents divided by the number of pairs in which *w* occurs in the first document of the pair. The first document of a pair can be picked randomly.

In order to filter out the unimportant terms, the term strength may be compared with the expected strength. If the term strength is not at least two standard deviations greater than the average term strength, then the term is removed from the documents.

This approach does not require any initial target categories, but it can be directly used for the feature selection in the supervised classification as well (Yang, 1995). It is particularly suited for similarity based methods such as the clustering because the discriminative nature of the features is defined on the basis of the similarities among the documents and the similar documents belong to the same category.

3.1.1.3 Entropy-based ranking

In the entropy-based ranking approach (Dash & Liu, 1997) the quality of a term is measured by the reduction of the entropy when the term is removed from the collection. The entropy of a term *w* in the collection of *M* documents is defined as

$$E(w) = -\sum_{i=1}^{M}\sum_{j=1}^{M}\left[s_{ij}\log(s_{ij}) + (1-s_{ij})\log(1-s_{ij})\right]. \quad (1)$$

The similarity s_{ij} between documents d_i and d_j when the term *w* is filtered out is computed as

$$s_{ij} = 2^{-\frac{r_{ij}}{\bar{r}}}. \quad (2)$$

Here r_{ij} stands for the distance between documents d_i and d_j when the term *w* is filtered out, $\bar{r}$ is the average distance between the pairs of documents after the removal of the term *w*. The definition of s_{ij} implies that $s_{ij} \in \langle 0;1 \rangle$. A pair of documents with the average distance has the similarity of one half. The resultant entropy *E(w)* then describes the variability of the similarity of documents after the term *w* is removed. The terms with the low entropy are filtered out form the collection.

Note that the computation of the term entropy *E(w)* is computationally intensive. The derivation of the entropy itself requires $o(M^2)$ operations plus we must add the distance computation requirements. Hence the entropy-based ranking is impractical for large collections and the sampling methods must be considered (Dash & Liu, 1997).

3.1.1.4 Term contribution

The term contribution selection method (Liu et al., 2003) is based on the fact that models often rely on document similarity. The typical example is the document clustering where the similar documents are grouped together. Therefore the contribution of a term can be viewed as its contribution to the document similarity. In the case of the commonly used cosine similarity the similarity between two documents is computed as the dot product of their

normalized frequencies. Then the contribution of a term to the similarity of two documents is the product of their normalized frequencies in these documents. To determine the contribution of a term the products need to be summed over all document pairs in a collection.

Note that only the extraction of all document pairs in a collection of M documents requires $o(M^2)$ operations. Hence sampling methods must be considered for larger collections. The second disadvantage of the term contribution selection method is the fact that it favors highly frequent terms without a regard to the specific discriminative power of a term.

3.1.1.5 Concept decomposition using clustering

While the dimensionality reduction is often used as preprocessing step for the document clustering, the clustering itself can be used as a feature selection approach known as the concept decomposition. The concept decomposition (Dhillon & Dharmendra, 2001) exploits any clustering algorithm applied on the original representation of documents. The frequent terms in the centroids of the resultant clusters are selected for the reduced document representation.

This condensed conceptual representation allows for the second step in the clustering task as well as for other tasks such as the classification. The computational requirements of the concept decomposition depend mainly on the selected clustering technique.

3.1.2 Feature extraction methods

3.1.2.1 Latent sematic indexing

The Latent Sematic Indexing (LSI) (Deerwester et al., 1990) analysis involves the Singular Value Decomposition (SVD) (Golub & Van Loan, 1996), a technique closely related to the Eigenvector Decomposition and the Factor Analysis (Forsythe et al., 1977). LSI as a feature extraction method attempts to overcome problems with the variability in the word usage by automatically organizing tokens into a sematic structure more appropriate for the information retrieval and other text mining tasks. LSI assumes that tokens contained in a document are incomplete and unreliable indicators of the document content. There is an underlying or latent structure in patterns of the token usage hidden behind the explicit document that is partially obscured by the variability of the word choice. The statistical approach used to reveal this latent structure gets rid of the obscuring noise and enables to represent documents in a new low dimensional feature space.

In LSI large and a sparse document-term matrix is decomposed into a set of orthogonal latent factors. Then only the most important ones are selected for a new document representation. The importance is usually measured by the variability of the original features explained by the factor. More formally the rectangular document-term matrix $\mathbf{D}$ of the size $M \times N$ is decomposed into the product of three new matrices as

$$\mathbf{D} = \mathbf{P}\boldsymbol{\Lambda}\mathbf{Q}^T . \tag{3}$$

The matrices $\mathbf{P}$ and $\mathbf{Q}$ of the sizes $M \times L$ and $N \times L$ respectively have orthogonal columns; $\boldsymbol{\Lambda}$ is a square diagonal matrix.

$$\begin{aligned} &\mathbf{P}^T\mathbf{P} = \mathbf{I} \\ &\mathbf{Q}^T\mathbf{Q} = \mathbf{I} \\ &\lambda = diag(\lambda_1, \lambda_2, \ldots, \lambda_L) \end{aligned} \tag{4}$$

This is the Singular Value Decomposition (SVD) of matrix **D**, L=rank(**D**) and λ_i are non-negative singular values. The column vectors $\mathbf{p}_i$ and $\mathbf{q}_i$, i=1…L, of the matrices **P** and **Q** are also referred as the left and right singular vectors satisfying equations

$$\begin{aligned} \mathbf{D}\mathbf{D}^T\mathbf{P} &= \mathbf{P}\boldsymbol{\Lambda}^2 \\ \mathbf{D}^T\mathbf{D}\mathbf{Q} &= \mathbf{Q}\boldsymbol{\Lambda}^2 \end{aligned}. \tag{5}$$

Note that λ_i^2 are the eigenvalues of both the document covariance matrix $\mathbf{D}\mathbf{D}^T$ and the term covariance matrix $\mathbf{D}^T\mathbf{D}$. Let the singular values be sorted descending $\lambda_{i+1} \geq \lambda_i$. The Singular Value Decomposition enables a dyadic decomposition of the document-term matrix **D** as

$$\mathbf{D} = \sum_{i=1}^{L} \mathbf{p}_i \lambda_i \mathbf{q}_i^T \ . \tag{6}$$

If only K largest singular values are kept along with their corresponding columns of **P** and **Q,** the approximation of the original document-term matrix **D** can be defined as

$$\hat{\mathbf{D}} = \sum_{i=1}^{K} \mathbf{p}_i \lambda_i \mathbf{q}_i^T \ . \tag{7}$$

Eckart–Young theorem (Eckart & Young, 1936) confirms that it is the best approximation of matrix **D** by a matrix of rank K and it holds true

$$\left\| \mathbf{D} - \hat{\mathbf{D}} \right\|^2 = \sum_{i=K+1}^{K} \lambda_i^2 \ . \tag{8}$$

The matrix norm is Frobenius norm and the theorem tells that the proposed approximation is the closest one in the least squares sense. The pragmatic idea for this approximation is that the first K independent components capture the major associational structure of **D** and throws out the noise.

The Singular Value Decomposition of the document-term matrix **D** is usually performed with a training set of documents. However, this method can be deployed for a new or a test document as well using the derived set of orthogonal factors. If a new document is originally represented by a row vector **d**, its approximation in the new K-dimensional space is

$$\hat{\mathbf{d}} = \mathbf{d}\mathbf{Q}\boldsymbol{\Lambda}^{-1}. \tag{9}$$

The equation also describes the resultant linear projection from the original L-dimensional space to the new K-dimensional space which is achieved by the multiplication by the projection matrix $\mathbf{Q}\boldsymbol{\Lambda}^{-1}$. Similarly to the transformation of documents to the latent low-dimensional space LSI enables to project the original features to the same latent space using the projection matrix $\mathbf{P}\boldsymbol{\Lambda}^{-1}$. Utilizing this dual projection we can conclude that the terms which occur in similar documents will be near each other in the new low-dimensional latent

space even if they never co-occur in the same document. So the LSI representation captures also term to term associations that are important for the information retrieval.

Due to the possible projections of both the documents and the terms to the new latent space LSI enables important comparisons between the objects. We can explore distances between the pairs of documents, the pairs of terms and also between the document-term pairs. The comparisons among the terms offer also another interpretation of LSI. The new latent dimension can be viewed as a topic expressed in documents. These topics are generated by the observed the terms which correspond to their higher semantic meaning. Due to the revealed term associations LSI solves the problem with the polysemy and the synonymy that is present in the written text.

LSI approach was successfully applied in many text mining tasks such as the information retrieval (Deerwester et al., 1990), the text summarization (Yihong & Xin, 2001) or the classification (Háva et al., 2012). The main advantages of LSI are the language independence, the easy implementation and the possible interpretation of the latent space.

3.1.2.2 Non-negative matrix factorization

The Non-negative Matrix Factorization (NMF) (Xu et al., 2003) belongs to the methods which reveal a latent space that is suitable for a document representation. Similarly to the LSI, the NMF represents documents in a new system of dimensions that are extracted from the document-term matrix of training documents. While LSI offers a new system of orthogonal axes, this is not the case for NMF.

NFM is a feature extraction method which is well suited for the clustering. The vectors in the basis system of NFM correspond to cluster topics. Therefore the cluster membership of a document may be determined directly from the new reduced representation by examining the largest component of the document.

The new coordinates of any document are always non-negative. They are derived as an additive combination of the underlining semantic features. Hence the representation of a particular document makes a sense from an intuitive perspective.

Let **D** be the document-term matrix of the size $M \times N$. We wish to create new K dimensions from the underlining document collection. The NFM method attempts to determine two matrices **U** and **V** that minimize the objective function

$$J = \tfrac{1}{2}\left\|\mathbf{D} - \mathbf{U}\mathbf{V}^T\right\|^2. \qquad (10)$$

The norm of the matrix is the sum of all squared elements of the matrix (Frobenius norm). **U** and **V** are non-negative matrices of the sizes $M \times K$ and $N \times K$ respectively. The columns of $\mathbf{V}^T$ provide K basis vectors that correspond to K hidden topics.

By minimizing the objective function J we attempt to approximate the matrix **D** by the product $\mathbf{U}\mathbf{V}^T$. Hence a document vector **d** which is a row of **D** is approximated as $\mathbf{u}\mathbf{V}^T$ where **u** is the corresponding row of **U**. Therefore the document vector **d** can be rewritten as the approximate non-negative linear combination of the K columns of $\mathbf{V}^T$. The rows of **V** (or the columns of $\mathbf{V}^T$) correspond to K basis vectors derived using NMF form the original representation of the document collection.

If the value of K is relatively small compared to the dimensionality of the original collection, the rows of **V** discover the latent structure of the data. Furthermore, the non-negativity of

matrices **U** and **V** ensures that the documents are expressed as the non-negative combination of the hidden topics which enables the straightforward interpretation of the results.

Let us solve the optimization problem for the objective function *J*. Frobenius norm of any matrix **Q** can be expressed as

$$\|\mathbf{Q}\|^2 = \mathrm{tr}(\mathbf{Q}\mathbf{Q}^T). \tag{11}$$

Then the objective function *J* can be rewritten as

$$J = \tfrac{1}{2}\mathrm{tr}\left((\mathbf{D} - \mathbf{U}\mathbf{V}^T)(\mathbf{D} - \mathbf{U}\mathbf{V}^T)^T\right) = \tfrac{1}{2}\mathrm{tr}\left(\mathbf{D}\mathbf{D}^T\right) - \mathrm{tr}\left(\mathbf{D}\mathbf{U}\mathbf{V}^T\right) + \tfrac{1}{2}\mathrm{tr}\left(\mathbf{U}\mathbf{V}^T\mathbf{V}\mathbf{U}^T\right) \tag{12}$$

We have to solve the optimization problem with therespect to all entries u_{ij} and v_{ij} of the matrices **U** and **V**. In addition, since **U** and **V** are the non-negative matrices, we receive the constraints

$$\begin{aligned} u_{ij} \geq 0, i = 1 \ldots M,\ j = 1 \ldots K \\ v_{ij} \geq 0, i = 1 \ldots N,\ j = 1 \ldots K \end{aligned}. \tag{13}$$

This constrained non-linear optimization problem can be solve using Lagrange multipliers. Let **α** and **β** are the matrices of Lagrange multiplies of the same dimensionality as **U** and **V**. Then Lagrange expressions for the non-negativity constraints equals to

$$\begin{aligned} \sum_{i=1}^{M}\sum_{j=1}^{K}\alpha_{ij}u_{ij} = \mathrm{tr}(\boldsymbol{\alpha}\mathbf{U}^T) \\ \sum_{i=1}^{N}\sum_{j=1}^{K}\beta_{ij}v_{ij} = \mathrm{tr}(\boldsymbol{\beta}\mathbf{V}^T) \end{aligned}. \tag{14}$$

Then we can express Lagrangian optimization as

$$L = J + \mathrm{tr}(\boldsymbol{\alpha}\mathbf{U}^T) + \mathrm{tr}(\boldsymbol{\beta}\mathbf{V}^T). \tag{15}$$

To solve the problem we have to express partial derivatives of *L* with the respect to the matrices **U** and **V** and equal them to zeros.

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{U}} = -\mathbf{D}\mathbf{V} + \mathbf{U}\mathbf{V}^T\mathbf{V} + \boldsymbol{\alpha} = 0 \\ \frac{\partial L}{\partial \mathbf{V}} = -\mathbf{D}^T\mathbf{U} + \mathbf{V}\mathbf{U}^T\mathbf{U} + \boldsymbol{\beta} = 0 \end{aligned} \tag{16}$$

The solution of these equations leads to the iterative updating rules for **U** and **V**

$$u_{ij} = \frac{(\mathbf{DV})_{ij} u_{ij}}{(\mathbf{UV}^T\mathbf{V})_{ij}}$$
$$v_{ij} = \frac{(\mathbf{DU})_{ij} v_{ij}}{(\mathbf{VU}^T\mathbf{U})_{ij}} \quad . \tag{17}$$

The objective function J continuously improves when updating **U** and **V** using these rules and converges to the optimum.

NMF can also be used to express terms in the new latent space. As the columns of **V** determine the new dimensions for documents, the columns of **U** can be viewed as new dimensions for terms. Hence NMF is also useful for the condensation of training data because it enables to substitute the original M documents by the new K ones.

It has been shown that NMF is equivalent to the graph-structure based document clustering technique named Spectral Clustering (Ding et al., 2005). An analogous and more universal technique called Concept Factorization (Xu & Gong, 2003) can be also applied for an input matrix with negative entries.

3.1.2.3 Probabilistic latent semantic indexing

Despite the remarkable success of Latent Sematic Indexing (LSI) the method has the deficit in its unsatisfactory statistical foundations. This deficit overcomes Probabilistic Latent Semantic Indexing (pLSI) (Hofmann, 1999) because it introduces a simple generative model of the data that takes the advantage of the likelihood principle of the parameter estimation. pLSI is a statistical model with latent variables that is also called the Aspect Model. pLSI model assumes that documents and tokens are conditionally independent given unobserved topics. The approach offers to estimate a joint distribution of the triplets [document, topic, token] and thus enables to assign the topic probabilities for each document or to assign the most probable topic to a token in a particular document. The number of latent topics K must be selected before the estimation of the model and is usually significantly smaller than the size of the vocabulary N.

Let us have a collection D of M documents $\{d_1, d_2,\ldots,d_M\}$. The documents include the words from the vocabulary $V=\{w_1,w_2,\ldots,w_N\}$. The word order in a document is not taken into the account (the bag-of-words approach), but the co-occurrence of words is driven by an unobserved topic variable $Z=\{z_1,z_2,\ldots z_K\}$. The model can be viewed as the generative one following this three-step process:

- Select a document d with the probability $p(d)$.
- Pick a latent topic z with the probability $p(z|d)$.
- Generate a word w with the probability $p(w|z)$.

From the resultant triplet $[d,z,w]$ only the pair $[d,w]$ is observed while the latent topic z is unknown. The process can be also described by causal Bayesian network from Figure 4. Using the chain rule for the decomposition of the joint probability according to Bayesian network, we get the probability of the triplet $[d,z,w]$ in the form

$$p(d,z,w) = p(z \mid d)\, p(w \mid z)\, p(d)\,. \tag{18}$$

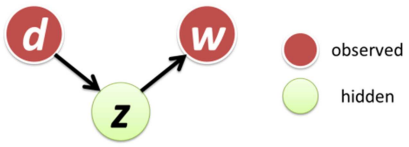


Figure 4: The causal schema of the probabilistic latent semantic indexing.

Note that Bayes formula applied to $p(z|d)$ enables to rewrite the joint probability as

$$p(d,z,w) = p(d \mid z)\,p(w \mid z)\,p(z)\,. \tag{19}$$

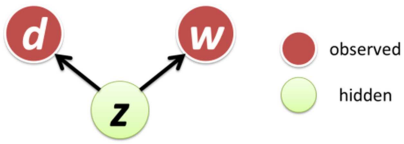


Figure 5: The rewritten schema of the probabilistic latent semantic indexing.

The formula (19) describes Bayesian network from Figure 5, hence both presented networks for the causality models are equivalent when estimating the joint probability. To derive the formula describing the Aspect Model one has to sum over the possible choices of z which could generate the observed pairs $[d,w]$.

$$p(d,w) = \sum_{z \in Z} p(d \mid z)\,p(w \mid z)\,p(z) = p(d) \sum_{z \in Z} p(w \mid z)\,p(z \mid d) \tag{20}$$

The Aspect Model is based on two assumptions. Firstly, the observed pairs $[d,w]$ are generated independently that corresponds with the bag-of-words approach. Secondly, the conditional independence assumption is made that conditioned on the latent class z, words w are generated independently of the specific document d. Hence the word distributions $p(w|d)$ are obtained by the combination of the aspects $p(w|z)$. The documents are then characterized by the specific mixture of the aspects with the weights $p(z|d)$.

To estimate the model we have to maximize the log-likelihood function

$$l = \sum_{d \in D} \sum_{w \in V} n(d,w) \log p(d,w)\,, \tag{21}$$

where $n(d,w)$ denotes the number of times a word w occurs in a document d. Due to the latent nature of the model the expectation-maximization algorithm (EM) (Dempster et al., 1977) must be used to estimate the desired probabilities. In the E-step the probability of the topic z behind the word w in document d is estimated as

$$p(z \mid d,w) = \frac{p(z)\,p(d \mid z)\,p(w \mid z)}{\sum_{z' \in Z} p(z')\,p(d \mid z')\,p(w \mid z')}\,. \tag{22}$$

The factors from the E-step are estimated in the M-step using the estimated probability and observed counts as

$$p(w \mid z) = \frac{\sum_{d \in D} n(d,w)\, p(z \mid d,w)}{\sum_{w' \in V} \sum_{d \in D} n(d,w')\, p(z \mid d,w')}$$
$$p(d \mid z) = \frac{\sum_{w \in V} n(d,w)\, p(z \mid d,w)}{\sum_{d' \in D} \sum_{w \in V} n(d',w)\, p(z \mid d',w)} \,. \qquad (23)$$
$$p(z) = \frac{\sum_{d \in D} \sum_{w \in V} n(d,w)\, p(z \mid d,w)}{\sum_{d \in D} \sum_{w \in V} n(d,w)}$$

The alternation of the E-step and the M-step defines a convergent procedure that approaches a local maximum of the log-likelihood function. Then among the estimated probabilities of the generating topic $p(z|d,w)$ we can substitute each word w in the document d by its the most probable generative topic z. Due to the fact that the number of topics is considerably smaller than the size of vocabulary N, this substitution leads to an important dimensionality reduction.

The Aspect Model can be also rewritten in a matrix notation. Let the conditional probabilities $p(d|z)$ and $p(w|z)$ create matrices $\mathbf{P}$ and $\mathbf{Q}$ of sizes $M \times K$ and $N \times K$ respectively. Similarly let the probabilities $p(z)$ create the diagonal square matrix $\mathbf{\Lambda}$ of the size $K \times K$. Then the joint probability of the document d and the word w

$$p(d,w) = \sum_{z \in Z} p(d \mid z)\, p(w \mid z)\, p(z) \qquad (24)$$

from the Aspect Model (19) can be written as the matrix product

$$\mathbf{D} = \mathbf{P}\mathbf{\Lambda}\mathbf{Q}^T \,, \qquad (25)$$

where the matrix $\mathbf{D}$ is the $M \times N$ matrix of the join probabilities $p(d,w)$. From the above formulas the correspondence between LSI and pLSI is apparent. The left and right eigenvectors of $\mathbf{D}$ from (5) correspond to the conditional probabilities $p(d|z)$ and $p(w|z)$ and the singular values correspond to the probabilities $p(z)$.

Despite this similarity, there is also a fundamental difference between pLSI and LSI in the objective function utilized to determine the optimal solution. In LSI it is Frobenius norm, which corresponds to an implicit additive Gaussian noise assumption on counts. In contrast, pLSI relies on the likelihood function of the multinomial sampling that aims to maximize the predictive power of the model. It offers the important advantages in the interpretation of results; the matrices include well-defined probabilities and the factors have the clear probabilistic meaning in terms of mixture component distributions.

On the contrary, the main disadvantage of pLSI is the lack of generalization. pLSI introduces a dummy index d of documents in the training set to the model. Thus d is the multinomial random variable with M possible values and the model learns the topic mixtures $p(z|d)$ only for those documents on which it is trained. Hence pLSI cannot be a universal generative document model because there is no straightforward way how to assign the probabilities to previously unseen documents.

A further problem that pLSI introduces also stems from the use of the distribution indexed by training documents. The number of parameters to be estimated grows linearly with the number of training documents. The parameters for a *K*-topic pLSI model are two *K*-multinomial distributions of size *M* and *N*. It results in *KM+KN* parameters. The linear growth in parameters suggests that the model is prone to overfitting and the problem should be solved for example by the subsequent smoothing.

3.1.2.4 Latent Dirichlet allocation

Latent Dirichlet Allocation (LDA) (Blei et al., 2003) overcomes the problems with the lack of generalization of pLSI by treating the topic mixture weights as a *K*-parameter hidden random variable. It does not introduce a large set of individual parameters which are explicitly linked to the training documents. LDA is a well-defined generative model and generalizes easily to new documents. For the vocabulary of the size *N* a *K*-topic LDA model needs only *K+KN* parameters.

LDA is another method that exploits the bag-of-words assumption; the order of words in a document is not important. Moreover, LDA also assumes that documents are exchangeable; the order of the documents in a collection is unimportant. De Finetti representation theorem (de Finetti, 1974) establishes that any collection of exchangeable random variables has a mixture distribution representation. Hence LDA introduces the mixture model that captures the exchangeability of both words and documents. The mentioned assumption of exchangeability is not equivalent to the assumption that the random variables are independent and identically distributed. Rather the exchangeability can be interpreted as a conditional independence, where the conditioning is with the respect to an underlying latent parameter of a probability distribution. Thus, while the exchangeability is clearly the major simplifying assumption that leads to computationally efficient methods, it does not necessarily lead to approaches that are restricted to simple frequency counts or linear combinations. The model can capture a significant intra-document statistical structure via the mixing of distributions.

Let us have a collection *D* of *M* documents $\{d_1, d_2,\ldots,d_M\}$. The documents include words from the vocabulary $V=\{w_1,w_2,\ldots,w_N\}$. LDA is the generative probabilistic model of the collection *D* that generalizes well for unseen documents. The basic idea is that the documents are represented as random mixtures over latent topics $Z=\{z_1,z_2,\ldots z_K\}$, where each topic is characterized by the distribution over words. The matrix **B** of the size $K\times N$ is the matrix of word probabilities; each row is the word distribution for the different topic. The non-negative *K*-dimensional vector $\boldsymbol{\alpha}$ includes the proportions of topics in the whole collection.

The following generative process describes the idea behind LDA:

- For each document in the collection choose $\boldsymbol{\theta}\sim Dir(\boldsymbol{\alpha})$.
- For each word in the selected document:
  - Choose a topic $z\sim Mult(\boldsymbol{\theta})$.
  - Choose a word $w\sim Mult(\mathbf{zB})$.

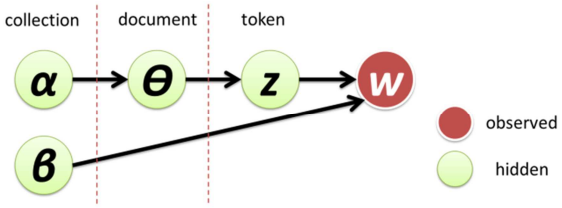


Figure 6: The causal schema of Latent Dirichlet Allocation model.

The vector **z** is a one-of-*K* binary vector; the product **zB** is the row from the matrix **B** with word probabilities conditioned by the topic *z*. The parameters *K*, **α**, **B** are treated as the fixed quantities. *K* is the dimensionality of Dirichlet distribution and must be set in advance. **α** and **B** are the hidden collection parameters to be estimated.

The *K*-dimensional Dirichlet random variable **θ** is the hidden property of each document. It is the probability vector that can take values from the (*K*−1)-simplex and it has the probability density on this simplex

$$p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) = \frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \prod_{i=1}^{K} \Theta_i^{\alpha_i - 1}, \tag{26}$$

where $\Gamma(x)$ is the Gamma function (the factorial generalized for non-integers). This Dirichlet distribution is the distribution on the simplex that is the conjugate to the multinomial distribution.

Note that the generative process of LDA does not operate with the number of documents and with the document lengths. These random values can be modeled separately using any arbitrary distributions; the choices do not influence the generative process.

Given the collection parameters **α** and **B** one can derive the joint distribution of the topic mixture **θ**, the set of topics **z**, and the set of words **w** on the document level as

$$p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w} \mid \boldsymbol{\alpha}, \mathbf{B}) = p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) \prod_{j=1}^{L} p(z_{(j)} \mid \boldsymbol{\theta})\, p(w_{(j)} \mid z_{(j)}, \mathbf{B}), \tag{27}$$

where *L* is the length of a particular document, *j* is the rank of a word in the document. The probability $p(\boldsymbol{\theta}|\boldsymbol{\alpha})$ is from (26), $p(z_{(j)}|\boldsymbol{\theta})$ is simply θ_i where *i* is the index of the topic behind j^{th} word ($z_{(j)}=z_i$), and $p(w_{(j)}|z_{(j)},\mathbf{B})$ is β_{ij}. Integrating over **θ** and summing over *Z* we obtain the marginal distribution of the document

$$p(\mathbf{w} \mid \boldsymbol{\alpha}, \mathbf{B}) = \int_{\boldsymbol{\theta}} p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) \prod_{j=1}^{L} \sum_{z_{(j)} \in Z} p(z_{(j)} \mid \boldsymbol{\theta})\, p(w_{(j)} \mid z_{(j)}, \mathbf{B}) d\boldsymbol{\theta}. \tag{28}$$

Finally multiplying the marginal probabilities of single documents, we obtain the probability of the whole collection

$$p(D \mid \boldsymbol{\alpha}, \mathbf{B}) = \prod_{d \in D} \int_{\boldsymbol{\theta}_d} p(\boldsymbol{\theta}_d \mid \boldsymbol{\alpha}) \prod_{j=1}^{L_d} \sum_{z_{(d)(j)} \in Z} p(z_{(d)(j)} \mid \boldsymbol{\theta}_d)\, p(w_{(d)(j)} \mid z_{(d)(j)}, \mathbf{B}) d\boldsymbol{\theta}_d \tag{29}$$

The idea of the generative process can be visualized using the causal network from Figure 6. The figure makes clear that there are three levels in the LDA representation. The parameters **α** and **B** are the collection level parameters, assumed to be sampled once in the process of generating all documents. The vectors $\boldsymbol{\theta}_d$ are document-level variables sampled once per document. Finally, the variables $z_{(d)(j)}$ and $w_{(d)(j)}$ are word-level variables and they are sampled once for each word in each document. In this 3-level model the topics are sampled repeatedly

within each document, hence each document is associated with multiple topics. Models that are similar to that shown on Figure 6 are referred as hierarchical models (Gelman et al., 2009), or more precisely as conditionally independent hierarchical models (Kass & Steffey, 1989).

The main inferential goal that we need to resolve in order to use LDA is that of computing the posterior distribution of the hidden probabilities **θ** and the generating topics **z** for a given document.

$$p(\boldsymbol{\theta}, \mathbf{z} \mid \boldsymbol{\alpha}, \mathbf{B}) = \frac{p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w} \mid \boldsymbol{\alpha}, \mathbf{B})}{p(\mathbf{w} \mid \boldsymbol{\alpha}, \mathbf{B})} \tag{30}$$

The denominator in the terms of the model parameters takes form

$$p(\mathbf{w} \mid \boldsymbol{\alpha}, \mathbf{B}) = \frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \int_{\boldsymbol{\Theta}} \left(\prod_{i=1}^{K} \theta_i^{\alpha_i - 1} \right) \left(\prod_{a=1}^{L} \sum_{i=1}^{K} \prod_{b=1}^{N} (\theta_i \beta_{ib})^{\delta_{ab}} \right). \tag{31}$$

δ_{ab} is the 1/0 function, it equals one if and only if $w_{(a)}=w_b$. Unfortunately, the above formula (31) is intractable due to the coupling between **θ** and **B** which is the obstacle in the estimation from (30). Although the posterior distribution is intractable for the exact inference, a wide variety of approximate inference algorithms can be considered for LDA, including Laplace approximation, the variational approximation, and Markov chain Monte Carlo. Even though LDA model cannot be estimated exactly due to its remarkable robustness for document modeling, there are several available implementations of variational methods in C, Java, or Matlab.

LDA may not be used for the modeling of documents only as described in the previous text; it has various potential extensions. For example, LDA can be applied to the continuous data or other non-multinomial data. As in the case of other mixture models, the multinomial emission probabilities $p(w|z)$ can be substituted by any more realistic distributions. For example, in the continuous variant of LDA Gaussian observables are used in the place of the multinomials. Another simple extension of LDA comes from allowing mixtures of Dirichlet distributions in the place of the single Dirichlet of LDA. This offers a richer structure in the latent topic space and in particular allows a form of the document clustering.

3.1.3 Principal language dependent dimensionality reduction methods

There are many language dependent approaches that derive different features from a text. They successfully exploit the known vocabulary, the morphology or the syntax of a particular language. Their accuracy is generally greater than the accuracy of language independent algorithms, but they rely on often huge and specific linguistic resources in the form of vocabularies, rule sets or libraries. Hence a lot of the computing power is often needed and an adjustment to new languages is not trivial.

The language dependent algorithms that convert or tag texts vary from fundamental algorithms such as the stemming or the part-of-speech tagging to highly specialized ones such as the sentiment recognition. They offer different features to structurally represent documents but the features are mostly too specific for a particular natural language which disables to switch a solution into the different language. Let us briefly review only the basic known

approaches that effectively lower the dimensionality of a text and they are available for many languages.

3.1.3.1 Stemming

The stemming is the process for reducing words to their stem, base or root form. The **stem** need not be identical to the morphological root of the word; it is usually sufficient that related words map to the same stem, even if this stem is not the valid root. The main benefit of the stemming for the dimensionality reduction is that the different words of similar meanings are assigned by the same generating stem.

The stemming is not available for any natural language. For example, Chinese does not allow the stemming. On the contrary, the rule-based morphology of Indo-European languages enables to construct efficient rule-based stemmers. In this language group the root of a word is surrounded by prefixes and suffixes. In the processes of stemming we often focus only on the suffixes. The suffixes can be categorized into three main groups:

A-suffix (attached) has a form of different word. Words with A-suffixes are the compound words.

I-suffix (inflectional) creates an inflectional form of a word. It meets the morphological rules of a language, but exceptions may exist. Some I-suffixes also alter the root.

D-suffix (derivational) changes the meaning of a word or even its part-of-speech. Even though some morphological rules are available for the D-suffixes, vocabularies are necessary to recognize D-suffixes sufficiently.

The stemming algorithms can be divided into several categories as well. They differ in their precision, efficiency or performance. Even though the stemming is the language dependent procedure, the stemmers are usually reasonable fast and compact. We can distinguish four groups of the stemming algorithms:

Brute force algorithms do not rely on linguistic rules. A table including the pairs of word-stem is the core component of the brute force algorithms. They search the table for each input word to find the correct stem. It is rather labor intensive to develop the lookup table that covers the most of the vocabulary of a particular language. On the contrary, the pair list easily covers all the exceptions.

Suffix stripping algorithms exploit a relatively small list of linguistic rules to strip the suffixes from input words. Their development is rather simple, but the developer must have the sufficient knowledge of the morphology of a particular natural language. The suffix stripping algorithms hardly recognize the exceptional stems; their efficiency differs among languages depending on the perplexity of the language.

Stochastic algorithms exploit the probability theory and statistics. A statistical model is the main component of the stochastic stemmers. The model is adjusted on the training examples that include correct pairs word-stem. The statistical model is usually in the form of inferred association rules between stems and words. These rules are used to recognize the most probable stems of new words.

Hybrid algorithms combine the above mentioned approaches. They can use the lookup tables together with the expert or inferred rules. For example, in the first instance a small table including the exceptions is searched and then the rule set is applied to strip out regular suffixes.

The first stemmer was published in the late 1960s (Lovins, 1968). Probably the most popular stemmer was written by Martin Porter and firstly published in the 1980s (Porter, 1980). Many

implementations of the Porter stemmer are freely available, some of them with enhancements. Martin Porter extended his work in the 2000s when he released Snowball (Porter, 2000), a framework for writing stemming algorithms. Stemmers for several languages are now available in Snowball; it became a standard platform for the development of new stemmers.

3.1.3.2 Lemmatization

The lemma or the citation form is the grammatical form that is used to represent a word concerning its meaning. For example, the infinitive form is used as the lemma for wordforms of verbs. The lemmatization is then a process of the mapping the different inflected forms of a word to the lemma, so they can be analyzed as a single item.

The lemmatization can substitute the stemming. The difference is that the stemmer operates on a single word without the knowledge of its context. Therefore stemmers cannot discriminate between words that have different meanings depending on their parts-of-speech. However, the stemmers are typically easier to implement and they run faster comparing to the lemmatizers.

On the contrary to stems, lemmas are part-of-speech specific. The same wordform can be mapped to different lemmas depending on its part-of-speech that can be recognized only from the context. Hence the lemmatization is usually performed together with the part-of-speech tagging considering the features from the neighborhood of the word. The combination of the lemma with the part-of-speech is often called the lexeme of a word.

Similarly to the stemmers, the lemmatizers can be rule-based or dictionary-based. The rule-based algorithms exploit the property of the word together with the features extracted from its context to determine the correct lemma. The dictionary-based approaches rely on the dictionary of citation forms. The both approaches are combined in the hybrid lemmatizers. They search for the lemma in two steps. In the first step the set of possible lemmas is determined from the dictionary of citation forms. The second optional step is executed if more than one lemma can be assigned to the wordform due to its ambiguity. The most probable lemma is selected based on the word context and other known features.

3.1.4 Dimensionality reduction appendix

The above description of dimensionality reduction methods used in text mining comprises widely used approaches. These approaches are often modified or adjusted to better fulfill goals of a text mining task or to fit to a particular document processing pipeline. For example, the feature selection methods are usually modified to select variable sets with small inter-set correlation to avoid aspects of multi-collinearity for regression models. It results in common heuristic procedures of variable subset selection such as forward or stepwise selection equipped by text mining criteria for feature inclusion or exclusion.

The reduction methods are often combined in the text processing pipelines as well. A resultant multi-step dimensionality reduction offers a better control over the process and enables a finer adjustment of optional parameters.

The main purpose of the thesis is to propose an alternative feature extraction and reduction approach that exploits an order in which standard features appear in a text. The natural order of linguistic entities in a text is often neglected in text mining applications[6], but the order is critical for a human reader to understand ideas of a document. An omission of a word order can lead to serious mistakes in natural language processing tasks such as machine translation,

[6] In some advanced representations that exploit language dependent linguistic resousces the context encoding is restricted to an extraction of multi-word terms.

question answering or text summarization. Hence it is worth exploring how the contiguity of linguistic entities may help to improve a performance of standard text mining models.

4 Proposed context document representation

Any document is a sequence of vocabulary terms. The frequency of terms within a document is an important statistics for the document representation, but we have to take into account the term adjacency as well. Hence the document can be regarded as a container of ordered groups of the terms of the fixed length. The groups are referred as n-grams where *n* stands for the fixed length of the group. This perspective comes from the standard n-gram language modeling where the occurrence of a term in the text depends on the presence of (*n*-1) previous terms[7].

The usage of n-grams as the basic building blocks of texts further significantly magnifies the dimensionality problem that arises from the size of natural vocabularies. Natural languages consist of tens of thousands of words, considering the word n-grams the resultant dimensionality would be of a much higher magnitude. On the other hand, representative vectors of a reasonable dimensionality are desirable for the further document processing. Such vectors should comprise as much as possible information that enables to distinct among different documents or to merge the similar ones.

The giant dimensionality difference between the n-gram representation and the representation suitable for predictive models requires a multi-step dimensionality reduction. The proposed document representation utilizes three reductive steps:

1. The transformation of terms to latent topics. This reduction step does not take into account the order or the distance among the words in a text, but reveals the latent topics hidden behind the text exploiting the co-occurrences of terms in documents of a training collection.
2. The construction of document context networks of the topics. The context networks include the information about the topics´ neighborhood; the topics that occur closely in the text are strongly connected in the network.
3. The centrality vectors extraction. The vector of importances of the topics in the context network is derived as the final representation of a document. The centralities of the topics that reflect their positions in the context network are used to quantify the topics´ involvements together with their closeness in the text.

The detail description of this three-step process of the derivation of the reasonable vector representation of documents follows. Let us have the training collection $D=\{d_1,d_2,\ldots,d_M\}$ of M documents that is available to train the topic model from the first step. Then all three steps can be performed to obtain the proposed representation of any document regardless of the fact if it belongs or not to the training collection D.

Figure 7: The representations of a text document in the proposed processing pipeline.

4.1 Transformation of terms to latent topics

In the first step, the size of the vocabulary can be reduced by several methods that take into account the common appearance of the terms in the documents belonging to the training

[7] In real documents the length of the context that influences the appearance of particular terms is not probably fixed. The variable context length is difficult to introduce to the presented approach but the large enough context length may smooth these irregularities.

collection. They are often referred as the topic modeling methods. The popular topic modeling methods are described in chapter 3.1.2. We propose to use Latent Dirichlet Allocation (LDA) to substitute terms in a text by the topics in the first step. LDA exploits a flexible generative model of the training document collection. The derived model can be simply applied to new documents. LDA describes the collection as the whole using Dirichlet distribution and also offers a description of each document by fitting its multinomial distribution of the topics. Above all, it enables to substitute each term in a document by the topic which is exploited in the proposed representation. LDA is described in details in chapter 3.1.2.4.

To utilize LDA effectively several simple preprocessing procedures can be helpful. They include the term filtering and stemming. The term filtering deletes useless and unknown terms from the text. The useless terms include prepositions, conjunctions, particles etc. They are available in special lists called stop-word lists. Even though stop-word list usage is the language dependent procedure, it does not require vast resources. There are several hundreds of simple stop-words in each natural language, hence the filtering is reasonably fast. The stop-word filtering reduces the noise in the text that can be generated by the meaningless terms.

To recognize the same meaning of two or more different wordforms, we recommend to preprocess the text by stemming. The stemming removes word suffixes and optionally also prefixes that form the wordforms. It is not necessary to strip the words to the grammatical root; the main purpose of stemming is to unify all forms of the same word. This procedure further reduces the input vocabulary that enters to LDA. The stemming is also the language dependent procedure that does not require huge resources. The common stemmers are rule based ones; the stemmers usually include several hundreds of grammatical rules that trim the words to the common forms.

The lemmatization is an alternative to the stemming. The lemmatization enables to substitute a wordform by a basic grammatical form called the lemma. For example, all forms of nouns are substituted by the first case of their singular form. The lemmatization is tightly associated with a particular language and requires more computational resources. It is also rather difficult to acquire lemmatizers for uncommon languages. Hence the stemming is preferred to the lemmatization in the proposed process.

Similarly to the stop-word filtering non-linguistic entities should be filtered out from the text. The non-linguistic entities include numbers or URLs. Even though the non-linguistic entities appear often in texts, their actual forms are infrequent; there are many but different non-linguistic entities in the documents. Hence this filtering procedure is covered by more general exclusion of non-dictionary terms. LDA that follows can substitute the known terms[8] by latent topics, the unknown terms are always omitted. The vocabulary of known terms is built in advance usually using the training set of documents. The infrequent terms are generally excluded from the dictionary because their co-occurrence with other terms does not enable to estimate parameters of joint term distributions reliably. The usual vocabulary that enters to LDA consists of thousands or tens of thousands of terms.

After the input vocabulary is set, the LDA model can be trained for the given number of hidden topics. The number of the topics is fixed for all documents and has to be set in advance. The appropriate number of the topics that are further used instead of the terms varies from units to hundreds. The number of the topics implies the dimensionality of the final document representation because the topics form the vertices of the context networks. The length of final document vectors is the same as the given number of the topics.

[8] More precisely the stemmed terms in the proposed process.

The LDA model is adjusted on the training set of documents to estimate parameters of prior Dirichlet distribution which influences the individual distributions of the topics in the documents. The individual topic distribution can be then derived for any document either from the training collection or for new ones. Using the conditional probabilities of the terms given the topics, the terms are substituted by the topics. These conditional probabilities are also estimated in advance using the training set of the documents. The input for the second step then consists of preprocessed documents where the useless terms are omitted and the other terms are substituted by the nominal topics.

Formally the first step reduces the size of the dictionary for the consequent steps. The dictionary $V=\{w_1,w_2,\ldots,w_N\}$ now consists of N topic entries. Each document d can be now regarded as a sequence of vocabulary items $d=w_{(1)}w_{(2)}w_{(3)}\ldots w_{(L)}$; the terms are substituted by the topics The bracketed indexes express the ranks of the topics in the document d; L stands for the document length and it is the number of the recognized terms that were replaced by the topics.

The first step also provides the basic document representation where the counts of the vocabulary topics form the document vector $\mathbf{d}^T=(v_1,v_2,\ldots,v_N)$. This representation does not comprise the context of the document; a permutation of the terms within the document does not influence its basic representation by the vector **d**. It is the bag-of-words representation[9].

4.2 Construction of context networks

In the second step, the individual context network is built for each document. The network structure, that can be regarded as oriented graph with weighted edges, reflects the adjacency of topics in the document. The topics constitute vertices of the context networks; the set of vertices is the same for all documents. The context networks of two documents differ only in the strengths of edges that reflect the relations among the topics in the text. The topics that often appear nearby in the text are connected more strongly than the topics that appear further.

To assess the adjacency of the topics, a context window has to be defined. The context window covers a sequence of the topics of the fixed length K. The context window includes an uninterrupted subsequence of a text. For each topic in a document one can investigate its left or its right context window depending on the fact whether the topic is the last or the first topic of the context window. Sliding the context window through a text we can explore the neighborhood of each topic. The joint topics counts[10] from each position of the context window in a text are aggregated to form the weights in the context network.

The distance of the topics within the context window is unimportant; all topics are regarded as the neighbors of the last or the first topic. The counts of the neighbor pairs summed over all the context windows within a document then serve as the weights in the context networks. The order of the topics within the pair implies the direction of the weighted connection in the context network.

To be able to construct the comparable context window for starting or terminal topics, we suggest to add a reasonable number of dummy starting or final terms before or after the text respectively[11]. All the dummy terms are explicitly substituted by the same additional dummy

[9] The more precise name would be the bag-of-topics representation.
[10] The pairs of the most right or left topic with any other topic in the window are only regarded.
[11] For the length K of the context window, (K-1) dummy terms are inserted before or after the text.

topic[12]. This dummy topic makes a special vertex of the context network with solely outgoing or ingoing connections.

Optionally the text can be divided into sentences[13] and each sentence can be wrapped by the dummy topics separately. In such approach the original neighborhoods that exceed the sentence borders are excluded; only the term pairs that appear inside sentences contribute to the connection weights of the context network.

The weights in the context network which represents a document can be comprised to a square matrix **G**. The number of rows and columns equals to the given number of the topics N. The context network is the directed network, hence the matrix **G** is not symmetric, but it includes non-negative integers only. Rows represent the vertices where the connections origin, columns stand for the terminal vertices. The set of the vertices $V=\{w_1,w_2,\ldots,w_N\}$ together with the weight matrix **G** form the context network $G=\{V,\mathbf{G}\}$ for a particular document. The matrix document representation itself is not appropriate as the final representation that enables the further fluent document processing, but it is important for the derivation of the centralities in the third step.

4.3 Extraction of centrality vectors

The centrality is a measure that reflects a position of a vertex among other vertices in a network. If we compute the centralities of the topics in the context network, we get a vector of the same dimensionality as the number of the extracted topics N. The number of the topics is set in advance before LDA is applied to the text. It enables to control the dimensionality of the proposed representation. Considering the robustness of data mining models together with the richness of natural languages we recommend to compromise the number of the topics to the magnitude of tens.

A centrality score $c(w_i)$ of a vertex w_i is always derived regarding its incoming and outgoing connections; some centralities take into the consideration also other ties in the network. Hence the centrality score reflects the intensity of the topic as well as its typical position among the other topics. Depending on the selected centrality measure we can emphasize the document content (intensity) or the context within the text (position among others). The centralities $c_i=c(w_i)$ of the vertices w_i from the context network G form the new proposed document vector representation $\mathbf{c}^{\mathrm{T}}(G)=(c_1,c_2,\ldots,c_N)$.

In the further experiments we conducted tests for nine common centralities. The formal derivation of all tested centralities is described in detail in chapter 4.5. All of the selected centralities take into account the strengths of the ties and their directions; the only non-directional tested centrality is Degree. To compute any non-directional measure the matrix **G** of ties should be symmetrized by averaging or summing of the conjugate weights that are equivalents to the discarding the arrows in the network diagram. Degree depends on the connections of the particular vertex only hence it can be computed as the row sum plus the column sum of the matrix **G**. The directional centralities that consider only the connections of the vertex they are computed for include InDegree and OutDegree. They are derived as the sole row sum or column sum respectively. Hence Degree can be decomposed as InDegree plus OutDegree. The other considered directional centralities take into account a wider neighborhood of the vertex they are computed for. They include Eigenvector, Authority, Hub, PageRank, Closeness and Betweenness. Closeness and Betweenness centralities differ from the other centralities because they rely on the lengths of paths through the context network G.

[12] The dummy topic covers the dummy starting or final term only. The dummy topic is assigned directly it is not the output from LDA.

[13] The sentences can be recognized early in the document processing after the tokenization.

Hence to compute them one has to transform the weights to the distances to form the complement context network where the weights are substituted by the distances. Higher weights imply smaller distances and vice versa. The distances for such centralities are computed as the inverse of the weights in the proposed approach.

The magnitudes of the before mentioned centralities are influenced not only by the relations among topics but also by the length of the document. Longer documents imply higher weights in their context networks because the topics appear more frequently in the text.[14] The sizing of the weights influences the centralities as well. Degree, InDegree and OutDegree grow with the increasing length of the document. On the other hand, higher weights imply shorter paths, hence Closeness and Betweeness decrease with the increasing length of the document. Other centralities namely Eigenvector, Authority, Hub and PageRank rely on eigenvectors and eigenvalues of the matrices derived from the context network, hence the scale of the network weights may or may not influence its magnitude depending on the actual software implementation.

The standardized versions of the centralities can be defined together with the unstandardized ones. They are adjusted to the size of the network to be comparable among networks with the different number of the vertices. The number of the vertices in the proposed context network is fixed, hence the centralities standardized for the size of the context network are not very useful. On the other hand, the adjustment of the centralities to the total sum of weights in the weighted directed network would help to compare the documents of the different lengths because the weight sum is proportional to the document length. Unfortunately, this adjustment is not straightforward, so we propose to use the unstandardized centralities and to modify the way how the document vectors are compared. For the processing of the documents regarding their content and context controlling for the document length we propose to compare only angles among the centrality vectors. To do so the document vectors that consist of the unstandardized centralities can be standardized to unity length or appropriate proximity measure should be selected for the comparisons among the unstandardized vectors.[15]

4.4 Examples of obtaining representations

Example 1: The construction of the context network

Let us have a vocabulary V = {a,b}. A document d = aababababaa[16] is a product of a 3-gram generative model. We will try to estimate the parameters of the generative model and to complete the context network representing the document *d*.

Firstly, we derive all possible 3-grams that stand for cells of the transition matrix **T**. We also need to identify the set of all 2-grams that represent rows of a transition matrix. The length of the vocabulary $|V|$ is equal to two, hence there are 2^3 3-grams {aaa, aab, aba, abb, baa, bab, bba, bbb} and 2^2 2-grams {aa, ab, ba, bb} in our generative model. The matrices **Q** and **P** of joint and conditional probabilities have 8 cells each (4 rows and 2 columns) and these matrices are unknown. We can estimate them from the transition matrix **T** of the same size. The matrix **T** includes counts of transitions among 2-grams and the vocabulary terms.

[14] The total sum of weights equals to the product of the document length times the size of the context window.
[15] The cosine similarity is the standard choice when comparing the documents in text mining; it can be applied to the centrality vectors as well.
[16] The letters substitute topics extracted by LDA in the examples.

$$\mathbf{T} = \begin{pmatrix} 0 & 1 \\ 3 & 0 \\ 1 & 2 \\ 0 & 0 \end{pmatrix}$$

Even though the actual length of the document *d* is 9 terms, we exclude the first two terms from the consequent computation of transitions, hence we claim that the length of the document is 7[17]. It is also the sum of the counts in the matrix **T**. The maximum likelihood estimates of joint and conditional probabilities are then

$$\mathbf{Q} = \begin{pmatrix} 0 & 1/7 \\ 3/7 & 0 \\ 1/7 & 2/7 \\ 0 & 0 \end{pmatrix}, \mathbf{P} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 1/3 & 2/3 \\ 0 & 0 \end{pmatrix}.$$

They are very rough estimates of the probabilities and there are a lot of zeros and ones in the matrix **P.** Hence some method of probability smoothing should be taken into account to estimate practically useful probabilities. Fortunately, we do not require the estimation of the matrices **P** and **Q** to derive the context network hence we need not solve the smoothing problem.

Now we construct the context network **N** of the document *d*. To do so we have to count the numbers of different terms in the context windows of each term. We use the left context window of two adjacent terms to comply with the generative 3-gram model. It means we are going to count the frequencies of terms in the subsequence of two preceding terms for each term in the document *d*[18]. The counts form the square context network matrix **G** of the size 2×2.

$$\mathbf{G} = \begin{pmatrix} 4 & 4 \\ 4 & 2 \end{pmatrix}$$

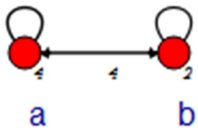


Figure 8: The visualization of the context network from the example 1.

The sum of all counts in **G** is equal to the length of the context window (*K*=2) times the length of the document (*L*=7). Using the matrix algebra the network matrix **G** can be obtained as a product of the transition matrix **T** and a fixed matrix **H** representing our context window of the length two.

[17] A more precise way to cope with the problem of start transitions is to extend our vocabulary by a dummy term *s* and put the reasonable number of terms *s* at the beginning of the document. The approach is not shown in the example because it extends further the number of possible n-grams and does not influence the proposed algorithm. However, in practical applications especially when coping with short texts the start terms should be considered.

[18] The nearly same context network would be obtained using the right context window. The resultant context networks would differ only in two starting or ending terms respectively.

$$\mathbf{G} = \mathbf{HT}$$

$$\mathbf{H} = \begin{pmatrix} 2 & 1 & 1 & 0 \\ 0 & 1 & 1 & 2 \end{pmatrix}$$

Rows of the matrix **H** represent the vocabulary terms and columns stand for all possible 2-grams. Integer values in **H** are then counts of the terms of the column 2-grams.

The resultant context network **G** has only two vertices and four oriented weighted edges including self-loops. The vertices represent the vocabulary terms, the edge weights are the sums of the transitions among the terms within all context windows in the document. Due to the small number of the vertices in the context network it is not appropriate to derive a centrality representation of our document *d* in this example. The next example is slightly more realistic and it ends with the selected centrality representations of a document.

Example 2: The derivation of the basic centralities from the context network.

Let us have a vocabulary V = {a,b,c,d}. A document d = bbacbadbacbaaadcba is a product of a 3-gram generative model. The size of the vocabulary N equals 4, the number of 2-grams M equals 16 and the length L of the document d is 16 (omitting the starting terms again). The left context window of the size K=2 will be used.

The matrix **T** of counts of transitions among the 2-grams and the vocabulary terms is of the size 16×4.

$$\mathbf{T}^T = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 3 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

A fixed matrix **H** that transforms the transition matrix **T** to a context network **G** has the form

$$\mathbf{H} = \begin{pmatrix} 2 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 2 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 2 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 2 \end{pmatrix}.$$

And the context network **G**=**HT** of the size 4×4 is then

$$\mathbf{G} = \begin{pmatrix} 3 & 3 & 3 & 3 \\ 7 & 0 & 2 & 1 \\ 3 & 3 & 0 & 0 \\ 1 & 2 & 1 & 0 \end{pmatrix}.$$

The basic centralities of the terms in the context network G can be calculated as the product of the weight matrix **G** and a matrix of a desired form. The InDegree and OutDegree centralities can be computed by the multiplication of **G** by a vector of ones from the left or from the right respectively. These products represent column or row sums of **G**. The Degree centrality is then the sum of InDegree and OutDegree. Hence we can easily conclude these centralities as

$$\mathbf{c}_{ID}^T = \mathbf{1}^T\mathbf{G} = \begin{pmatrix}14 & 8 & 6 & 4\end{pmatrix}$$
$$\mathbf{c}_{OD}^T = \mathbf{G}\mathbf{1} = \begin{pmatrix}12 & 10 & 6 & 4\end{pmatrix} \quad .$$
$$\mathbf{c}_{D}^T = \mathbf{c}_{ID}^T + \mathbf{c}_{OD}^T = \begin{pmatrix}26 & 18 & 12 & 8\end{pmatrix}$$

Note that the sum of all InDegrees as well as Outdegrees is equal to 32, which is the product of the document length and the context window size. The presented centralities are the linear combinations of the transition counts from the matrix **T**. Other more complex centralities generally cannot be calculated as a simple matrix product. Some centralities require eigenvalue and eigenvector computations while the others depend on distances in the context network. The next example illustrates such centralities in a larger detail.

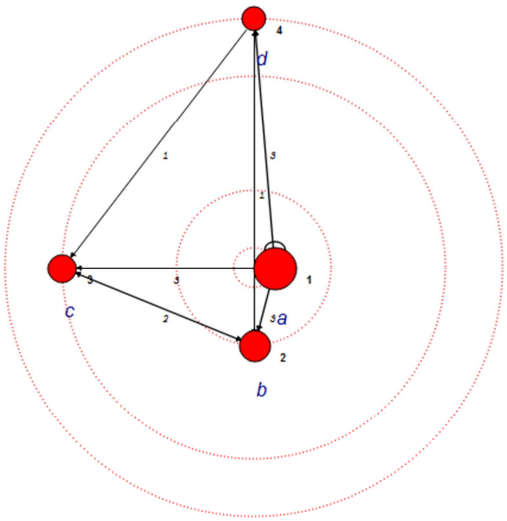


Figure 9: The visualization of the context network from the example 2. The sizes of the nodes are proportional to InDegree, the distance from the center is inversely proportional to Degree.

Example 3: The derivation of some advanced centralities from the context network.

Let us have the same vocabulary *V* and the document *d* as in the previous example. It has been already shown how the context network **G** is derived from the document *d*. Now let us present the centrality vectors that cannot be derived as a linear function of the matrix **G**.

Firstly we can compute the importance of the nodes in the whole context network. Regarding the fact that the importance of a node is influenced by the importances of its neighbors, we need to derive eigenvalues and eigenvectors. The Eigenvector centrality for a node is just an item of the eigenvector of the matrix **G**. Hence the proposed centrality representation of the document *d* is the eigenvector of its context network **G**. The eigenvector that is assigned to the largest eigenvalue is used because it is guaranteed that its items are non-negative real values. For our document *d* the Eigenvector centrality representation is

$$\mathbf{c}_{E}^T = \begin{pmatrix}0.63 & 0.61 & 0.41 & 0.25\end{pmatrix}.$$

If any eigenvector is multiplied by a constant, it is still regarded as the same eigenvector. Eigenvectors are usually provided as vectors with the unity length therefore no adjustment to the document length is necessary.

The more common centralities that express the importance of the node based on importances of its neighbors are Authority and Hub. Authority depends on the incoming ties and Hubs of the neighbors while Hub depends on the outgoing ties and Authorities of the neighbors. That

is why these centralities are usually provided together. Authorities of the nodes of our context network **G** are the items of an eigenvector of the matrix $\mathbf{G}^T\mathbf{G}$ and Hubs are the items of an eigenvector of the matrix $\mathbf{G}\mathbf{G}^T$. The matrices $\mathbf{G}^T\mathbf{G}$ and $\mathbf{G}\mathbf{G}^T$ are symmetric, hence they eigenvalues and eigenvector are real. Furthermore, they share the same eigenvalues; the eigenvectors are generally different. The eigenvectors that come with the largest eigenvalue are usually used as Authorities and Hubs. For our document *d* the Authority and Hub centrality representations are

$$\begin{aligned}\mathbf{c}_A^T &= (0.84 \quad 0.33 \quad 0.35 \quad 0.25)\\ \mathbf{c}_H^T &= (0.56 \quad 0.72 \quad 0.37 \quad 0.20)\end{aligned}.$$

Similarly to Eigenvectors, Authorities and Hubs are usually provided in the normalized form, hence any length adjustment is not necessary.

PageRank is another centrality that expresses the importance of a node based on its ties and the importance of neighbors. Apart from the fact that PageRank can be computed as a solution of a matrix equation, it can be also obtained as a result of a simulation process. If we simulate many random walks with breaks through the network, PageRank of a node is the probability that the random walk goes through the node. Hence the PageRank centralities are often provided as probabilities that sum to one. For our document *d* the PageRank centrality representation is

$$\mathbf{c}_{PR}^T = (0.39 \quad 0.27 \quad 0.20 \quad 0.14).$$

If the vector of PageRanks is adjusted to the unity length, we receive

$$\mathbf{c}_{PR}^T = (0.74 \quad 0.50 \quad 0.37 \quad 0.27).$$

The other centralities that are investigated in this work are based on distances or path lengths among the nodes in a network. To compute the representations of the document *d* using these centralities, the context network **G** has to be modified. Instead of weights that reflect how closely the vocabulary terms appear in the document we have to introduce the distances of the vocabulary terms. The inverse of the weight can be used as the distance. Hence our matrix **G** transformed to distances has the form

$$\overline{\mathbf{G}} = \begin{pmatrix} 1/3 & 1/3 & 1/3 & 1/3 \\ 1/7 & 0 & 1/2 & 1 \\ 1/3 & 1/3 & 0 & 0 \\ 1 & 1/2 & 1 & 0 \end{pmatrix}.$$

The zero entries in the matrix do not mean zero distances; they imply that the nodes are not connected. The Betweenness centrality for a node expresses the number of the shortest paths among the others nodes in the network that include this node. The Betweenness representation of the document *d* is

$$\mathbf{c}_B^T = (4 \quad 2 \quad 0 \quad 0).$$

Four shortest paths between pairs of terms from {b,c,d} go through the term a, two shortest paths between pairs of terms from {a,c,d} go through the term b etc. The Betweenness vector adjusted to unity length is

$$\mathbf{c}_B^T = (0.89 \quad 0.45 \quad 0.00 \quad 0.00).$$

The Closeness centrality express how close is a node to all other nodes in a network. Closeness is the inverse of Farness. Farness is the sum of distances to all other nodes. The small Closeness means that a node is not far away from other nodes. The Closeness representation of the document *d* is

$$\mathbf{c}_C^T = \begin{pmatrix} 3.00 & 2.74 & 2.25 & 1.42 \end{pmatrix},$$

and the adjusted version has the form

$$\mathbf{c}_C^T = \begin{pmatrix} 0.62 & 0.56 & 0.46 & 0.29 \end{pmatrix}.$$

It is hard to select the best universal representation of the document *d* among the proposed document vectors; the appropriateness may be influenced by a data mining task, by a modeling algorithm, by linguistic entities extracted from a text or by a domain of a document collection. Such investigation of the proposed document representations relevance is presented in the experimental part of the thesis.

4.5 Context network centralities

The described representation exploits selected common centralities originally proposed in Social Network Analysis theory (SNA). This chapter formally introduces these centralities and offers formulas to compute them.

The context network $G=\{V,\mathbf{G}\}$ consists of a set of N nodes or vertices $V=\{w_1,w_2,\ldots,w_N\}$ and a set of directed edges or ties between pairs of the nodes. The weight G_{ij} is assigned to the edge from the node w_i to the node w_j. If the edge between nodes w_i and w_j is missing, its weight G_{ij} is set to zero. The weights are arranged in a matrix **G**. The matrix **G** is the square of the size NxN, but not the symmetric one because the network G is directed. Rows of **G** represent nodes where the edges originate, columns stand for the nodes where the edges terminate.

The context network is derived for each document as an intermediate step in the process of extracting the proposed document representation. It can be viewed as a realization of the random matrix that is derived from other random matrix of transitions among possible n-grams (see chapter 5.2). The vertices from the set V are topics that were extracted from a text; the tokens in the text are substituted by the topics before the context network is built. The weights G_{ij} are proportional to the frequency of co-occurrence of topics w_i and w_j within the context window of the fixed length K.

The context network G offers statistics of its vertices w_i that are derived from the weight matrix **G**. These statistics are called centralities[19] because they serve as a description of the position of the vertex w_i among the other vertices. The proposed context networks for different documents consist of the same set of the vertices V, but they differ in the weights of the ties **G** that are specific for particular documents. Therefore the vertex centralities form the vector of the fixed length which serves as the proposed representation of a document.

The node centralities can be aggregated together to form a global property of the whole network. Usually the centrality variability within the network is examined to characterize the network structure[20]. Such variability indexes could be also used as scalar representations of documents. Unfortunately, the reduction of information that arises from the contraction of the context network into a single score is so enormous that the scalar representation of documents does not enable to process them accurately enough.

[19] Some authors distinguish centrality measures and prestige measures.

[20] For example the standard deviation of a centrality within the network can be used. More often the variability indexes compare the node centralities with the largest centrality in the network.

Depending on the actual centrality statistics, the proposed vector representation specifically comprises the adjacency of the topics within a document. Let us review the centrality measures that are tested in the experimental part of this work. They come from the general Social Networks Analysis theory (SNA). The presented centralities are not the only ones that can be used; SNA offers many others, but they are commonly used in other applications of social networks. Even though the centralities are often adjusted to the size of a network[21] to be comparable among the networks of the different sizes, the proposed representations do not use the adjusted versions because the number of the nodes is fixed for all documents in a processed collection. In following paragraphs only the centrality versions that apply for directed weighted networks are presented. For example, weighted InDegree is simply referred as InDegree.

4.5.1 InDegree

For a node w_i the sum of weights assigned to adjacent connections that terminate in the node w_i is called InDegree of the node w_i. A vertex with zero InDegree is called a source, as it is the origin of each of its incident edges. InDegree is often interpreted as a form of the popularity of the node w_i. The InDegree centrality can be computed as the column sum in the weight matrix **G**.

$$c_{ID}(w_j) = \sum_{i=1}^{N} G_{ij} \tag{32}$$

The sum of InDegrees of all vertices equals the sum of all entries of the weight matrix **G**. In the matrix notation the whole InDegree centrality vector that serves as the proposed document representation is computed as

$$\mathbf{c}_{ID}^{T}(G) = \mathbf{1}^{T}\mathbf{G}. \tag{33}$$

InDegree ranges from zero for the source nodes and has no upper limit for the weighted networks.

4.5.2 OutDegree

For a node w_i the sum of weights assigned to adjacent connections that originate in the node w_i is called OutDegree of the node w_i. A vertex with zero OutDegree is called a sink, as it is the end of each of its incident edges. OutDegree is sometimes called the branching factor of a node. OutDegree is often interpreted as a form of gregariousness of the node w_i. The OutDegree centrality can be computed as the row sum in the weight matrix **G**.

$$c_{OD}(w_i) = \sum_{j=1}^{N} G_{ij} \tag{34}$$

The sum of OutDegrees of all vertices equals the sum of all entries of the weight matrix **G**. The same is true for InDegrees, hence the sum of OutDegrees is equal to the sum of InDegrees. In the matrix notation the whole OutDegree centrality vector that serves as the proposed document representation is computed as

[21] Number of nodes in the network.

$$\mathbf{c}_{OD}(G) = \mathbf{G1}. \tag{35}$$

OutDegree ranges from zero for the sink nodes and has no upper limit for the weighted networks. If it holds $c_{OD}(w_i)=c_{ID}(w_i)$ for every node w_i, the network is called the balanced network.

4.5.3 Degree

Conceptually the simplest centrality is Degree, which is defined as the sum of weights assigned to all incoming and outgoing connections incident upon a node w_i. A vertex with zero Degree is an isolated vertex; there are no connections to or from the isolated vertex. The degree can be interpreted in terms of the immediate risk of a node for catching the information that is flowing through the network. The Degree centrality can be computed as the sum of InDegree and OutDegree.

$$c_D(w_k) = c_{ID}(w_k) + c_{OD}(w_k) = \sum_{i=1}^{N} G_{ik} + \sum_{j=1}^{N} G_{kj} \tag{36}$$

The sum of Degrees of all vertices equals twice the sum of all entries of the weight matrix **G** and also equals the sum of InDegrees plus the sum of OutDegrees. In the matrix notation the whole Degree centrality vector that serves as the proposed document representation is computed as

$$\mathbf{c}_D(G) = \left(\mathbf{1}^T \mathbf{G}\right)^T + \mathbf{G1} = \left(\mathbf{G}^T + \mathbf{G}\right)\mathbf{1}. \tag{37}$$

This formula also tells us that Degree is computed as the row or column sum of the symmetrized weight matrix. Degree ranges from zero for the isolated nodes and has no upper limit for the weighted networks.

4.5.4 Authority and Hub

Authority and Hub scores were introduced in the link analysis algorithm that rates web pages (Kleinberg, 1999). They were precursors to PageRank. The Authority and Hub centralities are usually computed together and they are referred as HITS[22]. A node with the high Hub score points to important Authorities and a node with the high Authority score links by important Hubs. This scheme therefore assigns two scores to each node in a network.

The intuition behind the algorithm arising from the scoring of webpages is the existence of a mutually reinforcing relationship between two different types of pages: Firstly Authorities, which are commonly cited regarding certain topics, thus they are informative and tend to exhibit a large InDegree; and secondly Hubs, which cite many related Authorities, thus they are useful resources for finding Authorities and tend to exhibit a large OutDegree.

The Authority and Hub centralities generalize the InDegree and OutDegree centralities because they take into account broader neighborhood of a node. They are defined in terms of one another in a mutual recursion. The Authority score of a node w_i is computed as the sum of the scaled Hub scores of nodes that point to w_i. A Hub score of a node w_i is the sum of the scaled Authority scores of the nodes that w_i points to.

[22] Hyperlink Induced Topic Search

The Hub and Authority scores can be calculated with a two-step iterative algorithm that can be transformed to the computation of eigenvectors of Hub and Authority matrices or even viewed as the Singular Value Decomposition (SVD) of the weight matrix **G**. The first step of the iterative algorithm is the Authority update: The Authority score of the node w_i is proportional to the weighted sum of the Hub scores of each node that points to w_i. Hence the node is given the high Authority score by being strongly linked to the nodes that are recognized as important Hubs.

$$\lambda_A \mathbf{c}_A(w_j) = \sum_{i=1}^{N} G_{ij} \mathbf{c}_H(w_i) \tag{38}$$

The second step of the iterative algorithm is the Hub update: The Hub score of the node w_i is proportional to the weighted sum of the Authority scores of each node w_i points to. Hence the node is given the high Hub score by strongly linking to the nodes that are recognized as important Authorities.

$$\lambda_H \mathbf{c}_H(w_i) = \sum_{j=1}^{N} G_{ij} \mathbf{c}_A(w_j) \tag{39}$$

These two steps (38)and (39) are repeated. The scaling factors λ_A and λ_H ensure the convergence of the process. The equations (38) and (39) can be expressed in the matrix notation as

$$\begin{aligned} \lambda_A \mathbf{c}_A(G) &= \mathbf{G}^T \mathbf{c}_H(G) \\ \lambda_H \mathbf{c}_H(G) &= \mathbf{G} \mathbf{c}_A(G) \end{aligned} . \tag{40}$$

Combining these two equations together we receive the final matrix equation for the Authority scores and rather similar equation for the Hub scores.

$$\mathbf{G}^T \mathbf{G} \mathbf{c}_A(G) = \lambda \mathbf{c}_A(G) \tag{41}$$

$$\mathbf{G} \mathbf{G}^T \mathbf{c}_H(G) = \lambda \mathbf{c}_H(G) \tag{42}$$

The equations (41) and (42) imply that the searching for the Authority and Hub scores results in deriving of eigenvectors of the matrices $\mathbf{G}^T\mathbf{G}$ and $\mathbf{G}\mathbf{G}^T$. The matrix $\mathbf{G}^T\mathbf{G}$ is called the Authority matrix and $\mathbf{G}\mathbf{G}^T$ is called the Hub matrix. The constant λ is the same in both equations (41) and (42) and is related to scaling factors from (45) as $\lambda = \lambda_A \lambda_H$. It indicates that the Authority and Hub matrices share the same eigenvalues while their eigenvectors are generally different.

The problem can be also considered as SVD of the weight matrix **G** of the form

$$\mathbf{G} = \mathbf{C}_H \boldsymbol{\lambda}^{1/2} \mathbf{C}_A^T . \tag{43}$$

The matrix $\mathbf{C}_H$ consists of the right eigenvectors of the Hub matrix $\mathbf{G}\mathbf{G}^T$ while the matrix $\mathbf{C}_A$ consists of the right eigenvectors of the Authority matrix $\mathbf{G}^T\mathbf{G}$. The matrix $\boldsymbol{\lambda}$ is the diagonal matrix of common eigenvalues[23].

SVD of **G** also enables to decompose the weight matrix as the weighted sum of separable matrices[24] that are formed by Kronecker product of the eigenvectors of Hub and Authority matrices.

$$\mathbf{G} = \sum_{i=1}^{N} \lambda_i^{1/2} \mathbf{c}_{Hi} \otimes \mathbf{c}_{Ai}^T \tag{44}$$

Among the solutions of (41) and (42) the eigenvectors that are implied by the largest common eigenvalue are used as the Hub and Authority scores. Hence the vectors of the Hub and Authority scores provide the strongest separable matrix from the decomposition of **G** given in the formula (44).

4.5.5 PageRank

PageRank was developed at Stanford University as a part of the research project that focused on the development of a new search engine (Page et al., 1999). PageRank is similar to the Eigenvector centrality concept. PageRank is used by Google to rank webpages in their search results. PageRank can be regarded as a link analysis algorithm that assigns one numerical score to each element of a hyperlinked set of webpages with the purpose of measuring the relative importance of the webpage within the set. The algorithm may be applied to any directed social network.

PageRank of the node w_i is defined recursively and it depends on the number and on the magnitude of PageRank centralities of all nodes that link to w_i. If the node w_i is connected by many nodes with high PageRank, it receives a high PageRank as well. PageRank of the node w_i is the probability that represents the likelihood that a randomly selected connection on the walk through the network refers to w_i. In the network of webpages PageRank of 0.5 means there is a 50% chance that a person clicking on a random link will be directed to the document with the 0.5 PageRank.

To compute PageRanks of the nodes in the network G the weight matrix has to be transformed to a stochastic matrix where the elements of each row sum up to 1. Let us define a matrix **H** that is derived from the weight matrix **G** by dividing each element by the appropriate row sum[25].

$$\mathbf{H} = diag(\mathbf{G1})^{-1}\mathbf{G} \tag{45}$$

The function *diag*() returns a diagonal matrix from the vector in its argument. Note that the stochastic matrix **H** is obtained by the multiplication of **G** by the diagonal matrix of inverse OutDegrees. The matrix **H** is the transition probability matrix in Markov process in which we search for a stationary distribution of probabilities of visiting the nodes of G; the PageRank centrality of a node w_i is the probability of arriving at w_i after a large number of transitions.

[23] The square root of the common eigenvalue is called the singular value.
[24] A matrix is separable if it can be written as an outer (Kronecker) product of two vectors.
[25] Assuming that all weights are non-negative and the row sums are positive. The zero row sum will be addressed and adjusted later in the text.

$$\mathbf{c}_{PR}(G)^T \mathbf{H} = \mathbf{c}_{PR}(G)^T \tag{46}$$

The equation shows that PageRank is a variant of the left-handed Eigenvector centrality for a modified weight matrix. It is ensured by Perron–Frobenius theorem that the largest eigenvalue of a stochastic matrix equals one; hence the eigenvalue or the scaling factor is not present in the equation (46).

The formula (46) is only the simplified version of PageRank computation. The calculation of PageRank is commonly adjusted for random transitions that are artificially added to the network. Therefore in the random walk through the network we can jump from the current node to any node with the constant probability $(1\text{-}b)/N$[26], where N is the number of nodes and b is a selected constant between zero and one[27]. The idea behind is that when browsing the Internet a user can either follow the hyperlinks on pages or to write a completely new address to his browser. Such adjustment also solves the problems of sinks that are the nodes without any outgoing connection. In Markov theory the sinks are absorbing states; it is impossible to leave the absorbing state. Such sinks would attract all PageRanks on themselves from non-sink nodes if the network was not enriched by the artificial random transitions.

The artificial transitions are added to all nodes in the network regardless of the fact the node is the sink or not. The probability of jumping to a randomly selected node in the network instead of using the original transitions is usually set to 0.25 or $b = 0.85$. Hence we have to solve the following matrix equation to compute the vector of PageRanks $\mathbf{c}_{PR}(G)$ which represents the stationary probabilities of visiting the nodes on a random walk through the network.

$$(1-b)\frac{1}{N}\mathbf{1}^T + b\mathbf{c}_{PR}(G)^T \mathbf{H} = \mathbf{c}_{PR}(G)^T \tag{47}$$

The PageRank vector that serves as the proposed document representation is then

$$\mathbf{c}_{PR}(G)^T = (1-b)\frac{1}{N}\mathbf{1}^T(\mathbf{I} - b\mathbf{H})^{-1}, \tag{48}$$

where $\mathbf{I}$ is the identity matrix. The PageRank vector $\mathbf{c}_{PR}(G)$ can be also derived as the vector of stationary probabilities of visiting network nodes in Markov process similarly as in the formula (46) but using a modified transition probabilities. If $\mathbf{c}_{PR}(G)$ include probabilities, it sums up to one. Using the matrix notation it holds

$$\mathbf{c}_{PR}(G)^T \mathbf{1} = 1. \tag{49}$$

Multiplying the first summand in (47) by (49) we can define the adjusted transition probabilities matrix as

$$\overline{\mathbf{H}} = (1-b)\frac{1}{N}\mathbf{1}\mathbf{1}^T + b\mathbf{H}. \tag{50}$$

[26] It assures that the adjusted Outdegree is always positive.
[27] b is the probability of preferring the original transitions over the artificially added ones.

It is also the stochastic matrix representing the process without sinks. This property is ensured by the first summand in (50) where $\mathbf{1}\mathbf{1}^T$ represents the square matrix of ones. Hence the PageRank document representation can be derived as the left-hand eigenvector of the matrix $\overline{\mathbf{H}}$ assigned to the eigenvalue that equals one.

$$\mathbf{c}_{PR}(G)^T \overline{\mathbf{H}} = \mathbf{c}_{PR}(G)^T \tag{51}$$

4.5.6 Eigenvector

Eigenvector centrality, regarded as a ranking measure, is a remarkably old method (Seeley, 1949; Leontief, 1941). The Eigenvector centrality is the measure of the broader influence of a node in a network. It assigns relative scores to all nodes in the network based on the concept that the connections to high-scoring nodes contribute more to the score of the node than the connections to low-scoring nodes. It is a natural extension of the InDegree or OutDegree centralities. For example the InDegree centrality awards weight centrality points for every link a node receives. But not all vertices are equivalent: some are more relevant than others; endorsements from important nodes count more. The Eigenvector centrality makes a node important if it is linked to other important nodes. The Eigenvector centrality differs from the InDegree and OutDegree centralities: a node with strong ingoing or outgoing connections does not necessarily have the high Eigenvector centrality because it might be that all linkers have the low Eigenvector centrality. Moreover, a node with the high Eigenvector centrality is not necessarily strongly linked because the node might have few but important linkers.

The Eigenvector centrality score of a vertex w_i can be defined as a solution of the following equation. The constant λ serves as the proportional factor that ensures that the equation has a finite solution.

$$\lambda c_E(w_j) = \sum_{i=1}^{N} G_{ij} c_E(w_i) \tag{52}$$

With a small rearrangement the equation can be rewritten in the matrix notation as the eigenvalue and eigenvector equation.

$$\mathbf{c}_E^T(G)\mathbf{G} = \lambda \mathbf{c}_E^T(G) \tag{53}$$

The centrality vector $\mathbf{c}_E$ that serves as the proposed document representation is the left-hand eigenvector of the weight matrix **G** associated with the eigenvalue λ. The solution is not unique; the matrix **G** provides generally more eigenvalues and eigenvectors depending on its rank. It is wise to choose λ as the largest absolute eigenvalue of matrix **G** and $\mathbf{c}_E$ is then the eigenvector associated with this eigenvalue. By virtue of Perron-Frobenius theorem, this choice guarantees that the associated eigenvector consists of real positive entries. This theorem assumes that the matrix **G** is irreducible. For the context network *G* it means that *G* is strongly connected. A network is said to be strongly connected if every vertex is reachable from every other vertex[28].

Note that the process how the context network *G* is derived does not guarantee that the weight matrix **G** is irreducible. The matrix **G** tends to be irreducible especially when the document *d*

[28] A special case of the irreducible matrix is a matrix where all entries are positive.

is short and the number of topics K is large. In the later described experiments the unreachable nodes receive zero Eigenvector centrality similarly as for InDegree or OutDegree.

The idea of the Eigenvector centrality formulated in (52) implies the left-handed eigenvector as a solution of (53). In such approach the Eigenvector centrality is a generalization of InDegree. The Eigenvector centrality may be also a generalization of OutDegree. If the centrality score of the node w_i is proportional to centrality scores of nodes connected by outgoing ties from w_i, it leads to the following equation for centralities $c_E(w_i)$.

$$\lambda c_E(w_i) = \sum_{j=1}^{N} G_{ij} c_E(w_j) \tag{54}$$

The equation can be again rewritten in the matrix notation as the eigenvalue and eigenvector equation. The centrality vector $\mathbf{c}_E$ is then the right-hand eigenvector of the weight matrix $\mathbf{G}$ associated with the eigenvalue λ.

$$\mathbf{G}\mathbf{c}_E(G) = \lambda \mathbf{c}_E(G) \tag{55}$$

The right-handed and left-handed eigenvectors are not generally the same for asymmetric matrices. The right-handed eigenvector of $\mathbf{G}$ can be derived as the left-handed eigenvector of $\mathbf{G}^T$ and vice versa. The right-handed eigenvectors are used more commonly and also most implementations of the Eigenvector centrality offer right-handed Eigenvector only. Hence in the experiments only the right-handed Eigenvector representation is explored.

4.5.7 Closeness

The Closeness centrality of a node was developed to reflect how close is the node to other nodes (Sabidussi, 1966). It expresses how effectively the node can interact with other nodes. Such interaction is influenced by the number of mediators and by the proximities between the directly connected mediators. Closeness of the node is a function of its distance to all other nodes in the network[29]. The closeness centrality is computed as the inverse of the sum of the distances between the node and all other nodes.

$$c_C(w_i) = \frac{1}{\sum_{j=1}^{N} d(w_i, w_j)} \tag{56}$$

The Closeness centrality can never be zero; it is always non-negative. The upper bound is not constrained in the weighted network. The higher values of Closeness imply that the node is tightly connected with the others. The inverse of Closeness is called Farness; it is just the sum of the distances to all other nodes.

Note that Closeness cannot be computed for isolated nodes. If there is the isolated node in the context network, zero Closeness centrality is assigned to it. The isolated node can occur in the context network when the topic that is represented by the isolated node is completely missing in a text.

[29] The distance means the length of the shortest path that connects the vertices.

4.5.8 Betweenness

The notion of the Betweenness centrality can be found in sociology (Freeman, 1977). The Betweenness concept of centrality of a node concerns how the node controls or mediates the paths between pairs of other nodes that are not directly connected. The Betweenness centrality measures the extent to which the node lies on the shortest path[30] between pairs of other nodes. Generally Betweenness is an indicator of the control over the information exchange within a network. The more often the node is located on the shortest paths between numerous node pairs, the higher is its potential to control the network interactions.

To compute the Betweenness centrality the function $s_{ij}(w_k)$ has to be introduced. This function returns the number of the shortest paths between nodes w_i and w_j that the node w_k is intersects. Summing $s_{ij}(w_k)$ across all pairs of nodes not including the node w_k, we receive the number of the shortest paths that are controlled by the node w_k[31].

$$c_B(w_k) = \sum_{i=1}^{N} \sum_{j=i+1}^{N} s_{ij}(w_k) \tag{57}$$

Betweenness is zero when the node w_k falls on no shortest path for all the pairs among the other nodes. It reaches the maximum value $(N\text{-}1)(N\text{-}2)$ in directed networks with N nodes when the node w_k falls on every shortest path for all node pairs, assuming that the only one shortest path exists between each pair.

[30] The shortest path is sometimes referred as geodesic path.

[31] Alternatively the Betweenness centrality can be defined relatively as the proportion of the shortest paths where the node w_i is present.

5 Theoretical evaluation

5.1 Goals of theoretical evaluation

Variability of document vectors within a collection is essential for document discrimination in standard text mining tasks. Diversity of document vectors is exploited by mining models to recognize similar and dissimilar documents for retrieval, classification or clustering. If the document diversity is inhibited by extracting unappropriated features, the mining models cannot discriminate documents well enough and their quality is lower than it could be. A context in which words or some higher linguistic entities appear in documents can be an important source of the document diversity that may be exploited by the text mining models if it is propagated to document vectors. The goals of this evaluation of properties of the proposed document representations presented in following chapters are:

- Find or approximate the probability distribution of the proposed document representations.
- Quantify the reduction of variability when context networks are substituted by centrality vectors.
- Compare the above variability reduction with a standard document representation that does not comprise any contextual information.

The presented evaluation does not include an assessment of the contribution of the proposed representation to the quality of text mining models because there are other significant factors that influence the model quality in a processing pipeline including modeling algorithm and its parameters. Hence comparisons of a performance of the proposed representation with a performance of a standard representation in selected text mining tasks are left to the experimental part of the thesis.

5.2 Properties of proposed document representations

Let us have the vocabulary $V=\{w_1,w_2,\ldots,w_N\}$ of the size $|V|=N$. The vocabulary terms w_i can be words or higher entities presented in texts[32]. A document d is a sequence of the vocabulary terms in the form of $w_{(1)}w_{(2)}w_{(3)}\ldots$ of the length L_d. The bracketed indexes denote the order of terms in the sequence. The non-vocabulary terms are omitted from the sequence.

The neighborhood of terms in a document cannot be ignored because the order in which terms appear in the sequence offers additional information about the document and it enables to distinguish the document with higher accuracy then the sole counts of terms[33]. Hence let us assume that each document in a collection is a container of groups of terms of the length of n that are referred as n-grams. These n-grams are products of a generative process. Each document can be described by a set of n-gram probabilities that constitutes the hidden property of the document. While the n-gram probabilities are unobservable, we can observe counts of n-grams within the document that are generated regarding to their hidden probabilities. The counts are the realization of the multinomial random variable. The n-gram probabilities can be estimated from these counts.

[32] They can be hidden or observed.

[33] We assume that the documents in a collection are grammatically correct. Instead of estimation of probability of a document or even its linguistic correctness which is the main goal of language models we want to propose an informative representation of documents for a further processing.

The n-gram probabilities would serve as the perfect representation of the document. Apart from the fact that n-gram probabilities are not directly observable, they suffer from another imperfection: There are N^n possible n-grams regarding to the vocabulary of N terms in our model, so the dimensionality of such a representation is unacceptably high. Hence we firstly organize the n-gram probabilities and counts into a well arranged matrix and then we propose the reduction of its size.

Similarly to traditional language modeling the n-gram probabilities compose a matrix **Q** of the size $N^{(n-1)}\times N$. The rows of **Q** represent all potential subsequences of terms of the length of (n-1). The columns represent the vocabulary terms. Each cell of **Q** includes the probability of the whole n-gram; the column term is the last term of the n-gram. Such a layout of the probabilities enables to switch easily from the unconditional to the conditional probabilities. The conditional probabilities are used in popular generative language n-gram models that exploit Markov chains.

In generative Markov process each term is generated regarding its preceding terms. More precisely, the probability of the next term is conditioned by its previous terms. In the standard n-gram language model (n-1) preceding terms influence the probability of the current term. The conditional probabilities form the matrix **P** that can be easily derived from the matrix **Q**. Similarly as for **Q**, the rows of **P** stand for preceding (n-1)-grams and the columns represent the final vocabulary terms of n-grams. The size of **P** is the same as the size of **Q** ($N^{(n-1)}\times N$). The item p_{ij} of **P** is the probability of the generation of the term w_j following the i^{th} term subsequence[34]. Obviously the row sums of **P** equal to one. Each document in the collection is generated using its matrix **Q** or **P** that is unknown. However, **Q** or **P** can be estimated from an observed transition matrix **T**. It is the matrix of the same size as **P** and includes frequencies of n-grams or transitions in the document. The maximum-likelihood estimates of the unconditioned probabilities take the form

$$q_{ij} = \frac{t_{ij}}{\sum_{k=1}^{M}\sum_{l=1}^{N} t_{kl}}. \tag{58}$$

The conditional probabilities can be derived from the unconditional ones as

$$p_{ij} = \frac{q_{ij}}{\sum_{l=1}^{N} q_{il}}. \tag{59}$$

The reverse derivation (unconditional **Q** from conditional **P**) is impossible. The reason is that the marginal row probabilities are lost when moving from **Q** to **P**. Hence the matrix **Q** is more informative and it will serve as the best but hidden representation of each document. The best observable representation of a document is then its matrix **T**.

[34] The alternative and equivalent approach is to introduce a square matrix **P** of the size $N^{(n-1)}\text{x}N^{(n-1)}$ including the conditional probabilities. The columns would represent the same subsequences of terms as the rows represent and **P** would be matrix of transition probabilities among (n-1)-grams. Obviously the most of these probabilities would equal zero, positive probabilities would appear only for (n-1)-grams with common ends and starts.

5.2.1 Representation distributions

The matrix **T** can be also considered as a realization of generative Markov process based on the matrix **P**. **T** is the matrix of random non-negative integers. The sum of all cells of **T** is equal to the document length L[35]. The distribution of the cells of **T** is multinomial with the parameters L and **Q**[36].

$$\begin{aligned} &\mathbf{T} \sim Mu(L, \mathbf{Q}) \\ &p(\mathbf{T}; L, \mathbf{Q}) = \frac{L!}{\prod_{i=1}^{M}\prod_{j=1}^{N} t_{ij}!} \prod_{i=1}^{M}\prod_{j=1}^{N} q_{ij}^{\,t_{ij}} \end{aligned} \tag{60}$$

If the matrix **T** comes from the multinomial distribution with the parameters L and **Q**, then each cell t_{ij} of **T** comes from the binomial distribution with the parameters L and q_{ij}.

$$\begin{aligned} &t_{ij} \sim Bi(L, q_{ij}) \\ &p(t_{ij}; L, q_{ij}) = \binom{L}{t_{ij}} q_{ij}^{t_{ij}} (1-q_{ij})^{L-t_{ij}} \end{aligned} \tag{61}$$

The counts of n-grams t_{ij} are not independent because they constitute the multinomial random matrix and they have to fulfill the condition

$$\sum_{i=1}^{M}\sum_{j=1}^{N} t_{ij} = L\,. \tag{62}$$

The properties of the random variables t_{ij} can be then summarized as

$$\begin{aligned} &t_{ij} \sim Bi(L, q_{ij}) \\ &E(t_{ij}) = Lq_{ij} \\ &\mathrm{var}(t_{ij}) = Lq_{ij}(1-q_{ij}) \\ &\mathrm{cov}(t_{ij}, t_{kl}) = -Lq_{ij}q_{kl}, i \neq k \vee j \neq l \end{aligned} \quad . \tag{63}$$

To derive a simplified representation of a document, the matrix **T** has to be transformed. We start with linear transformations. The linear transformations can be expressed by the multiplication of **T** by an appropriate matrix from the left or/and from the right. Regarding a general linear combination of binomial random variables, the resultant distribution is too difficult to derive [37]. Hence it is worth approximating the binomial distributions by distributions that can be easily combined. The binomial distribution of t_{ij} can be approximated

[35] Regarding the start of the document only L-(n-1) transitions can be observed. If the length of document is significantly larger than n, this difference can be neglected. Alternatively (n-1) artificial terms can be added before the start of the document and the vocabulary would be extended with such artificial term as well.

[36] The categories of this random categorical variable are not arranged into a vector but into the matrix. However it does not influence its multinomial distribution.

[37] The sum of independent binomial random variables has Poisson binomial distribution. But we deal with the dependent random variables.

either by Poisson or by the normal distribution. Concerning the covariances among n-gram counts, the better choice is the normal distribution because the matrix **T** has then the multivariate normal distribution that is better tractable for transformations than the multinomial Poisson distribution.

To correctly express the covariance matrix of the multivariate normal distribution of **T**, the matrix **T** of frequencies and the matrix **Q** of probabilities should be rewritten to vectors using the operation called vectorization. The vectorization of a matrix converts the matrix into a column vector that is obtained by stacking the columns of the original matrix on the top of one another. For example, the vectorization of the $M \times N$ matrix **T**, denoted by vec(**T**), is a $MN \times 1$ column vector $(t_{11},t_{21},\ldots,t_{M1},t_{12},t_{22},\ldots,t_{M2},t_{13},\ldots,t_{MN})^{\mathrm{T}}$.

Regarding the original covariances (63) among the binomial random variables t_{ij}, the matrix **T** can be approximated by multivariate normal distribution of the form

$$\begin{aligned} &\mathrm{vec}(\mathbf{T}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &\boldsymbol{\mu} = L\,\mathrm{vec}(\mathbf{Q}) \\ &\boldsymbol{\Sigma} = L\left(\mathrm{diag}(\mathrm{vec}(\mathbf{Q})) - \mathrm{vec}(\mathbf{Q})\mathrm{vec}(\mathbf{Q})^T\right) \\ &f(\mathrm{vec}(\mathbf{T})) = \frac{1}{(2\pi)^{\frac{MN}{2}}\det(\boldsymbol{\Sigma})^{\frac{1}{2}}}\exp\left(-\tfrac{1}{2}(\mathrm{vec}(\mathbf{T})-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathrm{vec}(\mathbf{T})-\boldsymbol{\mu})\right) \end{aligned} \quad , \qquad (64)$$

where function *diag()* forms a diagonal matrix from the given vector.

The covariance matrix $\sum$ of the size $MN \times MN$ cannot be rewritten as Kronecker product of two matrices of sizes $M \times M$ and $N \times N$ respectively, hence the matrix **T** cannot be regarded as a realization of the random matrix with the matrix normal distribution[38]. Even though **T** is the matrix, it is necessary to vectorize it to express correctly its distribution.

Because we want to obtain a new representation of a document by simplification of its matrix **T**, let us investigate the properties of some basic transformations of **T**. Firstly, the transformation of **T** to the proposed context network **G** can be realized by the multiplication by the appropriate matrix from the left. Then the extraction of some basic selected centrality vectors from the context network **G** can be performed by the multiplication of the network by a matrix from the right or from the left.[39]

If **T** comes from the approximated multivariate normal distribution, its linear transformation also comes from the multivariate normal distribution with the transformed vector of means and the transformed covariance matrix. To express the transformed parameters, the following formula for the vectorized matrix product should be taken into the consideration.

$$\mathrm{vec}(\mathbf{LTR}) = (\mathbf{R}^T \otimes \mathbf{L})\mathrm{vec}(\mathbf{T}) \qquad (65)$$

[38] The matrix normal distribution is a generalization of the multivariate normal distribution to matrix-valued random variable. A matrix **X** of the size $m \times n$ comes from matrix normal distribution $MN(\mathbf{M}, \mathbf{U}, \mathbf{V})$ if and only if $\mathrm{vec}(\mathbf{X}) \sim N(\mathrm{vec}(\mathbf{M}), \mathbf{U} \otimes \mathbf{V})$. The symbol $\otimes$ stands for Kronecker product. The matrix **M** of the size $m \times n$ is the matrix of means and there are two covariance matrices **U** and **V** in the matrix normal distribution of the sizes $m \times m$ and $n \times n$ respectively. They separately express the covariance among rows of **X** and columns of **X** respectively.

[39] Namely Degree, InDegree and OutDegree can be obtained as a linear combination of the co-occurrence frequencies. Other considered centralities require non-linear transformations of the matrix **T**.

The symbol $\otimes$ is Kronecker product of matrices[40]. If the multiplication is executed from the right or from the left only, then the matrix **L** or the matrix **R** from the above formula are the identity matrices. For the linear transformation of any random vector **t** holds

$$E(\mathbf{t}) = \boldsymbol{\mu}, \mathrm{var}(\mathbf{t}) = \boldsymbol{\Sigma} \Rightarrow E(\mathbf{At}) = \mathbf{A}\boldsymbol{\mu}, \mathrm{var}(\mathbf{At}) = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T. \tag{66}$$

Then exploiting (65) the transformed distribution of the product **LTR** has the form

$$\mathrm{vec}(\mathbf{LTR}) \sim N\left((\mathbf{R}^T \otimes \mathbf{L})\boldsymbol{\mu}, (\mathbf{R}^T \otimes \mathbf{L})\boldsymbol{\Sigma}(\mathbf{R}^T \otimes \mathbf{L})^T\right). \tag{67}$$

Using the parameters of the transformed distribution from the formula (64) and formula (65), the distribution can be rewritten as

$$\mathrm{vec}(\mathbf{LTR}) \sim N\left(L\,\mathrm{vec}(\mathbf{LQR}), L\left((\mathbf{R}^T \otimes \mathbf{L})\mathrm{diag}(\mathrm{vec}(\mathbf{Q}))(\mathbf{R}^T \otimes \mathbf{L})^T - \mathrm{vec}(\mathbf{LQR})\mathrm{vec}(\mathbf{LQR})^T\right)\right) \tag{68}$$

Note that comparing with the original distribution of **T** in (64) the first term of the covariance matrix lost the form of a diagonal matrix after the transformation while the second term of the covariance matrix and the mean vector are transformed similarly as the random matrix **T**. Let us illustrate the above mentioned linear transformations of **T** by several examples.

5.3 Examples of representation distributions

Example 1: The transformation of the n-gram representation to the (n-1)-gram representation.

The matrix **T** include the counts of transitions from (n-1)-grams to vocabulary terms. Then if we sum all the transitions from the same (n-1)-grams, we receive (n-1)-grams counts. We have to compute row sums of **T**. The row-sum operation can be realized by the multiplication by a column vector of ones of the size $N \times 1$ from the right. The left matrix remains the identity matrix **I** of the size $M \times M$[41]. The transformed (n-1)-gram representation has the following multivariate normal distribution.

$$\mathrm{vec}(\mathbf{T1}) \sim N\left(L\,\mathrm{vec}(\mathbf{Q1}), L\left((\mathbf{1}^T \otimes \mathbf{I})\mathrm{diag}(\mathrm{vec}(\mathbf{Q}))(\mathbf{1}^T \otimes \mathbf{I})^T - \mathrm{vec}(\mathbf{Q1})\mathrm{vec}(\mathbf{Q1})^T\right)\right) \tag{69}$$

The first term of the transformed covariance matrix can be simplified using the formula (65) and the vectorizations of **Q1** and **T1** can be omitted.

$$\mathbf{T1} \sim N\left(L\mathbf{Q1}, L\left(\mathrm{diag}(\mathbf{Q1}) - \mathbf{Q1}(\mathbf{Q1})^T\right)\right) \tag{70}$$

Regarding to the mean and to the covariance of the transformed matrix, we can conclude that the (n-1)-gram representation has the multivariate normal distribution that approximates a multinomial distribution with the parameters L and **Q1**.

---

[40] If **A** is a $m \times n$ matrix and **B** is a $p \times q$ matrix, then Kronecker product $\mathbf{A} \otimes \mathbf{B}$ is the $mp \times nq$ block matrix of products of all pairs of values from **A** and **B**.

[41] The (n-1)-gram representation can be also derived as the final state of stationary Markov proces discribed by transition matrix P extended to transitions from (n-1)-grams to (n-1)-grams.

$$\mathbf{T1} \to Mu\left(L, \mathbf{Q1}\right) \tag{71}$$

Example 2: The transformation of the n-gram representation to the term (1-gram) representation.

The matrix **T** includes the counts of transitions from (n-1)-grams to vocabulary terms. Then if we sum all the transitions to the same terms, we receive the term counts. We have to compute column sums of **T**. The column-sum operation can be realized by the multiplication by a row vector of ones of the size $1\times M$ from the left. The right matrix remains the identity matrix **I** of the size $N\times N$. The transformed 1-gram representation has the following multivariate normal distribution

$$\mathrm{vec}(\mathbf{1}^T\mathbf{T}) \sim N\left(L\,\mathrm{vec}\left(\mathbf{1}^T\mathbf{Q}\right), L\left(\left(\mathbf{I}\otimes\mathbf{1}^T\right)\mathrm{diag}\left(\mathrm{vec}\left(\mathbf{Q}\right)\right)\left(\mathbf{I}\otimes\mathbf{1}^T\right)^T - \mathrm{vec}\left(\mathbf{1}^T\mathbf{Q}\right)\mathrm{vec}\left(\mathbf{1}^T\mathbf{Q}\right)^T\right)\right) \tag{72}$$

The first term of the transformed covariance matrix can be simplified using the formula (65) and the vectorizations of $\mathbf{1}^T\mathbf{Q}$ and $\mathbf{1}^T\mathbf{T}$ can be substituted by transpositions.

$$\mathbf{T}^T\mathbf{1} \sim N\left(L\mathbf{Q}^T\mathbf{1}, L\left(\mathrm{diag}\left(\mathbf{Q}^T\mathbf{1}\right) - \mathbf{Q}^T\mathbf{1}\left(\mathbf{Q}^T\mathbf{1}\right)^T\right)\right) \tag{73}$$

Regarding the mean and the covariance of the transformed representation, we can conclude that the term representation has the multivariate normal distribution that approximates a multinomial distribution with the parameters L and $\mathbf{Q}^T\mathbf{1}$.

$$\mathbf{T}^T\mathbf{1} \to Mu\left(L, \mathbf{Q}^T\mathbf{1}\right) \tag{74}$$

Note that the term representation is not influenced by the term order or by the context. Representations that take into the account the context will be later compared with this representation.

Example 3: The transformation of the n-gram representation to the context network representation.

Let us consider the context of (n-1) foregoing terms. While the matrix **T** includes the counts of transitions from (n-1)-grams to vocabulary terms, the context network **G** should include the counts of transitions from the vocabulary terms to the vocabulary terms. The starting term in a transition must be included in the foregoing (n-1)-gram, the exact position of the starting term in the foregoing (n-1)-gram does not matter. Hence the transformation from the transition matrix **T** of the size $M\times N$ to the context network **G** of the size $N\times N$ can be realized by the multiplication of **T** by a fixed matrix **H** of the size $N\times M$ from the left. The columns of **H** represent (n-1)-grams while the rows stand for vocabulary terms. The integer values in **H** are counts of terms in (n-1)-grams. The matrix **H** can be decomposed into the sum of (n-1) Kronecker products of (n-1) factors. One of the factors is the identity matrix **I** of the size $N\times N$ and the other factors are transposed vectors of ones of the size $1\times N$. For example, if the size of the vocabulary $|V|=N=3$, the matrix **H** can be decomposed as

$$\mathbf{H} = \mathbf{1}^T\otimes\mathbf{1}^T\otimes\mathbf{I} + \mathbf{1}^T\otimes\mathbf{I}\otimes\mathbf{1}^T + \mathbf{I}\otimes\mathbf{1}^T\otimes\mathbf{1}^T. \tag{75}$$

Using the formula (68) the transformed context network representation has the following multivariate normal distribution

$$\begin{aligned} \mathrm{vec}(\mathbf{G}) &= \mathrm{vec}(\mathbf{HT}) \\ &\sim N\left(L\,\mathrm{vec}(\mathbf{HQ}), L\left((\mathbf{I}\otimes\mathbf{H})\mathrm{diag}(\mathrm{vec}(\mathbf{Q}))(\mathbf{I}\otimes\mathbf{H})^T - \mathrm{vec}(\mathbf{HQ})\mathrm{vec}(\mathbf{HQ})^T\right)\right) \end{aligned} \quad (76)$$

While the mean vector is transformed the same way as the transition matrix **T**, the first term of the covariance matrix cannot be reasonably modified. Hence we can conclude that the context network distribution is not similar to the multinomial or other common distribution. It has to be approximated by the multivariate normal distribution with the parameters from the above formula.

Example 4: The transformation of the n-gram representation to the InDegree centrality representation.

The InDegree centrality representation of a document is derived as a vector of InDegree centralities of the vertices from the context network *G*. The vertices of *G* represent vocabulary terms. The InDegree centrality of a vertex equals the sum of weights of its incoming edges. Hence each InDegree can be computed as the column sum in the weight matrix **G**. The InDegree representation can be derived from the context network by multiplication of the weight matrix **G** by the row vector of ones from the left. From the previous example we know that the matrix **G** is derived from the transition matrix **T** by multiplication by the rectangle matrix **H** from the left. Altogether, we have to investigate the properties of the matrix product $\mathbf{1}^T\mathbf{HT}$.

Any column sum of the matrix **H** equals the length of the context *K* which is set to (*n*-1). Hence the product $\mathbf{1}^T\mathbf{H}$ equals the constant vector filled by values (*n*-1). We can simplify the InDegree representation as

$$\mathbf{1}^T\mathbf{HT} = (n-1)\mathbf{1}^T\mathbf{T}. \quad (77)$$

The product $\mathbf{1}^T\mathbf{T}$ forms the term representation from example 2; the InDegree representation is the same as the term representation only multiplied by the constant (*n*-1). Similarly as in the second example we can conclude

$$vec(\mathbf{1}^T\mathbf{HT}) = (n-1)\mathbf{T}^T\mathbf{1} \sim N\left((n-1)L\mathbf{Q}^T\mathbf{1}, (n-1)^2 L\left(\mathrm{diag}(\mathbf{Q}^T\mathbf{1}) - \mathbf{Q}^T\mathbf{1}(\mathbf{Q}^T\mathbf{1})^T\right)\right) \quad (78)$$

Regarding the transformed mean and the transformed covariance, we can conclude that it is not the multivariate normal distribution that directly approximates the multinomial distribution, but it is the approximation of the multinomial distribution with the parameters *L* and $\mathbf{Q}^T\mathbf{1}$ multiplied by factor (*n*-1)[42].

$$\frac{\mathbf{T}^T\mathbf{H}^T\mathbf{1}}{n-1} \to Mu(L, \mathbf{Q}^T\mathbf{1}) \quad (79)$$

Example 5: The transformation of the n-gram representation to the OutDegree centrality representation.

[42] It does *not* holt that $\mathbf{X} \sim Mu(L, \mathbf{p}) \Rightarrow k\mathbf{X} \sim Mu(kL, \mathbf{p})$.

The OutDegree centrality representation of a document is derived as a vector of OutDegree centralities of the vertices in the context network *G*. The OutDegree centrality of a vertex equals the sum of weights of its outgoing edges. Hence OutDegree can be computed as the row sum in **G**. The OutDegree representation can be derived by the multiplication of the weight matrix **G** by a column vector of ones from the right. Together with the derivation **G** from **T** we have to investigate the properties of the matrix product **HT1**.

Using the formula (68) the OutDegree representation has the following multivariate normal distribution

$$\mathrm{vec}(\mathbf{HT1}) \sim N\left(L\,\mathrm{vec}(\mathbf{HQ1}), L\left(\left(\mathbf{1}^T \otimes \mathbf{H}\right)\mathrm{diag}\left(\mathrm{vec}(\mathbf{Q})\right)\left(\mathbf{1}^T \otimes \mathbf{H}\right)^T - \mathrm{vec}(\mathbf{HQ1})\mathrm{vec}(\mathbf{HQ1})^T\right)\right. \quad (80)$$

Unfortunately, the first term of the covariance cannot be reasonably modified. We can conclude that the distribution of the OutDegree representation is not similar to the multinomial or other common distribution. It has to be regarded as multivariate normal distribution with the parameters from the above formula.

Example 6: The transformation of the n-gram representation to the Degree centrality representation.

The Degree centrality representation of a document is derived as the vector of Degree centralities of the vertices in the context network *G*. The Degree centrality of a vertex equals the sum of weights of all its ingoing and outgoing edges. Degree can be computed as InDegree plus OutDegree, hence we can put together the results from the previous examples. In this sum InDegree from (77) has to be transposed to the column vector or OutDegree to the row vector:

$$\left(\mathbf{1}^T\mathbf{HT}\right)^T + \mathbf{HT1} = \left(\left(\mathbf{HT}\right)^T + \mathbf{HT}\right)\mathbf{1} \quad (81)$$

Note from the above formula that Degree is computed as the row sum of the symmetrized (undirected) context network[43]. To investigate the properties of the Degree representation, we need to rewrite the above formula using vectorization of the transition matrix **T**. Using formula (65) we can conclude

$$\mathrm{vec}\left(\left(\left(\mathbf{HT}\right)^T + \mathbf{HT}\right)\mathbf{1}\right) = \left(\mathbf{1}^T\mathbf{H} \otimes \mathbf{I}\right)\mathrm{vec}\left(\mathbf{T}^T\right) + \left(\mathbf{1}^T \otimes \mathbf{H}\right)\mathrm{vec}\left(\mathbf{T}\right). \quad (82)$$

Hence we have to investigate the properties of the sum of two normally distributed random vectors $(\mathbf{HT})^T\mathbf{1}$ and **HT1**. Note that vec(**T**) and vec($\mathbf{T}^T$) are the different vectors, but they are not independent. The distribution of vec(**T**) is described in (64) and similarly the distribution of vec($\mathbf{T}^T$) is

$$\mathrm{vec}(\mathbf{T}^T) \sim N\left(L\,\mathrm{vec}(\mathbf{Q}^T), L\left(\mathrm{diag}\left(\mathrm{vec}\left(\mathbf{Q}^T\right)\right) - \mathrm{vec}\left(\mathbf{Q}^T\right)\mathrm{vec}\left(\mathbf{Q}^T\right)^T\right)\right). \quad (83)$$

The sum of two normal distributions has also the normal distribution with the sum of original means. However, the variance matrix is not equal to the sum of original variance matrices

[43] The row sums of a symmetrized matrix equal the column sums.

because we sum dependent vectors. Generally, for two random vectors **X** and **Y** of the same size it holds

$$\mathrm{var}(\mathbf{X}+\mathbf{Y}) = \mathrm{var}(\mathbf{X}) + \mathrm{var}(\mathbf{Y}) + \mathrm{cov}(\mathbf{X},\mathbf{Y}) + \mathrm{cov}(\mathbf{Y},\mathbf{X}). \tag{84}$$

Using this formula we can derive the final parameters of the distribution of the Degree representation, but it results in rather complicated and uninformative formula for the variance matrix. The resultant normal distribution is not similar to the multinomial or other common distribution.

It is obvious from the previous examples that the approximation of the multinomial distribution of the transitions **T** by the normal distribution enables to investigate the properties of some selected representations obtained as a linear transformation of the matrix **T**. Regarding the centrality representations, it was shown that the InDegree representation is similar to 1-gram representation and comes from the multinomial distribution with the parameters L and $\mathbf{Q}^{\mathrm{T}}\mathbf{1}$ multiplied by the length of the context K. The distribution of the OutDegree and Degree representations can be approximated by the normal distributions as well, but the dependence among transition counts of **T** causes that the resultant covariance matrices are not similar to the multinomial distributions and their forms can be rather complicated.

More promising centrality representations include the ones where the centrality of a vertex in the context network does not depend on its own connections only, but on properties of its neighbor vertices as well. Unfortunately, such centralities cannot be derived as a linear transformation of the random matrix **T**, but their computation often implies an iterative process. They exploit the eigenvectors derivation or the investigation of paths in the context network. Hence the exact distributions for such centralities cannot be derived; the distributions can be approximated by simulations.

5.4 Relationship with original representation

Now let us investigate how the proposed representations maintain or reduce the original variability of random transition matrix **T**. Remember that we assume that the observed documents are products of the n-gram generative model where the length of the context, which influences the appearance of a term, is constant. Each document is then fully described by its matrix of n-gram probabilities **Q.** The proposed representations reduce the number of entries of **Q** and transform it into a vector. Hence we should investigate the relationship between the proposed vector representation and **Q**. Even though the n-gram probabilities are organized in the matrix **Q**, they should be regarded as the random vector that results from the vectorization of **Q**. Therefore we will investigate the relation between the pair of random vectors: the proposed representation and the vectorization of **Q**.

In the following paragraphs two approaches to the expression of the common variability are exploited. Firstly, the variability of nominal random variables is measured by the entropy and the relation between two representations is measured by their mutual information. Secondly, the original multinomial vectors are approximated by normal vectors again and the variability is expressed using covariance matrices. The relation between two representations is then measured by the variability reduction when comparing conditional and unconditional distributions.

5.4.1 Mutual information approach

The variability of a discrete random variable X that can take values from the set $\{x_1,x_2,\ldots,x_C\}$ with probabilities $\{p(x_1),p(x_2),\ldots,p(x_C)\}$ can be expressed by the notion of entropy as

$$H(X) = -\sum_{i=1}^{C} p(x_i)\ln(p(x_i)) \tag{85}$$

The entropy is non-negative, its upper-bound depends on the number of possible entries as $ln(C)$. Using the natural logarithm it is measured in nats[44]. To compare two distributions using the entropy approach, we can compare their entropy or we can use Kullback-Liebler distance. It measures how a random variable Y with probabilities $\{q(x_1),q(x_2),\ldots,q(x_C)\}$ divergates from a baseline random variable X from the previous paragraph. Kullback-Liebler distance is measured in the same units as the entropy is.

$$D(X \parallel Y) = D(p(x) \parallel q(x)) = \sum_{i=1}^{C} p(x_i)\ln\left(\frac{p(x_i)}{q(x_i)}\right) \tag{86}$$

Since Kullback-Liebler distance introduces the baseline distribution, it is not the symmetric measure and therefore it does not meet all the properties to be a real distance[45]. It is safer to use the term Kullback-Liebler divergence.

Using Kullback-Liebler divergence would be straightforward for the investigation how much the proposed document representation divergates from the original n-gram representation[46]. But the Kullback-Liebler divergence is defined for two distributions with the same outcomes; the distributions can differ in probabilities of their outcomes only. Hence we cannot measure $D(\text{vec}(\mathbf{Q})\|\mathbf{c})$ where $\mathbf{c}$ is a centrality representation of a document. The similar problem arises when comparing the pure entropy since its upper-bound is tied to the number of outcomes, hence the values $H(\text{vec}(\mathbf{Q}))$ and $H(\mathbf{c})$ are not comparable.

To overcome the problem with the different outcomes of two nominal distributions that we need to compare, the mutual information approach could be helpful. The mutual information measures the reduction of the entropy of one variable when the value of the second variable is known.

$$\begin{aligned} I(X,Y) &= H(X) + H(Y) - H(X,Y) \\ I(X,Y) &= H(X) - H(X \mid Y) = H(Y) - H(Y \mid X) \end{aligned} \tag{87}$$

To be able to determine the mutual information from this formula, one has to define joint and conditional entropies as[47]

[44] Other common units include bits for the logarithm base 2 and dits for the logarithm base 10.
[45] Kulback-Liebler distance is non-negative and equals zero for identical distributions.
[46] Similarly to Kullback-Liebler we can introduce the cross-entropy. The cross-entropy equals Kullback-Liebler divergence plus the entropy of the baseline distribution.
[47] Note that $H(X,X) = H(X) = I(X,X)$.

$$H(X,Y) = -\sum_{i=1}^{C_X}\sum_{j=1}^{C_Y} p(x_i, y_j)\ln\left(p(x_i, y_j)\right)$$
$$H(X \mid Y) = -\sum_{i=1}^{C_X}\sum_{j=1}^{C_Y} p(x_i, y_j)\ln\left(p(x_i \mid y_j)\right) \quad (88)$$

Note that the joint and conditional entropies are defined for two nominal variables with possibly different outcomes. The same holds for the mutual information. Unlike Kullback-Liebler divergence the mutual information is the symmetric measure and it does not rely on a baseline distribution. The units of the mutual information are the same as for the entropy. The relations among the entropies and the mutual information defined by formulas (87) can be illustrated using the following picture.

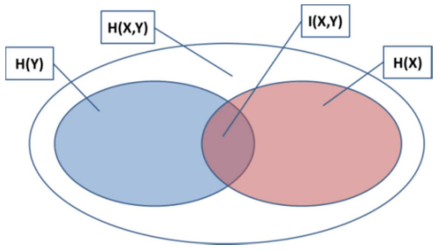


Figure 10: The Venn diagram clarifies the relationship between the entropies and the mutual information.

The formula that enables to compute the mutual information from the original distributions can be derived from formulas (87) and (88) as

$$I(X,Y) = \sum_{i=1}^{C_X}\sum_{j=1}^{C_Y} p(x_i, y_j)\ln\left(\frac{p(x_i, y_j)}{p(x_i)p(y_j)}\right). \quad (89)$$

Finally, the relation between the mutual information and Kullback-Liebler divergence is visible from the previous formula.

$$I(X,Y) = D\left(p(x,y) \| \; p(x)p(y)\right) \quad (90)$$

Note that it is the divergence of the joint distribution of two variables under the hypothesis they are independent from their real joint distribution. These two joint distributions have the same outcomes therefore Kullback-Liebler divergence is applicable.

To investigate the relation between the original representation of a document and the proposed representation, let us try to derive the mutual information between vectorized forms of the matrices **T** and **LTR** where **L** and **R** are again some appropriate matrices that enable to express selected proposed representations[48]. The distribution of **T** is described in formulas (60), (61), (62) and (63), hence we are able to express its entropy.

[48] Namely Degree, InDegree and OutDegree can be obtained as a linear combination of co-occurrence frequencies. Other considered centralities require non-linear transformations of the matrix **T**.

$$\mathbf{T} \sim Mu(L, \mathbf{Q})$$

$$H(\mathbf{T}) = -\sum_{\sum_{i=1}^{M}\sum_{j=1}^{N} t_{ij} = L} \frac{L!}{\prod_{i=1}^{M}\prod_{j=1}^{N} t_{ij}!} \prod_{i=1}^{M}\prod_{j=1}^{N} q_{ij}^{\,t_{ij}} \ln\left(\frac{L!}{\prod_{i=1}^{M}\prod_{j=1}^{N} t_{ij}!} \prod_{i=1}^{M}\prod_{j=1}^{N} q_{ij}^{\,t_{ij}} \right) \tag{91}$$

It has been already stated that the general distribution of the vectorized form of **LTR** is neither multinomial nor any common one. Therefore we are not able to directly express the entropies H(**LTR**), H(**T**|**LTR**) or H(**T**, **LTR**) that are necessary for the computation of the mutual information I(**T**,**LTR**) using the formula (87). Even the formula (89) is not useful to compute I(**T**,**LTR**) by the same reasons.

To be able to express the mutual information I(**T**,**LTR**), we again approximate the multinomial distribution of **T** by the multivariate normal distribution from the formula (64). The distribution of **LTR** is then the multivariate normal described in the formulas (67) and (68).

As we have approximated the nominal distributions by the continuous ones, we have to adapt the definitions of the mutual information and the entropies to continuous distributions. More specifically, we have to express the mutual information and the entropy for multivariate normal distributions and investigate how the linear transformation achieved by the rectangular (singular) matrices **L** and **R** influence these measures.

The discrete (Shannon) entropy from the formula (85) can be substituted by the differential entropy of a continuous random variable X with the probability density function $f(x)$ of the form

$$H(X) = -\int_{-\infty}^{\infty} f(x) \ln\left(f(x) \right) dx . \tag{92}$$

Kullback-Liebler divergence of a continuous random variable Y from a continuous random variable X has the form

$$D(X \parallel Y) = \int_{-\infty}^{\infty} f(x) \ln\left(\frac{f(x)}{g(x)} \right) dx . \tag{93}$$

Note that in this case of two continuous random variables we cannot tackle the problem with the number of outcomes that was mentioned in the case of two discrete random variables. The conditional differential entropy is again based on the joint and conditional distributions and is defined as

$$H(X \mid Y) = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y) \ln\left(f(x \mid y) \right) dx dy . \tag{94}$$

Using the general formula (87) we can conclude that the mutual information between two continuous random variables X and Y is

$$I(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\ln\left(\frac{f(x,y)}{f(x)f(y)}\right)dxdy\,. \tag{95}$$

In the context of Kullback-Leibler divergence the mutual information between two continuous random variables is again the divergence of the true joint distribution from the hypothetic joint distribution of independent random variables.

$$I(X,Y) = D(f(x,y)\,\|\,f(x)f(y)) \tag{96}$$

If we consider an n-dimensional random continuous vector **X** with the probability density function $f(\boldsymbol{x})$ instead of the one-dimensional random variable X, its differential entropy is

$$H(\mathbf{X}) = -\int_{\Re^n} f(\mathbf{x})\ln(f(\mathbf{x}))d\mathbf{x}\,. \tag{97}$$

Analogically, Kullback-Liebler divergence of two n-dimensional random vectors **X** and **Y** with probability density functions $f(\mathbf{x})$ and $g(\mathbf{y})$ is then

$$D(\mathbf{X}\,\|\,\mathbf{Y}) = \int_{\Re^n} f(\mathbf{x})\ln\left(\frac{f(\mathbf{x})}{g(\mathbf{x})}\right)d\mathbf{x}\,. \tag{98}$$

While Kullback-Liebler divergence can be computed for two vectors of the same dimensionality only, the conditional differential entropy is defined for random vectors that may differ in their dimensionalities. The conditional differential entropy of the n-dimensional random continuous vector **X** and the m-dimensional random continuous vector **Y** is

$$H(\mathbf{X}\,|\,\mathbf{Y}) = -\int_{\Re^m}\int_{\Re^n} f(\mathbf{x},\mathbf{y})\ln(f(\mathbf{x}\,|\,\mathbf{y}))d\mathbf{x}d\mathbf{y}\,. \tag{99}$$

Their mutual information using the formula (87) is then

$$I(\mathbf{X},\mathbf{Y}) = \int_{\Re^m}\int_{\Re^n} f(\mathbf{x},\mathbf{y})\ln\left(\frac{f(\mathbf{x},\mathbf{y})}{f(\mathbf{x})f(\mathbf{y})}\right)d\mathbf{x}d\mathbf{y}\,. \tag{100}$$

This mutual information again corresponds to Kullback-Leibler divergence of the true joint distribution from the hypothetic joint distribution of two independent random vectors.

$$I(\mathbf{X},\mathbf{Y}) = D(f(\mathbf{x},\mathbf{y})\,\|\,f(\mathbf{x})f(\mathbf{y})) \tag{101}$$

Note that the differential entropy does not share all the properties of Shannon entropy. The differential entropy can be negative since the probability density function can be greater than one. There is not also an upper limit for the differential entropy since the number of possible outcomes of a continuous random variable is unbounded.

The relation between the differential and Shannon entropies is noticeable if the continuous random variable X is approximated by its discretized form X^{Δ}.

$$X^{\Delta} = x_i \Leftrightarrow i\Delta \le x_i < (i+1)\Delta$$
$$H(X^{\Delta}) = -\sum_{i=-\infty}^{\infty} p(x_i)\ln(p(x_i)) = -\sum_{i=-\infty}^{\infty} \Delta f(x_i)\ln(\Delta f(x_i)) \quad (102)$$

The symbol Δ stands for the length of an equidistant interval on which X is considered to be approximately constant with the probability $f(x)\Delta$. If Δ is small enough, Shannon entropy differs from the discrete entropy by the factor $ln(\Delta)$.

$$H(X^{\Delta}) + \ln(\Delta) \to H(X) \quad (103)$$

The similar relation holds for the conditional entropy. Note if Δ comes to zero, the number of the outcomes of X^{Δ} increases and the entropy $H(X^{\Delta})$ increases as well while the factor $ln(\Delta)$ decreases to minus infinity. Using the definition of mutual information from the formula (87) we can state that

$$I(X^{\Delta}, Y^{\Delta}) = H(X^{\Delta}) - H(X^{\Delta} \mid Y^{\Delta}) \to H(X) - \ln(\Delta) - H(X \mid Y) + \ln(\Delta) = I(X,Y) \quad (104)$$

Now let us investigate how the linear transformations influence the entropies and the mutual information. If a continuous random variable X has the probability density function $f(x)$ then the random variable $Y=aX$ has the probability density function $g(y)$ of the form

$$g(y) = \frac{1}{|a|} f\left(\frac{y}{a}\right). \quad (105)$$

Using the formulas (92) and (94) and the transformation of the probability density function from the previous formula we get

$$\begin{aligned} H(aX) &= H(X) + \log(|a|) \\ H(aX \mid Y) &= H(X \mid Y) + \log(|a|) \end{aligned}. \quad (106)$$

And exploiting the formula (87) we can conclude that a linear transformation of a random continuous variable does not influence its mutual information with other random continuous variable.

$$I(aX, Y) = H(aX) - H(aX \mid Y) = H(X) + \log(|a|) - H(X \mid Y) - \log(|a|) = I(X,Y) \quad (107)$$

The above formula implies the mutual information between a random variable and its linear transformation as

$$I(aX, X) = I(X, X) = H(X). \quad (108)$$

Now let us consider random vectors that will be useful for investigation of relationships among the proposed document representations. If a continuous random vector **X** has the

probability density function $f(\mathbf{x})$, then the random vector $\mathbf{Y}=\mathbf{AX}$, where $\mathbf{A}$ is a regular matrix, has the probability density function $g(\mathbf{y})$ of the form[49]

$$g(\mathbf{y}) = \left|\det\left(\mathbf{A}^{-1}\right)\right| f\left(\mathbf{A}^{-1}\mathbf{y}\right). \tag{109}$$

Using the formulas (97)and (99) and the transformation of the vector probability density function from the previous formula, we get the following entropies for the linearly transformed random vector $\mathbf{X}$ by the regular matrix $\mathbf{A}$.

$$\begin{aligned} H(\mathbf{AX}) &= H(\mathbf{X}) + \log\left(\left|\det(\mathbf{A})\right|\right) \\ H(\mathbf{AX}\,|\,\mathbf{Y}) &= H(\mathbf{X}\,|\,\mathbf{Y}) + \log\left(\left|\det(\mathbf{A})\right|\right) \end{aligned} \tag{110}$$

And exploiting the formula (87) that holds also for vector random variables, we can conclude that the regular linear transformation of the random continuous vector does not influence its mutual information with some other random continuous vector.

$$\begin{aligned} I(\mathbf{AX},\mathbf{Y}) &= H(\mathbf{AX}) - H(\mathbf{AX}\,|\,\mathbf{Y}) \\ &= H(\mathbf{X}) + \log\left(\left|\det(\mathbf{A})\right|\right) - H(\mathbf{X}\,|\,\mathbf{Y}) - \log\left(\left|\det(\mathbf{A})\right|\right) = I(\mathbf{X},\mathbf{Y}) \end{aligned}. \tag{111}$$

As the special case we can derive the mutual information between the random vector $\mathbf{X}$ and its regular linear transformation as

$$I(\mathbf{AX},\mathbf{X}) = I(\mathbf{X},\mathbf{X}) = H(\mathbf{X}). \tag{112}$$

The provided formulas for the mutual information between continuous random variables and vectors are applicable regardless to their distributions. Let us provide the main formulas for the mutual information for variables with the normal distributions that are the proper approximations of our proposed document representations.

If a random variable X comes from the normal distribution with the mean μ and the variance σ, ^{2} its differential entropy can be derived using the formula (92).

$$\begin{aligned} X &\sim N\left(\mu,\sigma^2\right) \\ H(X) &= \frac{1}{2}\ln\left(2\pi e\sigma^2\right) \end{aligned} \tag{113}$$

Skipping the rather complicated formulas for Kullback-Liebler divergence and the conditional entropy of two normal random variables, their mutual information using (95) is

$$I(X,Y) = -\frac{1}{2}\ln\left(1-\mathrm{corr}(X,Y)^2\right). \tag{114}$$

The correlation $corr(X,Y)$ follows the standard definition of normalized covariance.

[49] If we consider a general invertible transformation function the determinant is then substituted by Jacobian of the inverse of the transformation function. Our formula covers only the linear transformation as the special case.

$$\mathrm{corr}(X,Y)=\frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)\mathrm{var}(Y)}} \tag{115}$$

Note that formula (114) provides the relation between two standard measures of the relation between two normal variables[50]. The known correlation can be easily transformed to the mutual information while the known mutual information does imply only the magnitude of the correlation but not the sign. Using the formulas (107) and (114) it can be verified that the linear transformation does not influence the magnitude of the correlation. On the other hand, the formula (114) is not suitable for the investigation of the special case $I(aX,X)$ because $|corr(aX,X)|=1$, but this case is covered by general formulas (108) and (113).

If $\mathbf{X}$ is a random vector that comes from the vector normal distribution[51] with the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\sum$, its differential entropy can be derived using the formula (97).

$$\begin{aligned}&\mathbf{X}\sim N(\boldsymbol{\mu},\boldsymbol{\Sigma})\\ &H(\mathbf{X})=\frac{1}{2}\ln\left((2\pi e)^n\det(\boldsymbol{\Sigma})\right)\end{aligned} \tag{116}$$

Comparing this formula with (113) we can observe that the variance from the one-dimensional case is now replaced by the determinant of the variance matrix[52]. The mutual information of two normal random vectors using the general formula (87) is then

$$I(\mathbf{X},\mathbf{Y})=-\frac{1}{2}\ln\left(\frac{\det(\mathrm{var}(\mathbf{X},\mathbf{Y}))}{\det(\mathrm{var}(\mathbf{X}))\det(\mathrm{var}(\mathbf{Y}))}\right). \tag{117}$$

The term $var(\mathbf{X},\mathbf{Y})$ in this formula represents the symmetric block variance-covariance matrix that can be obtained as the variance matrix of the joint distribution of $\mathbf{X}$ and $\mathbf{Y}$[53].

$$\mathrm{var}(\mathbf{X},\mathbf{Y})=\begin{pmatrix}\mathrm{var}(\mathbf{X}) & \mathrm{cov}(\mathbf{X},\mathbf{Y})\\ \mathrm{cov}(\mathbf{Y},\mathbf{X}) & \mathrm{var}(\mathbf{Y})\end{pmatrix} \tag{118}$$

The determinant of the block matrix $var(\mathbf{X},\mathbf{Y})$ can be modified using the following general formula that holds for invertible diagonal blocks.

$$\det\begin{pmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{pmatrix}=\det(\mathbf{A})\det(\mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B})=\det(\mathbf{D})\det(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C}) \tag{119}$$

[50] The provided relation between the correlation and the mutual information holds for normal distributed variables only; for random variables with other distributions the formula provides the lower bound of their mutual information.

[51] The form of the probability density function of the vector normal distribution is stated for the vectorized form of the random matrix in the formula (64).

[52] The determinant of any variance matrix cannot be negative because the covariance matrix is positive semi-definite.

[53] It is a distribution of the random vector that is constructed by merging the vectors $\mathbf{X}$ and $\mathbf{Y}$ into one column vector.

Exploiting the relation among conditional and unconditional variance and covariance matrices

$$\text{var}(\mathbf{X} \mid \mathbf{Y}) = \text{var}(\mathbf{X}) - \text{cov}(\mathbf{X}, \mathbf{Y})\text{var}(\mathbf{Y})^{-1}\text{cov}(\mathbf{Y}, \mathbf{X}) \tag{120}$$

we can conclude that the mutual information between two normal vectors equals[54]

$$I(\mathbf{X}, \mathbf{Y}) = -\frac{1}{2}\ln\left(\frac{\det(\text{var}(\mathbf{X} \mid \mathbf{Y}))}{\det(\text{var}(\mathbf{X}))}\right). \tag{121}$$

The ratio of the determinants in the logarithm is always from the interval <0;1> and can be explained as the relative proportion of the variability of the random vector **X** that is not explained by the vector **Y**. Comparing the formulas (114) and (121) we can conclude that one minus this ratio of the determinants is the generalization of the coefficient R-squared[55] for random normal vectors. This property will be taken into account later in the text.

We have already shown that a linear transformation of random variables does not influence their mutual information (remember formulas (107) and (108)). The same statement holds for the linear transformation of random vectors (remember formulas (111) and (112)) if the transformation matrix **A** is regular. In the proposed document representation we transform the vectorized form of the random transition matrix **T** (remember formula (65)) that is assumed to be normally distributed using the non-regular matrix $\mathbf{R}^T \otimes \mathbf{L}$[56]. The reason for non-regular transformation is that we try to propose an appropriate document representation in a low-dimensional space that is suitable for further processing of documents by predictive models. Therefore we have to investigate how non-regular transformations influence the mutual information. We should pay special interest to the formula for the mutual information $I((\mathbf{R}^T \otimes \mathbf{L})vec(\mathbf{T}), \text{vec}(\mathbf{T}))$ that measures the relation between the original and the proposed low-dimensional representation.

Let us start from the formula (117) that enables to compute the mutual information between two normally distributed random vectors. Generally, we need to derive the formula for the mutual information $I(\mathbf{AX}, \mathbf{X})$ that holds for **A** irregular. For any transformation matrices **A** and **B** the variance and covariance matrices of transformed vectors can be expressed as

$$\begin{aligned} \text{var}(\mathbf{AX}) &= \mathbf{A}\,\text{var}(\mathbf{X})\mathbf{A}^T \\ \text{cov}(\mathbf{AX}, \mathbf{BY}) &= \mathbf{A}\,\text{cov}(\mathbf{X}, \mathbf{Y})\mathbf{B}^T \end{aligned}. \tag{122}$$

The variance matrix of joint distribution of $(\mathbf{AX}, \mathbf{X})$ from the formula (117) is then

$$\text{var}(\mathbf{AX}, \mathbf{X}) = \begin{pmatrix} \mathbf{A}\,\text{var}(\mathbf{X})\mathbf{A}^T & \mathbf{A}\,\text{var}(\mathbf{X}) \\ \text{var}(\mathbf{X})\mathbf{A}^T & \text{var}(\mathbf{X}) \end{pmatrix}. \tag{123}$$

And exploiting the formula (119) for the determinant of a block matrix, we can conclude

[54] This formula for mutual information can be also derived directly from formulas (87) and (116).

[55] The coefficient of determination (R-squared) between two one-dimensional random variables equals to the squared correlation of these variables. R-squared can be interpreted as the percentage of the variability of one random variable explained by the second random variable.

[56] Some representations even use non-linear transformations.

$$\det(\mathrm{var}(\mathbf{AX},\mathbf{X})) = \det(\mathrm{var}(\mathbf{X}))\det\left(\mathbf{A}\,\mathrm{var}(\mathbf{X})\mathbf{A}^T - \mathbf{A}\,\mathrm{var}(\mathbf{X})\,\mathrm{var}(\mathbf{X})^{-1}\,\mathrm{var}(\mathbf{X})\mathbf{A}^T\right) = 0 \quad (124)$$

It implies using the formula (117) that the mutual information $I(\mathbf{AX},\mathbf{X})$ cannot be exactly determined for the irregular transformation matrix $\mathbf{A}$[57].

$$I(\mathbf{AX},\mathbf{X}) \to \infty \quad (125)$$

The same conclusion can be drawn using the conditional variances from the formulas (121) and (120) that imply

$$\det\left(\mathrm{var}\left(\mathbf{AX} \mid \mathbf{X}\right)\right) = 0\,. \quad (126)$$

Now we have come to the conclusion that the mutual information is not a useful theoretical measure of the quality of the proposed document representation at least in the cases when the transformation of the vectorized form of the transition matrix *vec*(**T**) is irregular. In other words, we cannot determine $I((\mathbf{R}^{\mathrm{T}}\otimes\mathbf{L})vec(\mathbf{T}),\mathrm{vec}(\mathbf{T}))$ for any document representation where the product $\mathbf{R}^{\mathrm{T}}\otimes\mathbf{L}$ yields to an irregular matrix.

$$I\left(\left(\mathbf{R}^T \otimes \mathbf{L}\right)\mathrm{vec}\left(\mathbf{T}\right), \mathrm{vec}\left(\mathbf{T}\right)\right) \to \infty \quad (127)$$

Unfortunately, the irregular transformations play the key role in the proposed approach because a significant reduction of dimensionality is desired. The impossibility to compare the theoretical mutual information of the different proposed document representation prompts to use the different approach. Even though the mutual information is the standard measure in the field of document processing, another approach must be employed to evaluate the reduction of information that yields from the proposed transformations from the transition representation of documents to low-dimensionality vectors.

We have already shown that it is suitable to approximate the exact multinomial distribution of the transition matrix **T** by the multivariate normal distribution from formula (64), hence we can pay attention to investigation of the correlation like comparisons of the proposed document representations.

5.4.2 Comparison of covariance matrices

The main issue of this approach is to select an appropriate measure of the association between the original transition matrix **T** of a document and its transformed form $(\mathbf{R}^{\mathrm{T}}\otimes\mathbf{L})vec(\mathbf{T})$. Similarly to one-dimensional continuous distributions, where the correlation and the covariance are commonly used, we try to propose and investigate the similar measures for multivariate continuous distributions. This approach will be applied to the multinomial normal distributions that approximate the unknown distributions of the proposed document representations.

Let us start with one-dimensional approach to commemorate the common measures used for the evaluation of relationships among continuous random variables. The variability of a random variable *X* will be measured by its variance *var*(*X*).

[57] Note that the formula (111) holds for the regular transformation matrix **A**.

$$\mathrm{var}(X) = \int_{-\infty}^{\infty} (x - E(X))^2 f(x) dx = E(X^2) - E(X)^2 \tag{128}$$

The symbol $E(X)$ stands for the expected value of X, $f(X)$ is the probability density function. The variance $var(X)$ is always non-negative. The variance is not upper-bounded, it is measured in the square units of random variable X[58]. The conditional variance $var(X|Y)$ is defined analogically using the conditional expected values $E(X|Y)$ and $E(X^2|Y)$ and the conditional probability density function $f(x|y)$.

The relation between two random continuous variables X and Y with the joint probability density function $f(x,y)$ will be measured by their covariance $cov(X,Y)$.

$$\mathrm{cov}(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x - E(X))(y - E(Y)) f(x,y) dx dy = E(XY) - E(X)E(Y) \tag{129}$$

The covariance $cov(X,Y)$ is not lower or upper-bounded, its magnitude is proportional to the common variability that is shared between X and Y. It is measured in the units of X times the units of Y. The covariance $cov(X,Y)$ can be standardized to the scale between minus one and one introducing the well-known correlation $corr(X,Y)$. The relation between $cov(X,Y)$ and $corr(X,Y)$ has been already stated in the formula (115). To avoid the direction of the relation between X and Y, we will use the square of correlation. Its values fall to the interval between zero and one and can be interpreted as the proportion of the variability of X explained by Y or vice versa.

$$R^2(X,Y) = \mathrm{corr}(X,Y)^2 \tag{130}$$

Indeed, it is well-known R-squared (coefficient of determination) known also from the linear regression that is generally defined using variances as

$$R^2(X,Y) = 1 - \frac{\mathrm{var}(X \mid Y)}{\mathrm{var}(X)}. \tag{131}$$

The conditional variance in the nominator can be replaced by the unconditional variance using the following relation between the variances.

$$\mathrm{var}(X \mid Y) = \mathrm{var}(X) - \frac{\mathrm{cov}(X,Y)^2}{\mathrm{var}(Y)} \tag{132}$$

To investigate R-squared between a continuous random variable and its linear transformation, we have to establish the transformations of variances and covariances. The formulas for the variance and the covariance of linearly transformed random variables can be easily derived from the basic formulas (128) and (129).

[58] Some authors use the dimensionless coefficient of variation that is defined as the ratio of the square root of the variance and the expected value.

$$\begin{aligned}\operatorname{var}(aX) &= a^2 \operatorname{var}(X) \\ \operatorname{cov}(aX, bY) &= ab \operatorname{cov}(X, Y)\end{aligned} \tag{133}$$

Using these formulas we can conclude that any linear transformation of random variables does not influence their R-squared. This conclusion is also obvious from the relation of R-squared and the correlation stated in the formula (130).

$$R^2(aX, Y) = 1 - \frac{\operatorname{var}(aX \mid Y)}{\operatorname{var}(aX)} = 1 - \frac{a^2 \operatorname{var}(X \mid Y)}{a^2 \operatorname{var}(X)} = R^2(X, Y) \tag{134}$$

R-squared for the random variable X and its linear transformation aX is then the same as $R^2(X,X)$ that equals one.

$$R^2(aX, X) = 1 - \frac{a^2 \operatorname{var}(X \mid X)}{a^2 \operatorname{var}(X)} = 1 \tag{135}$$

However, we need to investigate R-squared for two random normally distributed random vectors. Unfortunately, R-squared is not defined for random vectors, hence some generalized version of R-squared should be introduced. It has been already stated when investigating the properties of the mutual information that the determinant of the variance matrix is the multivariate generalization of the covariance. Hence we can define the measure $\lambda(\mathbf{X},\mathbf{Y})$ for random vectors $\mathbf{X}$ and $\mathbf{Y}$ that share the same properties as R-square.

$$\lambda(\mathbf{X}, \mathbf{Y}) = 1 - \frac{\det(\operatorname{var}(\mathbf{X} \mid \mathbf{Y}))}{\det(\operatorname{var}(\mathbf{X}))} \tag{136}$$

$\lambda(\mathbf{X},\mathbf{Y})$ is symmetric, its values fall into the interval between zero and one and can be interpreted as the proportion of the common variability shared by random vectors $\mathbf{X}$ and $\mathbf{Y}$. This coefficient is also known as Wilks' lambda and it is used for example in the discriminant analysis. The definition of Wilks' lambda can be also rewritten using the unconditional variance matrices and exploiting the formulas (119) and (120).

$$\lambda(\mathbf{X}, \mathbf{Y}) = 1 - \frac{\det(\operatorname{var}(\mathbf{X}, \mathbf{Y}))}{\det(\operatorname{var}(\mathbf{X}))\det(\operatorname{var}(\mathbf{Y}))} \tag{137}$$

The term $var(\mathbf{X},\mathbf{Y})$ is again the block variance matrix of the joint random variable $(\mathbf{X},\mathbf{Y})$ where the vectors $\mathbf{X}$ and $\mathbf{Y}$ are concatenated into one vector as shown in (118). Comparing this formula with (130) and (131), the left-hand side of the above formula can be regarded as a multidimensional analogy of the squared correlation coefficient.

To investigate $\lambda(\mathbf{X},\mathbf{Y})$ between a continuous random vector and its linear transformation, we have to establish the transformations of variance matrices and their determinants. Similarly to the formula (122) that hold for the variance $var(\mathbf{AX})$ the conditional variance matrix for the linearly transformed random vector $\mathbf{X}$ is

$$\operatorname{var}(\mathbf{AX} \mid \mathbf{Y}) = \mathbf{A} \operatorname{var}(\mathbf{X} \mid \mathbf{Y}) \mathbf{A}^T . \tag{138}$$

The formulas (122)and (138) hold for any matrix **A** with appropriate dimensions. To evaluate the determinants the following general formulas should be taken into account.

$$\begin{aligned}&\det\left(\mathbf{X}^T\right)=\det\left(\mathbf{X}\right)\\&\det\left(\mathbf{X}^{-1}\right)=\det\left(\mathbf{X}\right)^{-1}\\&\det\left(\mathbf{AX}\right)=\det\left(\mathbf{A}\right)\det\left(\mathbf{X}\right)\end{aligned} \tag{139}$$

In the above formulas it is assumed that both **A** and **X** are the square matrices, in $\det(\mathbf{X}^{-1})$ the matrix **X** is even invertible. Hence using any square matrix **A** we can conclude that the linear transformation performed by this square matrix does not influence Wilks' lambda.

$$\begin{aligned}\lambda\left(\mathbf{AX},\mathbf{Y}\right)&=1-\frac{\det\left(\mathrm{var}\left(\mathbf{AX}\mid\mathbf{Y}\right)\right)}{\det\left(\mathrm{var}\left(\mathbf{AX}\right)\right)}=1-\frac{\det\left(\mathbf{A}\,\mathrm{var}\left(\mathbf{X}\mid\mathbf{Y}\right)\mathbf{A}^T\right)}{\det\left(\mathbf{A}\,\mathrm{var}\left(\mathbf{X}\right)\mathbf{A}^T\right)}\\&=1-\frac{\det\left(\mathbf{A}\right)^2\det\left(\mathrm{var}\left(\mathbf{X}\mid\mathbf{Y}\right)\right)}{\det\left(\mathbf{A}\right)^2\det\left(\mathrm{var}\left(\mathbf{X}\right)\right)}=\lambda\left(\mathbf{X},\mathbf{Y}\right)\end{aligned} \tag{140}$$

Unfortunately, in the proposed document representations the original transition matrix **T** is transformed differently. Its vectorized form *vec*(**T**) is projected into a low-dimensional space, hence it cannot be assumed that the transformation is realized by a square matrix. Additionally, we need to evaluate the relation between the original vector representation and its transformation. Regardless of the shape of the transformation matrix **A** using the formulas (120)and (138) it holds

$$\mathrm{var}\left(\mathbf{AX}\mid\mathbf{X}\right)=\mathbf{A}\,\mathrm{var}\left(\mathbf{X}\mid\mathbf{X}\right)\mathbf{A}^T=\mathbf{A}\left(\mathrm{var}\left(\mathbf{X}\right)-\mathrm{cov}\left(\mathbf{X},\mathbf{X}\right)\mathrm{var}\left(\mathbf{X}\right)^{-1}\mathrm{cov}\left(\mathbf{X},\mathbf{X}\right)\right)\mathbf{A}^T=\mathbf{0} \tag{141}$$

Therefore evaluating $\lambda(\mathbf{AX},\mathbf{Y})$ using formula (136) we can conclude that any linear transformation realized by the matrix **A** yields to the same value of one.

$$\lambda\left(\mathbf{AX},\mathbf{X}\right)=1-\frac{\det\left(\mathrm{var}\left(\mathbf{AX}\mid\mathbf{X}\right)\right)}{\det\left(\mathrm{var}\left(\mathbf{AX}\right)\right)}=1 \tag{142}$$

This conclusion can be generalized also for some non-linear transformations. If the term **AX** is replaced by a vector function **f**(**X**), the conditional variance matrix *var*(**f**(**X**)|**X**) also equals zero matrix similarly to the formula (141).

Regarding the formula (142) for the case of our document representation, we are not able to evaluate how the proposed transformation of the transition matrix **T** stated in (65) reduces its variability.

$$\lambda\left(\left(\mathbf{R}^T\otimes\mathbf{L}\right)\mathrm{vec}\left(\mathbf{T}\right),\mathrm{vec}\left(\mathbf{T}\right)\right)=1 \tag{143}$$

The same results would be obtained if the linear projection of *vec*(**T**) represented by ($\mathbf{R}^T\otimes\mathbf{L}$) was replaced by the more general but unambiguous projection of *vec*(**T**). Hence we can

conclude that Wilks' lambda cannot provide the information how much variability is lost when any of the proposed document representations is selected.

However, the variability of the transition matrix **T** enables to distinguish between documents within a collection, hence it is desired to preserve its variability as much as possible in any document representation. On the other hand, we proved that neither the mutual information nor Wilks' lambda are the suitable measures to evaluate how much information or the variability of the original document representation included in the matrix **T** is lost when **T** is projected into a low-dimensional vector. If the distribution of *vec*(**T**) is approximated by the multivariate normal distribution, the mutual information diverges while Wilks' lambda always equals one. The multivariate normal approximation is necessary because the transformation of the original multinomial distribution of the matrix **T** is not tractable.

5.5 Conclusions summary

The main purpose of the theoretical evaluation of properties of the proposed document representations was to quantify the extent to which the information about a within document context is propagated to a document vector. To fulfill this goal the distribution of the n-gram transition matrix that is a carrier of the contextual information was substituted by the multivariate normal distribution to simplify consequent transformations. Then the distributions of the proposed representations that originate in transformations of the context network were estimated. Only some of the tested context network centralities offer document vectors with known distributions or distributions where the variability can be estimated.

Then the reduction of variability for the proposed representations with the known distributions was investigated in two approaches. In the first one, the concept of mutual information between the context network and the derived document vector was utilized. In the second approach, where variabilities of multivariate distributions are expressed by determinants of their covariance matrices, a portion of the original variability of the context network explained by the variability of the derived document vector is quantified.

Summarizing all the results collected in this chapter, it was shown that neither the mutual information approach nor the approach where covariance matrices are compared do not offer a satisfactory comparison of document representations. We are not able to precisely evaluate the proportion of the information that is lost by any of the proposed document representations. Hence our attention should be switched to experimental results. Let us investigate how the document variability within a document collection is preserved when transition matrices of the documents are transformed to any of the proposed low-dimensional vector representations. Additionally, it is worth exploring how the different representations influence the evaluation measures of standard text mining tasks, such as the classification of documents or the clustering.

6 Experiments

The theoretical assessment of the benefits of the proposed representations is not generally tractable (see the conclusions of chapter 5), hence we have to focus on comprehensive experiments. Instead of theoretical proves statistical tests are performed to explore the appropriateness of the proposed representations. Each experiment is repeated several times using a different random seed which influences the sampling.

The experiments are performed with artificially generated documents and also with real document collections. The generated texts enable to assess better the reduction of the contextual diversity of documents instead of theoretical evaluation from chapter 5 because they are generated to follow the assumptions about contextual ties within the documents that are borrowed from standard n-gram language models. On the other hand, the real downloaded documents offer more realistic assessment of the performance of the proposed representations in complex text mining tasks.

6.1 Goals of experimental evaluation

The purpose of the experiments is not only to assess the reduction of document diversity using different representations because the theoretical evaluation does not offer satisfactory estimates. The experiments also enable to test the performance of document representations in complex text mining tasks with examining effects of various combinations of parameters of the proposed representations such as the length of the context window or the vocabulary size. We can observe how these parameters influence the reduction of the diversity of documents within a collection and the performance of text mining models. The goals of experiments presented in the following chapters are:

- Assess the contextual document diversity reduction that arises from using the proposed representations and compare it with the diversity reduction that arises from using some standard representation.
- Test the usefulness of the proposed representations in real text mining tasks such as the information retrieval, the document classification or the clustering. Compare the performance of the proposed representation with the performance of a standard representation that does not comprise any contextual information.
- Perform the tests of usefulness of the proposed representations for various parameter settings and document collection types to obtain practical recommendations for using appropriate adjustments in different situations.

There are many combinations of the setting of all adjustable parameters in our experiments, hence it is not possible to investigate the effect of all these combinations. When examining the effect of one particular parameter, the other parameters are fixed to some designated standard values, thus reported parameter effects are valid for the standard values of the other parameters only. On the other hand, each experiment is carried out several times using a different random seed to support presented conclusions by statistical testing.

6.2 Methodology

The experiments may be divided into two parts depending on a source of documents. Firstly, the artificial documents are used. The artificial document is a sequence of generated topics; it is not necessary to generate a sequence of tokens or words because they would be transformed back to the sequence topic before further processing. The artificial sequences meet the

assumptions about the n-gram contextual ties within a document; they are generated based on the known transition matrices. Secondly, the collections of downloaded documents created by human writers are used as primary input data. All necessary preprocessing steps are performed with these real documents to receive useful sequences of topics. Moreover, the sentence borders identified in the original documents are propagated to the topic sequences to enable to ignore the adjacency of topics from different sentences or paragraphs.

All investigated document representations are then derived from the topic sequences; the topics put together form a dictionary. Other linguistic entities than topics could be used as well to derive the proposed representations, but a relatively small size of the topic dictionary lowers the resource requirements for the presented experiments.

The standard document representation used for experimental comparisons is the bag-of-topic representation that does not comprise any contextual information; it is based solely on topic frequencies within a document. The proposed and tested document representations are derived as centralities of context networks which are constructed from the topic sequences applying a given context window. The bag-of-topic representation and the proposed representations are of the same dimensionality, hence they enable to examine and compare the effect of encoding contextual information into document vectors.

The first experiments concern the investigation of the preservation of the document diversity. They should substitute the theoretic evaluation from chapter 5. The preservation of the document diversity is investigated using the SSTRESS measure. It describes how proximities between document pairs within a collection are disrupted when documents are projected from n-grams to any desired representation. These experiments are performed on both generated and downloaded documents.

The initial test of practical exploitation of the contextual information embedded in the proposed document vectors is performed as the recognition of the documents with randomly permuted topics among other artificial documents that were generated from the given n-gram distributions. This test is designed as a supervised binary classification; the permuted documents are labeled. The classification is then evaluated by the F-measure.

Before wider testing the usefulness of the proposed representation in common text mining tasks, a specific binary classification of real documents was examined in a special task where the contextual ties are apparently important: the goal was to recognize machine translated documents among other documents written by human authors. The test is again designed as a supervised binary classification; the translated documents are labeled and the benefit of the contextual representation is assessed by the F-measure.

The appropriateness of the proposed representations for the information retrieval task is investigated on generated documents. The documents that should be retrieved are generated based on the same n-gram distribution as a query while documents that should not be retrieved are generated from different n-gram distributions. The F-measure is used again to evaluate the information retrieval task.

The most common text mining task is the classification. The real downloaded documents are labeled, hence the nominal classification is evaluated on them. The nominal classification is evaluated using the F-measure; the micro-averaging is used to combine F-measures of all target categories.

Last experiments concern the clustering. They are conducted on both generated and downloaded documents. Artificial documents within each presumed cluster are generated from the same n-gram distribution; the distributions between clusters differ. The normalized mutual information serves as a measure for the supervised evaluation of clustering of the

generated documents. The clustering of the downloaded collections is evaluated by the normalized mutual information as well because the document labels are exploited in the evaluation process.

The effects of several adjustable parameter settings are investigated in all experiments with both real and downloaded documents. The parameters common for all tasks include the topic vocabulary size and the length of the left context window. In addition, the generated documents enable to investigate the effects of the length documents or the number of documents in a collection.

The experiments are conducted for each tested document representation separately. Each experiment is repeated several times using a different random seed. It enables to eliminate random variations and to submit conclusions as results of the statistical testing. A new collection is generated for each repetition of the experiment with the simulated documents while the random seed is used to fixing the partitioning to test and training sets in supervised experiments with the downloaded collections.

6.3 Experimental setup

6.3.1 Processing of downloaded documents

The downloaded documents come from three different collections of press releases (chapter 6.3.1.1). Each collection offers the document categories that can be exploited in any supervised evaluation. The collections differ in many parameters including the language, the number of documents, the average document length or the number of document categories to cover the possible diversity of real collections in our evaluation process.

The documents were processed by the proposed pipeline described in chapter 6.3.1.2. The pipeline outcome included not only the proposed document vectors, but also the documents where the original tokens were substituted by topics detected by LDA together with their transitions matrices between the topic (n-1)-grams and the topics. It enabled to estimate topic n-gram probabilities for each document[59]. The estimated n-gram probabilities serve as the original and most informative document representation in the following comparisons because they are assumed to form the distribution for the unobserved generative document process.

6.3.1.1 Downloaded collections

Three default collections of downloaded documents were used in the delivered experiments. An additional special collection was derived from the Czech collection for the task of the recognition of machine translated documents. Each collection is language homogenous; it includes documents of the same language. We decided to test several collections of different languages to investigate how the language influences the results. These collections differ also in many other parameters than the language to cover the wide range of possible properties of other potential collections.

The documents in each collection are labeled; the labels can be used for any supervised learning and evaluation. The labels in three default collections represent categories under which the documents were published on a news server. The fourth special collection is labeled by an indicator of the machine translation.

[59] The maximal likelihood estimates are applied.

ID	language	documents	tokens	average length	source	URL
CZ1	Czech	24 095	10 994 678	456	Novinky.cz	www.novinky.cz
EN1	English	13 285	8 106 108	610	BBC	www.bbc.co.uk
GE1	German	2 501	985 410	394	Wiener Zeitung	www.wienerzeitung.at
CZ2	Czech	3 097	1 229 492	397	Novinky.cz	www.novinky.cz

Table 2: The properties of four experimental collections of downloaded documents.

Before the processing of texts the document collections were transformed to the plain text format and tokenized. The tokens were filtered by a stop-word list in each collection. Then the tokens that included numbers were discarded. The tokens were then stemmed using language dependent stemmers. The last filter is based on the vocabulary. The vocabulary was extracted for each collection from the train documents after the partitioning. Only the stemmed tokens that appear in at least two training documents were included.

collection	stop-word list size	vocabulary size[60]
CZ1	1 174	87 583 – 88 412
EN1	185	21 498 – 21 598
GE1	231	13 880 – 14 477
CZ2	631	19 398 – 20 039

Table 3: The sizes of vocabularies that were extracted from the downloaded collections.

6.3.1.1.1 Extraction of collection CZ1

The first experimental collection consists of newspaper articles written in the Czech language. They were downloaded from the news portal novinky.cz. in October 2013. The documents were downloaded by WinHTTrack Website Copier crawler. Only the press releases that belong to specified categories were downloaded, other pages were discarded. Categories were available as a part of URL. The original document taxonomy includes 16 categories on the first level; some first level categories include the second level categories. For the classification and clustering experiments the second level categories were omitted.

For further processing the HTML pages were transformed to flat text files. Pictures and other undesirable objects were filtered out and HTML codes were parsed using Python script. Only the text from the selected tags was retrieved to filter out adverts and other unimportant text. The HTML ampersand references were decoded to its original characters and the encoding was changed to cp-1250 before saving as the flat files. The text files are identified by their original numeric identifiers given by novinky.cz. The collection includes 24 thousand documents with 10 million tokens.

[60] The size of vocabulary varies due repeating the whole experiment five times with different random seeds. The collection is partitioned to training and test sets in the each run. The vocabulary is extracted using the training documents only.

ID	category CZ	category EN	documents
1	auto	car	453
2	bydleni	habitation	1486
3	cestování	travel	2034
4	domácí	domestic news	3104
5	ekonomika	economy	1237
6	finance	finance	320
7	Internet a PC	Internet and PC	625
8	kariéra	career	220
9	koktejl	cocktail	1162
10	krimi	crime	2131
11	kultura	culture	2041
12	filmový festival	film festival	208
13	vánoce	Christmas	111
14	věda, školy	science, education	309
15	zahraniční	foreign news	3298
16	žena	woman	5356
total			**24095**

Table 4: The categories of the collection CZ1

6.3.1.1.2 Extraction of collection EN1

The second collection contains English newspaper articles from BBC. They were downloaded from bbc.co.uk in October 2013 using WinHTTrack Website Copier crawler. Webpages other than the newspaper articles were discarded. The articles were categorized using their file names. The original document taxonomy includes 10 categories on the first level. Domestic news and world news could be further divided into subcategories up to the third level. For the classification and clustering experiments only the first level categories were considered.

The pictures and other undesirable objects were filtered out and the HTML code was transformed to a flat text using Python script. Only the text from the selected tags was retrieved to filter out the unimportant text and the ampersand references were decoded before saving as the flat files. The text files are identified by their original numeric identifiers given by bbc.co.uk. The collection includes 13 thousand documents with 8 million tokens.

ID	category	documents
1	business	1415
2	education	163
3	entertainment, arts	319
4	health	424
5	in pictures	47
6	magazine	853
7	science, environment	576
8	technology	501
9	UK	3540
10	world	5447
total		**13285**

Table 5: The categories of the collection EN1

6.3.1.1.3 Extraction of collection GE1

The third collection is the collection of German newspaper articles from Austrian newspaper Wiener Zeitung. They were downloaded from wienerzeitung.at in October 2013 similarly as the previous collections. The original document taxonomy includes 8 categories on the first level; other levels were ignored in the classification and clustering experiments.

Similarly to the previous collection, Python script was used to retrieve only the relevant text of news articles. The resultant text files are identified by their original numeric identifiers given by wienerzeitung.at. The German collection is the smallest one with two and half thousand documents and 1 million tokens.

ID	category GE	category EN	documents
1	Europa	Europe	340
2	Kultur	culture	1029
3	Österreich	Austria	103
4	Sport	sport	324
5	Wahlen	elections	146
6	Welt	world	237
7	Wien	Vienna	328
8	Wirtschaft	economy	71
total			**2578**[61]

Table 6: The categories of the collection GE1

6.3.1.1.4 Extraction of collection CZ2

The additional Czech experimental collection is a subset of the collection CZ1 (chapter 6.3.1.1.1). It consists of selected documents from the category of domestic news. Approximately 12% of documents were automatically translated using Microsoft translator API service [62]. The documents were translated from Czech to English and then back from English to Czech. Hence all documents in the collection are in Czech, but some of them are products of the translation engine. The translated documents are labeled by an indicator; the indicator serves as a target attribute for further development of classifiers.

The text files in the collection are identified by their original numeric identifiers given by novinky.cz. The collection in total includes 3 thousand documents with over one million tokens divided into two categories.

ID	category	documents
1	machine translated	384
2	original	2 713
total		**3 097**

Table 7: The categories of the collection CZ2

6.3.1.2 Downloaded document processing pipeline

6.3.1.2.1 Learning pipeline description

A learning pipeline includes the process that develops necessary objects for a subsequent processing pipeline. While the learning pipeline exploits the training set of documents, the purpose of the processing pipeline is to derive an appropriate vector representation of new documents[63]. The learning pipeline does not provide a representation of input documents; the main outputs of the learning pipeline include the vocabulary and the model that substitutes vocabulary items by topics.

[61] Some documents from collection GE1 are assigned to multiple categories hence the total is larger than the number of documents in the collection.
[62] The same translator is also available as Bing browser translator
[63] Training and testing documents are eligible for the processing pipeline as well.

Figure 11: The learning pipeline.

The input to the learning pipeline consists of a collection of crawled text documents stored as local files[64]. If the format of stored documents is different from the raw text file, they are transformed to the raw text format using the selected encoding. In the case of HTML page only the content of relevant tags is kept[65]. The optional document category assignment is not important at the beginning of the process because the primary goal of both pipelines is to represent documents regarding their content and the context only without any further assumptions.

An index database is created for each collection to simplify and speed up the further processing. The index database is an ordered and possibly indexed list of lowercased tokens together with data about their locations within documents. The token location data consist of the identification of the document, the sentence and the position within the sentence where the token occurs. The other properties of tokens such as stems can be added to the index database during the process. The index database includes both training and testing documents, hence it is exploited by the procedures from both learning and processing pipelines.

ID index	ID doc	token	ID sentence	position within sentence	stem	vocabulary item	topic
5099	CZ1_102154	celý	24	2	celý	true	6
5100	CZ1_102154	komplex	24	3	komplex	true	9
5101	CZ1_102154	budov	24	4	bud	true	9
5102	CZ1_102154	přilehlých	24	6	přilehlých	true	9
5103	CZ1_102154	zahrad	24	7	zahrad	true	9
5104	CZ1_102154	získala	24	8	získ	true	6
5105	CZ1_102154	armáda	24	9	armád	true	7

Table 8: The sample from the index table enhanced by the vocabulary flag and the topic assignment.

The index database is built in the beginning of the pipeline where documents are tokenized and simplified by filtering of stop-words. The tokenization comprises breaking documents strings into tokens and sentences (segmentation), alternatively some higher text units such as paragraphs or chapters are recognized in the documents. Sentence boundaries are stored in the index database together with tokens or words. The filtering of stop-words is a simple language dependent step where frequent but meaningless words are filtered out from all documents. The stop-words lists are usually publicly available for common languages and they typically include hundreds of words such as determiners, prepositions or conjunctions. The non-word tokens that stand for numbers, e-mail addresses or URIs are removed as well to retain only the tokens that represent meaningful words. These non-linguistic entities are recognized using the regular expressions.

[64] Documents need not be stored locally; any other storage can be used as well.
[65] It applies also to documents in other formats that include the metadata and/or some irrelevant text.

The last optional step which can influence the index database in the proposed pipeline is the stemming. The stems are results of trimming of prefixes and suffixes of words. The stemming procedure is language dependent, but does not require advanced machine learning approaches. Usually a set of rules is applied to each individual word separately to receive its stem. The rules can have the form of regular expressions. The final stem may not be the same as the linguistic root of the word. However, it does not have the great negative impact on the further processing. The stemmers are available for many natural languages and they usually comprise hundreds of rules. The stemming significantly reduces the number of wordforms that occur in documents. Especially the collections of inflectional languages such as Czech exhibit a huge number of wordforms and the stemming simply enables to significantly reduce their initial dimensionality. Even though the stemming is language dependent procedure, the implementation of rules is relatively easy. The stemming reduces the size of vocabulary at the beginning of the process and brings the considerable reduction of processing requirements for further steps.

The next step is the selection of vocabulary tokens. This step can be bypassed if any external vocabulary is available. The internal vocabulary is extracted from the training documents only; this step is performed in the learning pipeline and is not present in the processing pipeline. The only stems that occur at least in two training documents are incorporated into the vocabulary[66]. The table of selected vocabulary stems also includes *idf* frequencies of stems derived from the training set of documents to further accelerate the possible computation of the *tf-idf* representation for processed documents. The index database is then enhanced by the vocabulary filter because only the vocabulary entries are taken into account in the further processing.

ID token	token	df	gf	idf
61013	šed	31	49	6.300
61014	šeď	8	8	7.655
61015	sedá	58	103	5.674
61016	šedá	42	49	5.997
61017	sedač	155	280	4.691

Table 9: A sample from vocabulary table of stemmed tokens.

The non-vocabulary tokens are omitted from the further processing. The training documents consisting from the vocabulary stemmed tokens are now processed by latent Dirichlet allocation procedure (LDA) (see chapter 3.1.2.4) to learn the model for the topics assignment. LDA performs a huge dimensionality reduction which is necessary if we need to form n-grams that further significantly increase the number of processed features. The number of extracted topics is set in advance[67]. The goal of LDA in the learning pipeline is to develop a topic model that will fluently substitute stems by topics in the processing pipeline. Hence the training documents are exploited to tune all essential global parameters of the model. LDA model and the vocabulary table constitute the main outputs from the learning pipeline that are necessary to form the processing pipeline.

6.3.1.2.2 Processing pipeline description

The purpose of the processing pipeline is to assign a proposed vector representation to new incoming documents[68]. The new incoming documents can be processed independently one by one; the relations among documents do not influence the output of the processing pipeline. In

[66] This parameter is subject to change.
[67] It is modified and investigated within the experimental setup.
[68] Testing and training documents can be vectorized as well.

the experiments the processing pipeline is applied to the test set of documents to further evaluate the appropriateness of the proposed document representations.

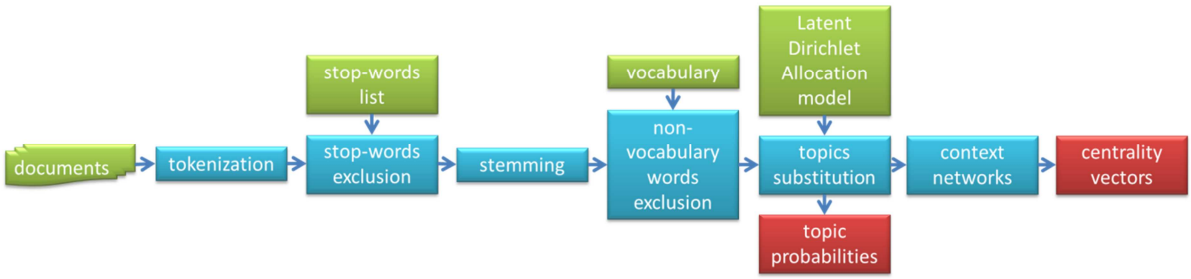


Figure 12: The processing pipeline.

The processing pipeline exploits the objects such as the vocabulary or LDA model for topics assignment that have been developed over the training set of documents in the learning pipeline. The initial preprocessing steps, namely tokenization, stop-word exclusion and stemming are the same in both pipelines[69]. Then the external vocabulary or the internal vocabulary built over the training documents in the learning pipeline is used to filter out non-vocabulary stems. It is necessary to keep the vocabulary stems only in documents because LDA model cannot assign the topics to the unknown stems.

The LDA substitution is the next step, the LDA model that has been derived over the training documents in the learning pipeline is exploited. Each document is regarded by the LDA model as the specific bag-of-topics. The vector of topic probabilities is stored for each document as well because it serves further as a benchmark representation for various comparisons. However, the assignment of topics to stems is our main output from the LDA model. The stems are substituted by the topics for the further processing and the assigned topics are added to the index database.

While all previous steps do not take into account any order of tokens, stems or topics, the building of a context network strongly relies on the adjacency of the topics within a document. For the topic on the particular position in a document a set of its adjacent topics is retrieved. It includes the topics that fall into its context window. The context window consists of token positions in a document that are within the selected vicinity from the investigated position in the document. The length of the context window is set in advance[70] and it reflects the fixed distance between the positions in the document on which the co-occurrence of tokens is supposed to be nonrandom. The order of topics within a particular context window does not matter; all topics in the window are regarded as neighbors of the topic on the position to which the context window is assigned.

The context window can be of three types: left, right or symmetric[71]. The left context window includes the selected number of positions left from the investigated position. The right and symmetric context windows are defined analogically. For the further counting of topics pairs[72] the selection of the context window type has the negligible effect, especially in the texts they are far longer than the length of the context window.

[69] See the learning pipeline description in the previous chapter.
[70] It is another parameter that is modified and investigated through the experiments.
[71] Two-sided asymmetric context windows are not taken into account.
[72] The topic pair consists of the topic on the investigated position and one topic from its context window.

The token positions that were not assigned by a topic should be taken into account as well. They contain non-vocabulary stems[73]. Hence a special topic coded by -1[74] is assigned to such stems. This assignment is not performed by LDA. The other special topics are introduced to reflect starts and ends of text units. The text unit is a part of the document where the context is important. We take into account the context within the sentences in the proposed document representation, hence the context windows include only the positions that belong to the same sentence[75]. The sentences within the document are recognized already in the tokenization step; the affiliation of token positions to sentences is stored in the index database. On the other hand, the start or the end of a sentence can influence the adjacency of the border tokens, hence it is worth including the information about sentence borders into the proposed representation. To do so two artificial positions are added to each sentence. The dummy start position is inserted before each sentence and the dummy end position is inserted after each sentence as well. These two positions are assigned by special topics coded as -3 for starts and -2 for ends. All special topics {-3,-2,-1} are then present in the further representations.

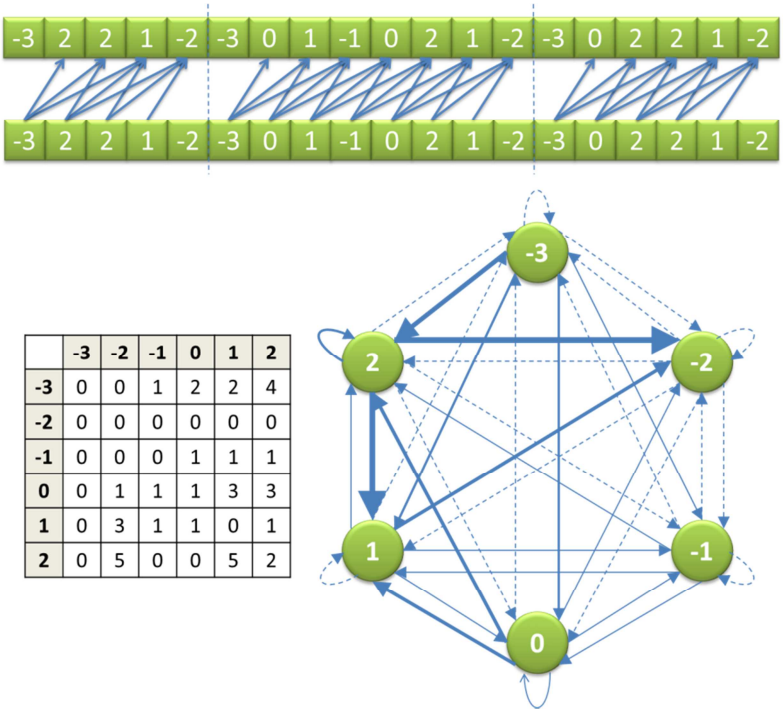
tokenized sentence		Jen	na	blatnících	objevíme	malé	náznaky	křivek	.	
token rank		1	2	3	4	5	6	7	8	
stop-word		Y	Y	N	N	N	N	N	Y	
stem				blatnících	objevím	malé	náznak	křivek		
vocabulary stem				N	Y	Y	Y	Y		
topic	-3			-1	3	3	1	3		-2
topic rank	1			2	3	4	5	6		7

Table 10: The example of the settings of the left context window of the length = 3. The original Czech sentence "Jen na blatnícíh objevíme malé náznaky křivek." is tokenized; the non-word tokens are discarded together with stop-words. The remaining tokens are stemmed. The topic is assigned to each vocabulary stem by the LDA model. Non-vocabulary stems are assigned by the topic coded -1. Special topics coded -2 and -3 are added to the border positions to reflect the start and the end of the sentence. All assigned topics are ranked to set the context windows properly. The context window can be established for each ranked position. The light gray background color highlights the left context window of the original word "křivek".

The identification of the context window for each position in the document enables to construct a context network for the document. The context network of the document is the network with oriented edges. Additionally, a numeric non-negative weight is assigned to each edge of the context network. Vertices of the context network are fixed, they represent the extracted topics[76]. The edge weights represent the counts of topic pairs within the document. The topic pairs are counted for each possible context window within the document and summed together. The process can be also described as the sliding by the context window through the document and continuously counting the topic pairs[77]. Each topic pair consists of the topic on the investigated position and one topic from its context window[78]. The order of topics within the pair is important because it determines the direction of the edge in the context network. All topic pairs from the context window on the particular position within the document include the same topic on the second position when the left context window is used

[73] The positions of stop-words are discarded in the beginning of the pipeline. On the contrary to non-vocabulary stems the stop-words have no impact on further document processing.
[74] The regular topics that are assigned to stems by the used implementation of LDA use numerical codes starting from zero.
[75] The left context windows at the beginning of the sentence are shorter. The same holds for the right context windows at the end of the sentence.
[76] The special topics {-3,-2,-1} are included as well.
[77] More precisely the context widow slides continuously only within the sentences and it jumps over the sentence borders.
[78] The order of topics within particular context window is not taken into account.

and similarly, all topic pairs include the same topic on the first position when the right context window is applied.



	-3	-2	-1	0	1	2
-3	0	0	1	2	2	4
-2	0	0	0	0	0	0
-1	0	0	0	1	1	1
0	0	1	1	1	3	3
1	0	3	1	1	0	1
2	0	5	0	0	5	2

Figure 13: The construction of the context network from the sequence of topics. The document consists of three sentences; the tokens were already substituted by the topics and the sentence borders were marked by the special topics. The topic sequence is copied twice in the picture to see the construction of topics pairs using the left context window of the length of three. Each arrow represents a topic pair. The counts of all topic pairs form the weight matrix of the context network. The arrow thickness in the context network diagram is proportional to the weight. The missing connections with zero weight are dashed.

The sum of all weights in the context network is proportional to the length of the document[79] and to the length of the context window. The distribution of weights within the context network reflects the patterns of topics presented in the document. Hence the context of the document together with the document content is described by the relations among the nodes in the context network. Such relations are determined by the edge weights. The strengths of incoming or outgoing connections of a node reflect the presence of the topic in the text and the distribution of weights within the network reflects the adjacencies of topics that are influenced by the context. Unfortunately, the whole context network is not the appropriate document representation that can be used as an input for mining models because it is described by the multidimensional weight matrix. The mining models expect document vectors as their inputs. However, the main properties of the context network can be captured by centrality statistics.

The centrality statistics of each node is derived from the weight matrix and the statistics is somehow proportional to the importance of the node. The importance can be evaluated using different criteria, hence we can select from the wide range of centralities. The centralities further experimentally investigated include Degree, InDegree, OutDegree, Eigenvalue, Authority, Hub, PageRank, Closeness and Betweenness (see the chapter 4.5). Closeness and

[79] The magnitude of the sum is also influenced by the structure of sentences within the document. Shorter sentences imply the smaller weight sum because smaller number of context windows is present in the document. The sum of weights is also influenced by the presence of stop-words, punctuation and non-word tokens such as numbers or e-mails. These tokens are discarded before the document is indexed.

Betweenness centralities rely on the distances among the nodes rather than on the weights, hence the weights are inverted before Closeness and Betweenness are computed. Other centralities exploit directly the weight matrix.

The centralities of all nodes of the context network form the final document vector. The dimensionality of the proposed document vector is fixed and is determined by the number of extracted topics which is the parameter of LDA model[80]. Some centralities may depend on the total sum of the weights that is proportional to the document length. Therefore the centrality vectors are normalized to have the length of one for some experimental comparisons where the document length should not be taken into account[81].

6.3.2 Generation of and processing of simulated documents

Each simulated document is a sequence of vocabulary items. The size of vocabulary of a natural language is usually large comprising tens of thousands of words, but in the proposed representations we use topics instead of words[82]. Hence it is not necessary to generate original documents that include words in the simulations; we directly generate documents where the words are already substituted by topics. The size of the vocabulary of topics is reasonably lower, it is usually of the order of units or tens.

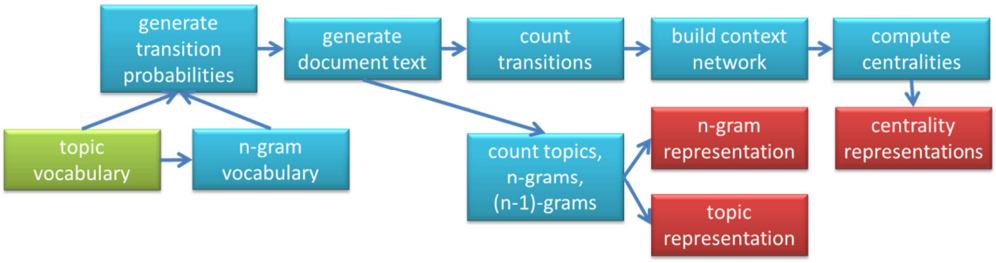


Figure 14: The generation and procesing of the simulated documents.

The simulated documents are sequences of topics. The topic sequences are generated regarding the n-gram probabilities. n-gram is a short sequence of *n* topics. Each document is fully described by its n-gram probabilities. The probabilities of n-grams are arranged in a matrix where rows represent (n-1)-grams and columns represent topics. Such arrangement easily enables to obtain conditional probabilities of the topics knowing a foregoing (n-1)-gram. These conditional probabilities enable to generate the next topic of the sequence knowing last *n*-1 topics.

Each document is generated given the topic vocabulary, the n-gram probabilities and the document length. After the first n-gram is generated, each subsequent topic is added regarding the conditional probabilities of the topics. The process is repeated until the number of topics in the sequence equals the given document length.

Note that n-gram probabilities are known for each simulated document, hence they could be exploited for comparisons. Although we assume that the n-gram probabilities fully describe real documents, they are unobservable and can be only estimated from the observed n-gram

[80] Three special topic {-3,-2,-1} slightly increase the dimensionality.
[81] For example Bayesian classifiers do not rely on the cosine similarity that naturally normalizes the lengths of vectors.
[82] LDA is used with real documents to transform words to topics. It is an initial dimensionality reduction that does not rely on the order of words within a document. A subsequent dimensionality reduction is included in a derivation of centralities of context networks.

frequencies in the document processing. Hence the n-gram probabilities estimated[83] from the n-gram frequencies are used in the performed tests.

After the sequence of topics is generated to its required length, it is further processed the same way as the downloaded documents after their stems are substituted by topics (see the processing pipeline from chapter 6.3.1.2.2). The only simplification is the absence of special topics in the generated sequences. They do not include special topics {-3,-2,-1} because any generated sequence is not divided into sentences or paragraphs and all topics in a generated sequence are vocabulary topics.

6.4 Evaluation of experiments

6.4.1 Document diversity

A dimensionality reduction that transforms the original n-gram document vectors[84] into a different target representation affects distances among documents. The distances among the documents are important for the predictive models that perform standard text mining tasks such as the clustering, the classification or the document retrieval. The relevance of the projection from the original space to the target space relates to the notion of distance preservation. The distance preservation is often an optimization criterion used by general dimensionality reduction methods (Lee & Verleysen, 2007). Any statistics that measures the distance destruction or preservation can be used as the fitness function or for the purpose of evaluation of a projection. The simple example of such a criterion was proposed in (Paukkeri et al., 2011). The nearest neighbor of each data vector is calculated in the original feature space. After the evaluated dimensionality reduction is performed, the nearest neighbors are searched again. The ratio of the preserved neighbors serves as the evaluation measure of the distance preservation.

The well-known example of exploiting of a distance destruction criterion in searching for an optimized non-linear low-dimensional projection is the multidimensional scaling method (MDS). The multidimensional scaling is an exploratory approach that enables to visualize multidimensional data (Cox & Cox, 2000) (Borg & Groenen, 2005). The primary outcome of MDS is a spatial configuration, in which the objects are represented as points in a low-dimensional space arranged in the way that their distances correspond to the similarities of the original objects: similar objects are represented by the points that are close to each other, dissimilar objects by the points that are far apart. The common criterion of the distance destruction used in MDS as the fitness function will serve as the evaluation statistics of the proposed document representation. Such a simple non-negative criterion enables us to express the changes in the vicinity of document vectors when constructing the proposed document representations.

Let the original distances[85] between M objects are given by a symmetric distance matrix **R** of the size MxM, whose diagonal elements equal zero. The projected points to a low-dimensional space form another symmetric distance matrix **S** of the size MxM. We need to express the discrepancy between the matrices **R** and **S**. The classical metric approach (Torgerson, 1952) has fallen from favor and more modern formulations of the metric MDS have introduced two popular discrepancy measures between original and projected distances. The STRESS criterion (Kruskal, 1964), proposed for the nonmetric MDS, is based on the squared errors

[83] Maximal likelihood estimates are applied.
[84] In accordance with statistical language model we believe that n-gram probabilities fully describe each document.
[85] In the MDS theory more general dissimilarities are considered.

between all entries of **R** and **S**. The newer and more popular SSTRESS criterion for the nonmetric MDS (Takane et al., 1977) is based on the squared errors between the squared original and the squared projected distances.

Generally, the objective of MDS is to find a projection that minimizes the following discrepancy criterion.

$$\sum_{i=1}^{M}\sum_{j<i} w_{ij}\left((r_{ij}^{\,2})^k - (s_{ij}^{\,2})^k\right)^2 \rightarrow \min \tag{144}$$

This general formula enables to take into account the importance of particular distances by introduction of the weights w_{ij}[86], permits negative similarities and enables to select between STRESS (k = 1/2) and SSTRESS (k = 1). For the evaluation of the proposed document representations in the experimental part of the thesis, the non-weighted SSTRESS measure is used and **R** and **S** are distance matrices with non-negative entries. Hence the general fitness function (144) simplifies to just SSTRESS evaluation statistics.

$$SSTRESS = \sum_{i=1}^{M}\sum_{j<i}\left(r_{ij}^{\,2} - s_{ij}^{\,2}\right)^2 . \tag{145}$$

The SSRESS measure is non-negative and is not upper-bounded. Its magnitude depends on the number of entries of the distance matrices **R** and **S** and on the scale of distances as well. Hence the correct comparison of different proposed document representation requires the following settings:

- Set the single distance measure for all comparisons.
- Normalize the sizes of document vectors to be on the comparable scale.
- Select the baseline document representation to which all proposed representations will be evaluated.
- Compare only the proposed representations of the same dimensionality.

All these requirements are fulfilled in the experiments. The popular cosine similarity is used as the distance measure. The high-dimensional context network document representation is used as the reference representation that forms the distances in **R** and all proposed centrality vectors that form matrices **S** share the same dimensionality.

The cosine similarity serves as the document distance[87] because it is independent on the scale of document vectors, it includes the normalization of the vectors. It is a common measure used in the document retrieval and other text mining tasks. Only an angle between the document vector pair influences the distance, hence document lengths do not affect the cosine similarity.

[86] The weights are also used to accommodate missing data in MDS.

[87] The cosine similarity does not comply with the usual distance definition. It is rather similarity measure. For example, the cosine similarity of two same documents does not equal zero. But it is the common measure for document comparisons.

$$r(\mathbf{d}_i, \mathbf{d}_j) = \frac{\mathbf{d}_i \bullet \mathbf{d}_j}{\|\mathbf{d}_i\| \|\mathbf{d}_j\|} = \frac{\sum_{k=1}^{N} v_{ik} v_{jk}}{\sqrt{\sum_{k=1}^{N} v_{ik}^2} \sqrt{\sum_{k=1}^{N} v_{jk}^2}} \tag{146}$$

The above formula for the cosine similarity compares two document vectors $\mathbf{d}_i$ and $\mathbf{d}_j$ of N dimensions. If the document vectors include only non-negative components, the cosine similarity takes values between zero and one.

6.4.2 Document classification and retrieval

The classification of documents to the categories that are known in advance is a very common text mining task, hence the evaluation of classifiers is well developed and unified. The document search known as the information retrieval is a special case of the document classification task from the evaluation point of view. We assume that the system of document categories is established and testing documents are labeled by their categories. The information retrieval is then a binomial classification regarding a given query. Each document either satisfies or dissatisfies the query.

Minor complications can be caused by multiple labels when each document can belong to more than one category and a classifier can assign the document to several categories as well (Tsoumakas & Katakis, 2007). It does not appear in the presented simulations, but multiple labels can be found in one downloaded experimental collection. The idea of evaluation of an ambiguous classification is to split a complex classifier into several simple binomial classifiers. Each binomial classifier is evaluated separately and the results are then put together. Two frequently used folding methods are known as micro-averaging and macro-averaging. In the macro-averaging approach, the mean of evaluation statistics of all binomial classifiers is presented as the final evaluation statistics. In the micro-averaging approach, the counts of correct and incorrect predictions of all binomial classifiers are summed together and the evaluation statistics is computed from these sums. The micro-averaging is preferable in the case of unbalanced categories because it naturally comprises category frequencies. These two approaches can be also regarded as a generalization of the evaluation of a binomial classifier to the evaluation of a multinomial classifier; the multiple labels are then only the special case of multinomial classification.

Each simulated document in the presented experiments is a member of just one category; the number of categories is the parameter of the experimental setup. The simulations of the information retrieval and the recognition of permuted documents imply just two categories classification; other simulations are multi-categorical. The experimental collections of downloaded documents have the fixed number of categories; the number of categories is the experimental parameter in the simulations. Each document belongs to just one category, only the German collection of real documents includes several documents with multiple labels. The problem in the German collection is overcome by the duplication of such documents; each copy has just one different label. The duplication is performed after a document is assigned to a training or test set, hence all copies of the same document belong to the same set. Then the multinomial classifier which predicts just one category for each document is developed and evaluated. Such duplication is equivalent to the micro-averaging.

The evaluation of a classifier always starts with the construction of the misclassification matrix[88]. Let us have a set of categories $A=\{a_1,a_2,\ldots,a_{|A|}\}$. Each testing document is labeled by the category x from A and also the document is assigned to the category y from A by the evaluated classifier. Altogether the pair $[x,y]$ is known for each testing document. The misclassification matrix is then a square crosstab that includes counts of the $[x,y]$ pairs of the test documents.

The misclassification matrix of the binomial classifier is the fourfold table of frequencies[89]. However, the misclassification matrix of the multinomial classifier can be also transformed to a set of fourfold tables[90]. One fourfold table is constructed for each category, hence the evaluation is performed over the set of $|A|$ tables. For the category a each classification result represented by the pair $[x,y]$ can be assigned to one of the following four cells: true negative (TN), false positive (FP), false negative (FN), true positive (TP). Positive or negative refers to the assignment to the particular category done by the classifier. The incorrect classifications are referred as false, the correct classifications are referred as true. Formally, while constructing the fourfold table for the category r, each document is assigned to one of four cells using the following schema.

$$\begin{aligned} TN &: x \neq a \wedge x = y \\ FP &: x \neq a \wedge x \neq y \\ FN &: x = a \wedge x \neq y \\ TP &: x = a \wedge x = y \end{aligned} \tag{147}$$

		prediction	
		y≠a	**y=a**
target category	**x≠a**	*TN*	*FP*
	x=a	*FN*	*TP*

Table 11: The fourfold evaluation frequency table for the category r. x represents the document label and y stands for the assigned category.

Over each fourfold evaluation table a large number of evaluation statistics can be constructed. The basic and commonly used statistics are the precision P and the recall R.

$$\begin{aligned} P &= \frac{TP}{FP+TP} \\ R &= \frac{TP}{FN+TP} \end{aligned} \tag{148}$$

From the information retrieval point of view, the precision[91] is the fraction of the retrieved documents that are relevant, while the recall[92] is the fraction of the relevant documents that are retrieved. The precision and the recall are not independent; a binary classifier that exhibits the high precision often suffers by the low recall and vice versa. Hence it is worth combining

[88] The misclassification matrix is sometimes referred as the confusion matrix or the coincidence matrix.
[89] For example the evaluation of the information retrieval directly offers such fourfold table.
[90] Due this fact the evaluation methods that were originally developed for the information retrieval and the binomial classification are applied with minor modifications to the nominal classification documents.
[91] The precision is sometimes referred as the positive predictive value.
[92] The recall is also known as the sensitivity.

these two evaluation measures. The single measure that combines both is their harmonic mean, the traditional F-measure.[93]

$$F = \frac{2PR}{P+R} \tag{149}$$

The precision and the recall are evenly weighted in the harmonic mean, hence the mean is also referred as F_1-measure. F_1-measure is a special case of the more general F_β-measure. β stands for the non-negative real ratio of weights of the precision and the recall in the weighted harmonic mean.

$$F_\beta = \frac{(\beta^2+1)PR}{\beta^2 P+R} \tag{150}$$

Only F_1-measure is used in the evaluation of the presented simulations and experiments. However, the experiments often take the advantage of multinomial classifiers, hence it is necessary to combine the evaluation results of particular binomial classifiers into a main fourfold table where the cells of particular fourfold tables are summed up.

6.4.3 Clustering of documents

A huge effort has been spent to improve algorithms for the document clustering (Aggarwal & Zhai, 2012). However, the evaluation of the clustering outcomes has not been fully standardized, yet. Indeed, the construction of quality measures for unsupervised techniques is not straightforward and it is still an emerging area. Fortunately, the evaluation of the clustering of experimentally simulated documents enables to exploit the information about a distribution of n-grams from which each document was generated. More precisely, documents generated from the same distribution of n-grams are expected to belong to the same cluster. The experimental setup also simplifies our evaluation: the number of n-gram distributions from which a collection is generated is always set to be the same as the number of clusters in the consequent k-means algorithm. Therefore the evaluation is similar to the situation when the documents are labeled with the target category. The labeling information is not taken into account during the development of the clustering model, but it is exploited for the evaluation purposes.

The general objective function in the clustering evaluation formalizes the goal of attaining the high within-cluster similarity of documents together with the low between-cluster similarity. In the situations when the category labels are known the document similarity is measured by similarity of their labels rather than by similarity of their input vectors in the evaluation phase. The idea behind the evaluation statistics is to estimate how the cluster assignment implies the label purity of clusters. The useful evaluation statistics include the purity, the normalized mutual information, the rand index or F-measure[94] (Vinh et al., 2010).

To compute the purity, each cluster is assigned to the category which is the most frequent in the cluster. The accuracy of this assignment is then measured by counting the proportion of correctly assigned documents.

[93] Sometimes referred as the balanced F-score.

[94] The F-measure is constructed the same way as well-known F-measure used for evaluation of information retrieval but the fourfold misclassification matrix is constructed differently as shown later in the text.

$$PR = \frac{1}{M}\sum_{j=1}^{|B|}\mathrm{mod}(b_j) \tag{151}$$

The function $mod(b_j)$ stands for the count of modal labels in the cluster b_j, $|B|$ is the number of clusters, M is the number of documents in the collection. The purity PR takes values between zero and one[95].

The normalized mutual information (NMI) is computed as the mutual information (MI) of two categorical variables: the category label and the assigned cluster. The upper-limit of MI is influenced by the number of categories, hence it makes sense to normalize it to a standard scale[96]. The mutual information $I(A,B)$ between categories $A=\{a_1,a_2,\ldots,a_{|A|}\}$ and clusters $B=\{b_1,b_2,\ldots,b_{|B|}\}$ is computed as

$$I(A,B) = \sum_{i=1}^{|A|}\sum_{j=1}^{|B|} p(a_i,b_j)\ln\left(\frac{p(a_i,b_j)}{p(a_i)p(b_j)}\right). \tag{152}$$

The probabilities in the above formula are maximum likelihood estimates derived as joint resp. marginal relative frequencies of categories and clusters.

MI measures the amount of the information by which our knowledge about the categories increases when we are told what the clusters are and vice versa. The minimum of MI is zero if the clustering is random with respect to the category labels. In that case, knowing that a document is in a particular cluster does not give us any new information about its label. The maximum of MI is reached for the clustering that perfectly recreates the category labels[97]. The maximum of MI is equal to the average of entropies of the investigated variables. Hence the normalization of our MI by dividing by the average entropy fixes the problem since the entropy of the clustering tends to increase with the number of clusters[98]. The entropies of categories and clusters are

$$\begin{aligned} H(A) &= -\sum_{i=1}^{|A|} p(a_i)\ln(p(a_i)) \\ H(B) &= -\sum_{j=1}^{|B|} p(b_j)\ln(p(b_j)) \end{aligned}. \tag{153}$$

The normalization of the mutual information from (152) can be then written as

$$NMI = \frac{I(A,B)}{\frac{H(A)+H(B)}{2}}. \tag{154}$$

[95] If the number of clusters is not fixed, the purity will not enable to compare solutions with the different number of clusters correctly. Small clusters tend to exhibit the higher purity.

[96] The non-normalized mutual information exhibits the same problem as the purity. It does not penalize large cardinalities and thus does not formalize the bias that, other things being equal, fewer clusters are better.

[97] The maximum of MI is also reached, if such clusters in are further subdivided into the smaller ones.

[98] The maximum entropy of the categorical variable with $|B|$ categories equals $\ln(|B|)$.

NMI is then always a number between 0 and 1. Because NMI is normalized, it can be used to compare the experiments with the different numbers of clusters.

An alternative approach to the evaluation of clustering is to view it as a series of decisions, one for each of the $M(M\text{-}1)/2$ pairs of documents in the collection. We want to assign two documents to the same cluster if they share the same label. A true positive (TP) decision assigns two documents from the same category to the same cluster; a true negative (TN) decision assigns two documents from different categories to different clusters. There are also two types of errors we can commit. A false positive (FP) decision assigns two documents from different categories to the same cluster and a false negative (FN) decision assigns two documents from the same category to different clusters. Putting all together we receive the standard fourfold misclassification matrix with total count

$$TN + FP + FN + TP = \frac{M\left(M-1\right)}{2}. \tag{155}$$

Let f_{ij} stands for the frequency of documents that belong to a category a_i, i=1,2,…,$|A|$, and were assigned to the cluster b_j, j=1,2,…,$|B|$. The marginal frequencies of documents are then marked as f_{i+} (categories), f_{+j} (clusters) and f_{++} (total). The basic relations among the document frequencies and the document pair frequencies from the fourfold misclassification matrix are then

$$\begin{aligned} TN + FP + FN + TP &= \binom{f_{++}}{2} \\ FN + TP &= \sum_{i=1}^{P}\binom{f_{i+}}{2} \\ FP + TP &= \sum_{j=1}^{Q}\binom{f_{+j}}{2} \\ TP &= \sum_{i=1}^{P}\sum_{j=1}^{Q}\binom{f_{ij}}{2} \end{aligned} \quad . \tag{156}$$

The other cell frequencies as well as the marginal frequencies in the misclassification matrix result from these equations:

$$\begin{aligned} TN + FP &= \frac{1}{2}\sum_{i=1}^{P} f_{i+}\left(f_{++} - f_{i+}\right) \\ TN + FN &= \frac{1}{2}\sum_{j=1}^{Q} f_{+j}\left(f_{++} - f_{+j}\right) \\ FN &= \frac{1}{2}\sum_{i=1}^{P}\sum_{j=1}^{Q} f_{ij}\left(f_{i+} - f_{ij}\right) \\ FP &= \frac{1}{2}\sum_{i=1}^{P}\sum_{j=1}^{Q} f_{ij}\left(f_{+j} - f_{ij}\right) \\ TN &= \frac{1}{2}\sum_{i=1}^{P}\sum_{j=1}^{Q} f_{ij}\left(f_{++} - f_{i+} - f_{+j} + f_{ij}\right) \end{aligned} \quad . \tag{157}$$

Knowing the frequencies in the misclassification matrix of the document pairs, we can use the standard evaluation measures known from the information retrieval or the document clustering tasks.

The rand index (RI) of the clustering is an analogy to the absolute accuracy:

$$RI = \frac{TN + TP}{TN + FP + FN + TP}. \tag{158}$$

F-measure of the clustering is than the harmonic mean of the precision and the recall that implies the standard formula

$$\begin{aligned} F &= \frac{2PR}{P+R} \\ P &= \frac{TP}{FP+TP} \\ R &= \frac{TP}{FN+TP} \end{aligned}. \tag{159}$$

Note that even though F-measure for the clustering is computed using the same formula as F-measure for the document classification or F-measure for the information retrieval, these three F-measures are derived using the different number of total counts. The total count of the information retrieval F-measure equals the number of testing documents. The total count of the nominal classification F-measure equals the number of testing documents multiplied by the number of categories. And finally, the last total count of the clustering F-measure equals the number of all possible testing document pairs. Hence although any F-measure takes values between zero and one, we should always distinct among these measures; the results among the different text mining tasks should not be mutually compared.

6.5 Experimental assessment of benefits of context encoding

6.5.1 Experimental setup for estimation of information reduction

To explore how an original diversity of documents is affected by the proposed representations, the distances among documents have to be computed. The document distances in the original n-gram feature space are compared with the distances in the proposed feature space. The original distances among documents should be preserved as much as possible using the proposed representations. The full distance preservation can be hardly reached because the dimensionality of the proposed feature spaces is significantly lower than the original dimensionality of the n-gram representation. SSTRESS measure (145) is used as an evaluation score that describes the preservation of cosine distances (146).

To compare the diversity of M documents $M*(M-1)/2$, distances have to be computed for the collection; the number of possible distances grows quadratically with the size of the document collection. Therefore generated collections of reasonable sizes were examined; the maximal collection size of 100 generated documents implies 4450 cosine similarities for each representation.

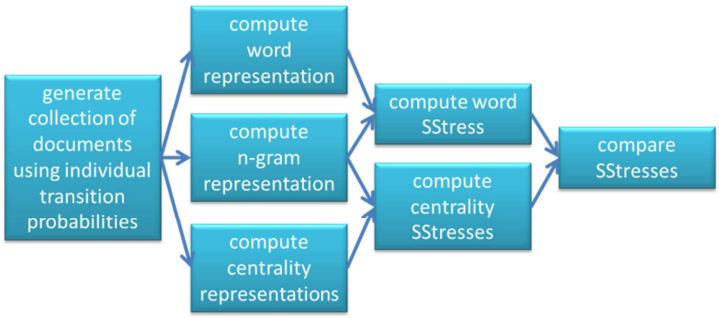


Figure 15: The estimation of the the reduction of the diversity of the simulated documents.

The collection sizes together with other parameters of the experimental setup concerning the generated collections are depicted in the following table. Each combination of the parameters was evaluated ten times; a different random seed was used for each run to generate the collection documents. The repeated evaluations were used to estimate the significance of the difference between SSTRESS scores between the proposed and a standard representations.

parameter	**values**
topic vocabulary size	2, 3, 5, 10
length of left context window	1, 2, 4
number of documents	10, 20, 50, 100
document length	10, 50, 100, 1000
number of repetitions	10
total number of collections	1 920
tested representations	*(n-1)-gram*, *bag-of-topics*, Authority, Betweenness, Closeness, Degree, Eigenvector, Hub, InDegree, OutDegree, PageRank
baseline representation	n-gram
total number of experiments	21 120

Table 12: The tested values of simulation parameters in the task of exploration of diversity reduction using generated collections. Bag-of-topics and (n-1)-gram representations serve as benchmarks for comparisons with the centrality representations.

The n-gram representation is considered as an original representation to compute SSTRESS for each tested representation; it is the vector of the estimated n-gram probabilities for a document[99]. The bag-of-topics representation served as the standard representation for the comparisons; it is a vector of the estimated topic probabilities for a document[100]. Hence SSTRESS of the bag-of-topics representation is compared with SSTRESS of the proposed representations to evaluate the appropriateness of the proposed representations.

[99] The dimensionality of the n-gram representation is significantly higher than the dimensionality of all proposed representations.
[100] The bag-of-topics representation is of the same dimensionality as all proposed representations.

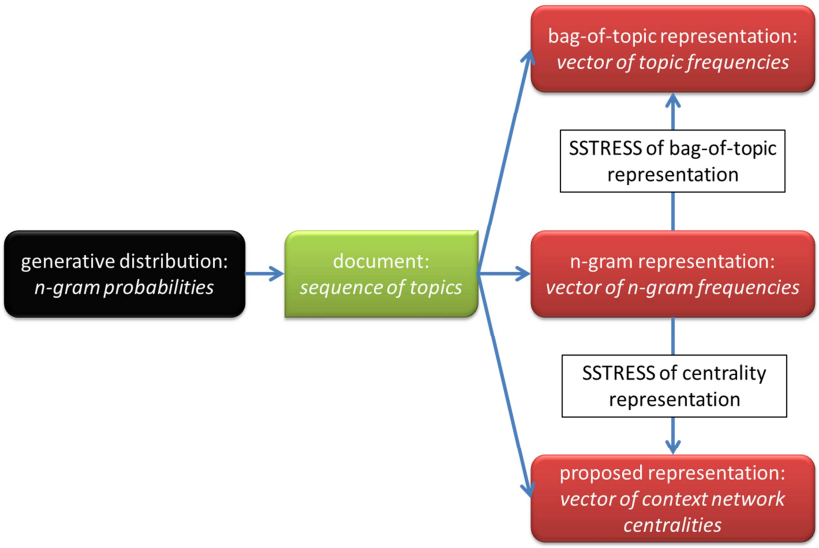


Figure 16: The computation of SSTRESS scores that are used to compare the loss of diversity.

The experimental setup for experiments with downloaded collections is similar to the setup of experiments with artificially generated documents. The change of the diversity of documents is explored in an unsupervised way using SSTRESS measure and cosine distances.

The comparison of the diversity in the whole downloaded collections is too resource consuming because the number of distances among documents increases quadratically with the size of the collection. Therefore the random sample of 100 documents from the test set[101] was used for the comparisons in each experimental run that implies the processing of 4450 cosine similarities for each representation.

The parameters of the experimental setup for downloaded collections are depicted in the following table. Each combination of the parameters was evaluated five times; a different random seed was used for each run to select the different sample of documents from each collection[102]. The repeated evaluations were used to estimate the significance of the difference between SSTRESS scores of the proposed and the standard representations.

parameter	**values**
collection	GE1, EN1, CZ1
topic vocabulary size	5, 10, 20, 50, 100
length of left context window	1, 2, 4, 9
tested representations	*bag-of-topics*, Authority, Betweenness, Closeness, Degree, Eigenvector, Hub, InDegree, OutDegree, PageRank
baseline representations	n-gram, bag-of-topics
number of repetitions	5
total number of experiments	5700[103]

Table 13: The values of evaluation parameters in the task of exploration of diversity reduction using the downloaded collections. The bag-of-topics representation serves as the benchmark for comparisons with the centrality representations.

Again the n-gram representation is considered as the original generative representation[104] of real documents to compute SSTRESS for each tested representation; it is a vector of estimated

[101] The train set of documents that includes approximately 70% of the documents is used to select the vocabulary tokens and to learn LDA.
[102] The random seed also influences the partitioning of a collection.
[103] It does not make sense to evaluate the projection from bag-of-topics to bag-of-topics.

n-gram probabilities in a document. The bag-of-topics representation served as the standard representation for the comparisons; SSTRESS of the bag-of-topics representation is compared with SSTRESS of the proposed representations to evaluate the overall appropriateness of the proposed representations.

6.5.2 Preservation of document diversity in generated collections

The maintenance of the document diversity is reported as the percentage improvement of SSTRESS (145). The original document diversity is expressed by the proximities among the documents in the space of n-grams. SSTRESS measures how the proposed vectorizations of the context network affect the document proximities. Smaller SSTRESS means the better preservation of the proximities in the proposed representation. The base value for the presented percentages is SSTRESS of the bag-of-topic representation. Positive percentages correspond to the SSTRESS decrease and vice versa.

The selection of the centrality that is used as a simplification of the context network is critical. The centrality determines how the content information and the context information are mixed into the final document vectors. The original document diversity is primarily caused by the contextual adjacency of topics. The centrality measure that encodes the context topic relation well implies better SSTRESS. The most promising centralities include Authority, Hub and Eigenvector. These centralities are based on complex relations within the wider neighborhood of the topic in the context network. On the other hand, the centralities that are proportional mainly to the topic frequencies, which means that they encode primarily the document content, do not usually perform better than the benchmark bag-of-topic representation. They include Degree, OutDegree and Indegree. Namely InDegree implies the same SSTRESS as the benchmark representation[105] because it is equal to the topic frequency times the length of the context window[106].

Represen-tation	Mean of SSTRESS difference (%)	Sig.	Sign scheme
Authority	2.5	<0.01	+++
Betweenness	***84.6***	<0.01	+++
Closeness	***-14.9***	<0.01	---
Degree	-0.1	<0.01	--
Eigenvector	1.5	<0.01	+++
Hub	2.4	<0.01	+++
InDegree	0.0	1.00	o
OutDegree	-0.1	0.21	o
PageRank	-6.7	<0.01	---

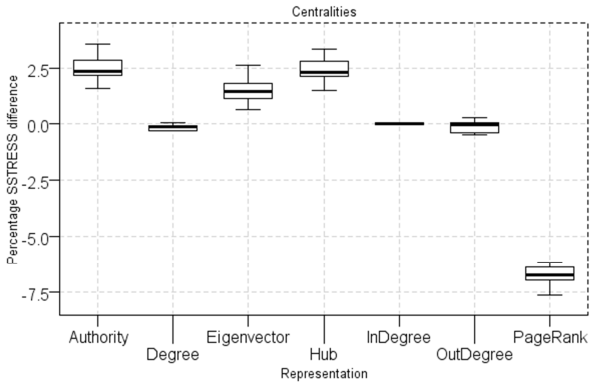


Figure 17: The SSTRESS change by the representation
Closeness and Betweenness are excluded from the graph because of the magnitude of their SSTRESS difference
vocabulary size = 5, n-gram length = 3, number of documents = 100, document length = 100

The centralities that rely on the path lengths should be used very carefully. While Closeness was nearly always significantly worse than the benchmark, Betweenness often performed best

[104] The bag-of topic representation was used as the original representation as well but the results are not presented here because they may confuse the reader. The bag-of topic representation is used as the standard representation instead in the comparisons of SSTRESS. See Figure 16.
[105] It holds for the left context window. If the right context window is used then OutDegree implies the same SSTRESS as the bag-of-topic representation.
[106] The minor differences can be caused by handling the starts of documents.

among all other centralities. The improvement of Betweenness is often incomparable better, but its success strongly depends on the experimental setup. Betweenness is very unsuitable when the vocabulary size is small and when the context window is short.

Represen-tation	Vocabulary size	n-gram length	Mean of SSTRESS difference (%)	Sig.	Sign scheme
Between-ness	2	2	-524.6	<0.01	---
	2	3	-82.8	<0.01	---
	2	5	64.1	<0.01	+++
	3	2	-93.7	<0.01	---
	3	3	44.5	<0.01	+++
	3	5	96.6	<0.01	+++
	5	2	58.4	<0.01	+++
	5	3	84.6	<0.01	+++
	5	5	87.6	<0.01	+++
	10	2	82.9	<0.01	+++
	10	3	84.8	<0.01	+++
	10	5	89.9	<0.01	+++

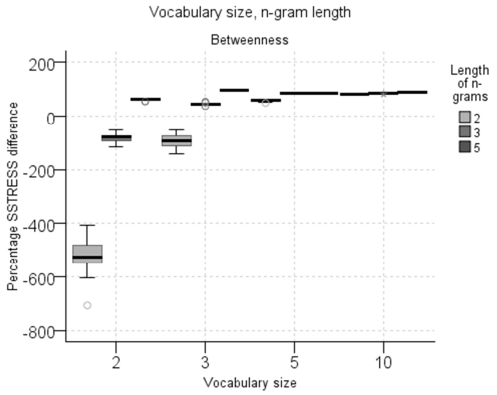


Figure 18: The SSTRESS change by the vocabulary size and the n-gram length in the Betweenness representation
number of documents = 100, document length = 100

PageRank centrality also relies on the possible paths through the context network. Similarly to Closeness, its observed performance was nearly always worse than the benchmark. The dependence of SSTRESS improvement on the setup parameters is very similar for PageRank and Closeness. They exhibit better results for longer documents and for larger context windows, but they are never better than the benchmark.

Represen-tation	Document length	Mean of SSTRESS difference (%)	Sig.	Sign scheme
PageRank	10	-18.27	<0.01	---
	50	-9.55	<0.01	---
	100	-6.74	<0.01	---
	1000	-3.72	<0.01	---

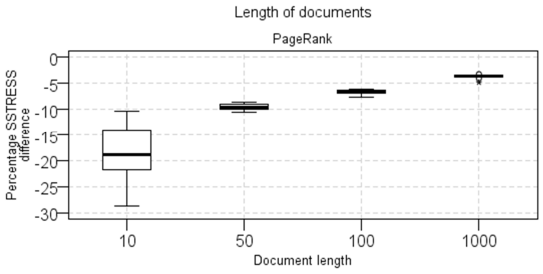


Figure 19: The SSTRESS change by the document length in the PageRank representation
vocabulary size = 5, n-gram length = 3, number of documents = 100

Represen-tation	n-gram length	Mean of SSTRESS difference (%)	Sig.	Sign scheme
Closeness	2	-57.06	<0.01	---
	3	-14.87	<0.01	---
	5	-7.08	<0.01	---

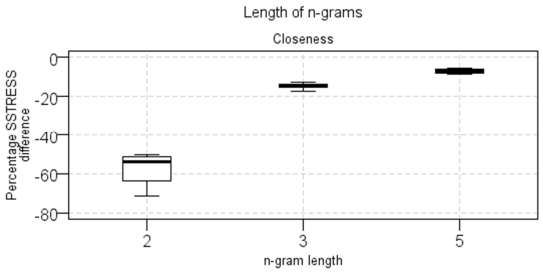


Figure 20: The SSTRESS change by the n-gram length in the Closeness representation
vocabulary size = 5, number of documents = 100, document length = 100

Even though Betweenness can significantly outperform all other centralities, the recommended centrality for preservation of the diversity of documents would be Authority. The usefulness of Betweenness is too variable; it is necessary to test carefully its performance

for particular data and setup. On the contrary, Authority does not improve SSTRESS so significantly, but it is the safe centrality because its performance was better than the benchmark for all experimental setups. The most promising results for Authority were obtained for shorter documents and using the short context window.

Represen-tation	Vocabulary size	n-gram length	Mean of SSTRESS difference (%)	Sig.	Sign scheme
Authority	2	2	14.5	<0.01	+++
	2	3	6.2	<0.01	+++
	2	5	2.6	<0.01	+++
	3	2	20.3	<0.01	+++
	3	3	4.5	<0.01	+++
	3	5	0.2	<0.01	++
	5	2	19.9	<0.01	+++
	5	3	2.5	<0.01	+++
	5	5	0.1	0.03	+
	10	2	19.6	<0.01	+++
	10	3	3.5	<0.01	+++
	10	5	1.2	<0.01	+++

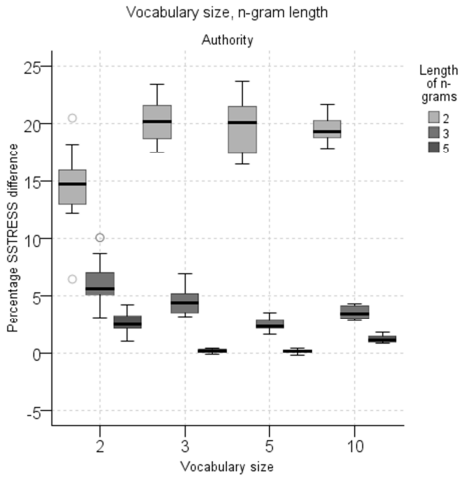


Figure 21: The SSTRESS change by the vocabulary size and the n-gram length in the Authority representation
number of documents = 100, document length = 100

Represen-tation	Number of documents	Document length	Mean of SSTRESS difference (%)	Sig.	Sign scheme
Authority	10	10	10.8	<0.01	++
	10	50	3.1	<0.01	++
	10	100	1.9	<0.01	+++
	10	1000	0.6	<0.01	+++
	20	10	13.7	<0.01	+++
	20	50	2.8	<0.01	+++
	20	100	1.8	<0.01	+++
	20	1000	0.9	<0.01	+++
	50	10	10.5	<0.01	+++
	50	50	2.9	<0.01	+++
	50	100	2.3	<0.01	+++
	50	1000	0.8	<0.01	+++
	100	10	10.9	<0.01	+++
	100	50	3.6	<0.01	+++
	100	100	2.5	<0.01	+++
	100	1000	1.1	<0.01	+++

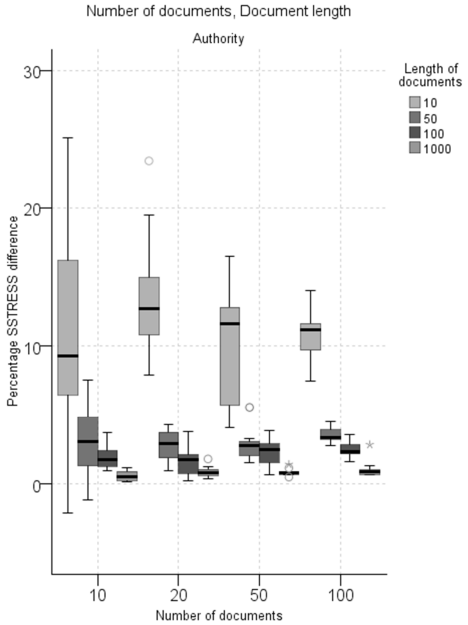


Figure 22: The SSTRESS change by the number of documents and the document length in the Authority representation
vocabulary size = 5, n-gram length = 3

6.5.3 Preservation of document diversity in downloaded collections

The preservation of the document diversity within the downloaded collection is again reported by the percentage improvement of SSTRESS (145). The document diversity is characterized by the proximities among the documents. SSTRESS measures how the proposed vectorization of the context networks affects the document proximities. Smaller SSTRESS means the better preservation. The base value for the presented percentages is SSTRESS of the bag-of-topic representation. Positive percentages correspond to the SSTRESS decrease and vice versa.

The centrality significantly influences the preservation of the diversity. Seven out of nine proposed representations significantly perform better than the benchmark bag-of-topic representation; they encode the contextual topic relations well. Only PageRank and Closeness centralities worsen SSTRESS. Especially Closeness is undesirable one; its SSTERSS difference is negative in all experimental setups. The centralities that are based on the complex contextual relations, namely Authority, Hub and Eigenvector, perform slightly better than the centralities that are proportional to the adjacent topic frequencies, namely Degree, OutDegree and Indegree.

Represen-tation	Collec-tion	Mean of SSTRESS difference (%)	Sig.	Sign scheme
Authority	CZ1	42.321	0.03	+
	EN1	44.295	0.03	+
	GE1	34.927	0.03	+
Between-ness	CZ1	54.723	0.03	+
	EN1	68.986	0.03	+
	GE1	73.916	0.03	+
Closeness	CZ1	***-429.887***	0.03	-
	EN1	***-222.450***	0.03	-
	GE1	***-138.639***	0.03	-
Degree	CZ1	40.534	0.03	+
	EN1	42.767	0.03	+
	GE1	33.493	0.03	+
Eigen-vector	CZ1	42.849	0.03	+
	EN1	44.418	0.03	+
	GE1	36.566	0.03	+
Hub	CZ1	43.098	0.03	+
	EN1	44.987	0.03	+
	GE1	36.402	0.03	+
InDegree	CZ1	38.433	0.03	+
	EN1	41.111	0.03	+
	GE1	30.785	0.03	+
OutDegree	CZ1	39.050	0.03	+
	EN1	41.683	0.03	+
	GE1	31.923	0.03	+
PageRank	CZ1	-10.859	0.03	-
	EN1	5.283	0.03	+
	GE1	-11.234	0.03	-

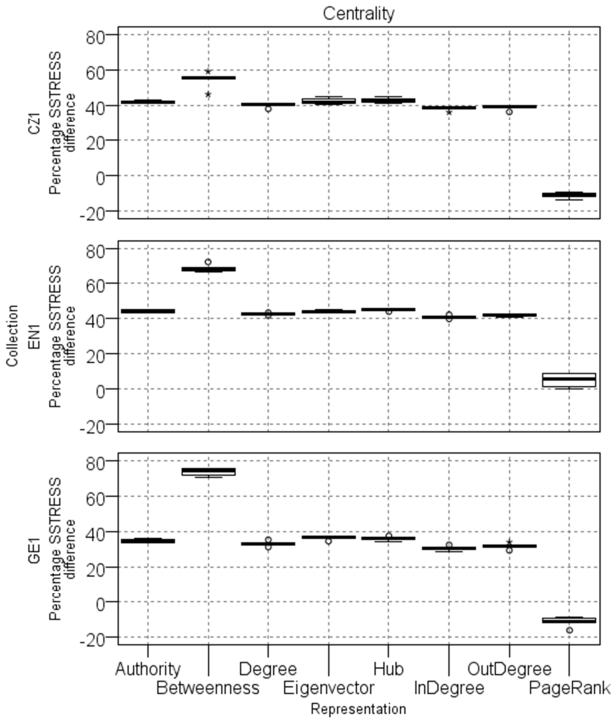


Figure 23: The SSTRESS change by the representation
Closeness is excluded from the graph because of the magnitude of its SSTRESS difference
vocabulary size = 5, n-gram length = 3

The centralities that exploit the path lengths within the context network exhibit a large variability of SSTRESS improvement depending on the experimental setup. While Closeness is always significantly worse than the benchmark, Betweenness often performed best among all other centralities. The improvement of Betweenness is often much better than the improvement of other centralities, but its deployment must be carefully revised considering other parameters like the vocabulary size, the size of the collection and the language.

Represen-tation	Collec-tion	Vocabulary size	Mean of SSTRESS difference (%)	Sig.	Sign scheme
Between-ness	CZ1	5	54.7	0.03	+
		10	51.9	0.03	+
		20	41.2	0.03	+
		50	29.4	0.03	+
		100	7.4	0.03	+
	EN1	5	69.0	0.03	+
		10	67.6	0.03	+
		20	61.7	0.03	+
		50	43.5	0.03	+
		100	17.8	0.03	+
	GE1	5	73.9	0.03	+
		10	67.7	0.03	+
		20	19.4	0.03	+
		50	-244.6	0.03	-
		100	-399.5	0.03	-

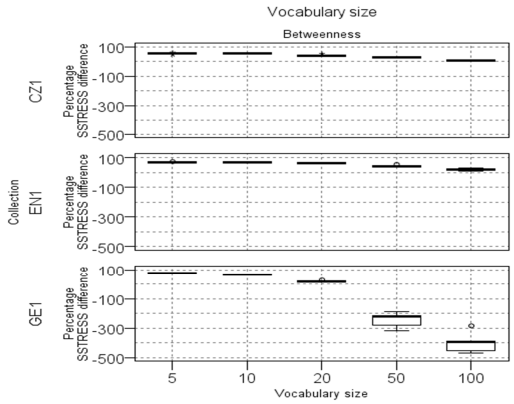


Figure 24: The SSTRESS change by the vocabulary size in the Betweenness representation n-gram length = 3

PageRank centrality exhibits the similar behavior as Closeness; its observed performance was mostly worse than the benchmark. PageRank is the inappropriate option, especially when the vocabulary size is large.

Represen-tation	Collec-tion	Vocabulary size	Mean of SSTRESS difference (%)	Sig.	Sign scheme
PageRank	CZ1	5	-10.9	0.03	-
		10	-23.4	0.03	-
		20	-50.4	0.03	-
		50	-161.0	0.03	-
		100	-300.8	0.03	-
	EN1	5	5.3	0.03	+
		10	-6.0	0.06	o
		20	-30.1	0.03	-
		50	-133.5	0.03	-
		100	-396.9	0.03	-
	GE1	5	-11.2	0.03	-
		10	-43.0	0.03	-
		20	-123.4	0.03	-
		50	-498.9	0.03	-
		100	-799.8	0.03	-

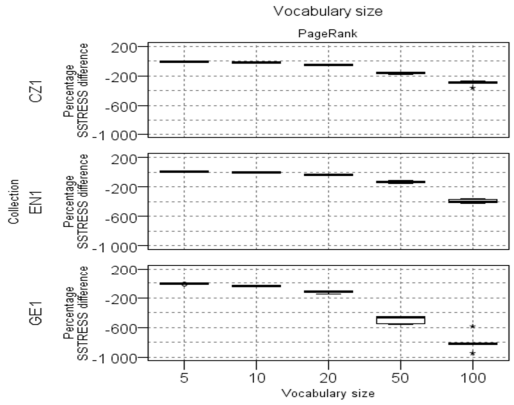


Figure 25: The SSTRESS change by the vocabulary size in the PageRank representation n-gram length = 3

The vocabulary size influences SSTRESS of all representations. Larger vocabulary sizes worsen SSTRESS with the exception of Closeness; the dependence between SSTRESS and the vocabulary size is not monotonic for Closeness. For other centralities the steepness of the decrease of SSTRESS difference by the vocabulary size is language dependent. The largest decrease was observed for the German data; the vocabulary size of the order of tens implies even worse performance than the bag-of-words representation on the German collection.

Represen-tation	Collec-tion	Vocabulary size	Mean of SSTRESS difference (%)	Sig.	Sign scheme
Eigen-vector	CZ1	5	42.8	0.03	+
		10	34.8	0.03	+
		20	23.8	0.03	+
		50	-4.9	0.03	-
		100	-35.6	0.03	-
	EN1	5	44.4	0.03	+
		10	44.6	0.03	+
		20	39.8	0.03	+
		50	17.6	0.03	+
		100	-20.4	0.03	-
	GE1	5	36.6	0.03	+
		10	22.0	0.03	+
		20	-17.3	0.03	-
		50	-198.9	0.03	-
		100	-314.6	0.03	-

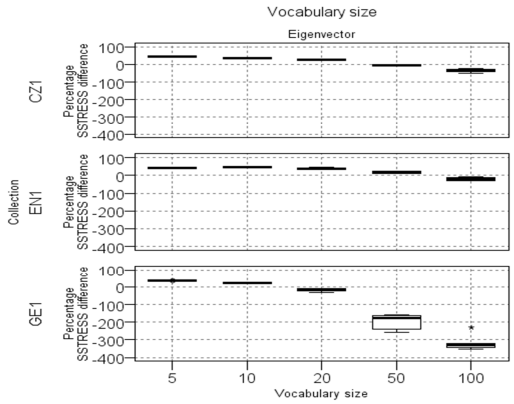


Figure 26: The SSTRESS change by the vocabulary size in the Eigenvector representation n-gram length = 3

Represen-tation	Collec-tion	Vocabulary size	Mean of SSTRESS difference (%)	Sig.	Sign scheme
Closeness	CZ1	5	-429.9	0.03	-
		10	-763.9	0.03	-
		20	-1131.1	0.03	-
		50	-966.8	0.03	-
		100	-447.9	0.03	-
	EN1	5	-222.5	0.03	-
		10	-390.7	0.03	-
		20	-640.8	0.03	-
		50	-916.5	0.03	-
		100	-690.3	0.03	-
	GE1	5	-138.6	0.03	-
		10	-275.0	0.03	-
		20	-458.8	0.03	-
		50	-480.2	0.03	-
		100	-146.3	0.03	-

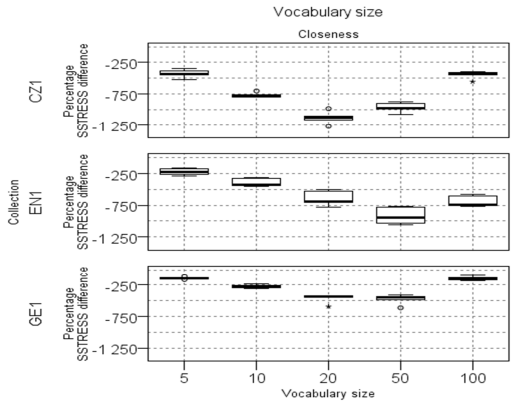


Figure 27: The SSTRESS change by the vocabulary size in the Closeness representation n-gram length = 3

The length of the context window influences SSTRESS as well. The larger context implies the worse SSTRESS with the exception of PageRank and Closeness. Closeness performs always badly, but its performance improves for larger context windows. The dependence between SSTRESS of PageRank and the length of the context window is language dependent.

Represen-tation	Collec-tion	n-gram length	Mean of SSTRESS difference (%)	Sig.	Sign scheme
Eigen-vector	CZ1	2	63.9	0.03	+
		3	42.8	0.03	+
		5	25.2	0.03	+
		10	18.1	0.03	+
	EN1	2	64.5	0.03	+
		3	44.4	0.03	+
		5	30.0	0.03	+
		10	26.2	0.03	+
	GE1	2	59.3	0.03	+
		3	36.6	0.03	+
		5	26.6	0.03	+
		10	25.2	0.03	+

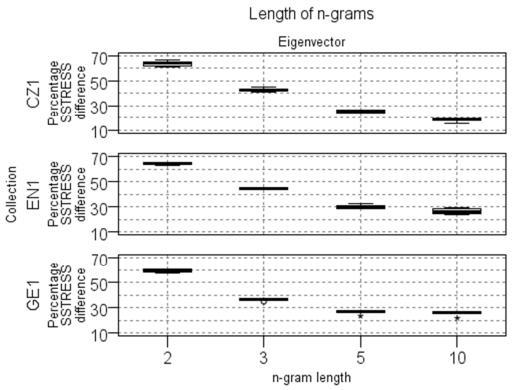


Figure 28: The SSTRESS change by the n-gram length in the Eigenvector representation vocabulary size = 5

Represen-tation	Collec-tion	n-gram length	Mean of SSTRESS difference (%)	Sig.	Sign scheme
Closeness	CZ1	2	-769.3	0.03	-
		3	-429.9	0.03	-
		5	-234.7	0.03	-
		10	-172.6	0.03	-
	EN1	2	-374.8	0.03	-
		3	-222.5	0.03	-
		5	-144.5	0.03	-
		10	-124.8	0.03	-
	GE1	2	-280.8	0.03	-
		3	-138.6	0.03	-
		5	-92.1	0.03	-
		10	-84.9	0.03	-

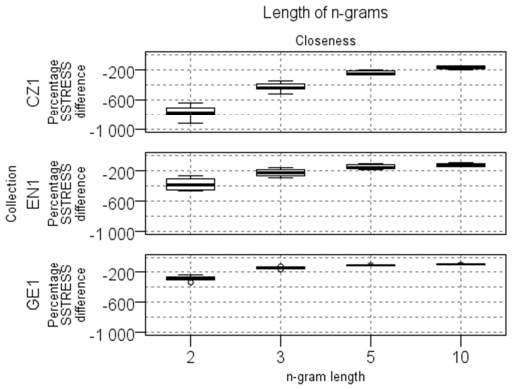


Figure 29: The SSTRESS change by the n-gram length in the Closeness representation vocabulary size = 5

The general improvement of STRESS varies depending on the collection. The most of seven successful representations performed best on the English data and worst on the German data. Only Betweenness, which is the trickiest representation, exhibits the best average results on the German collection and it is the worst on the Czech collection. Such language dependency probably reflects contextual grammar rules that are different in the tested languages. The relations among SSTRESS and values of the experimental parameters are language dependent as well. The strongest relation among SSTRESS and experimental parameters was observed on the German collection; especially the decrease of the performance by the vocabulary size is much steeper than for other collections.

6.5.4 Experimental setup for recognition of permuted documents

The goal of this simple binary classification task is to recognize documents that do not follow any n-gram distribution. The recognition of the generating n-gram distribution is possible only if any contextual information about a topic order is present in document vectors. Therefore the classification experiments should prove that the proposed representations contain such contextual information in a way that it can be exploited by classifiers. These experiments are conducted on generated documents only because they fully comply with the assumption about the n-gram generative process.

Initially, collection documents are generated following the known n-gram distribution. Then topics are randomly permuted in a half of documents; the permuted documents are labeled by

an indicator. The permutation does not influence document topic frequencies, hence the baseline bag-of-topic representation remains unchanged. A consequent Bayes classifier is built to recognize the permuted documents based on the presented input document vectors.

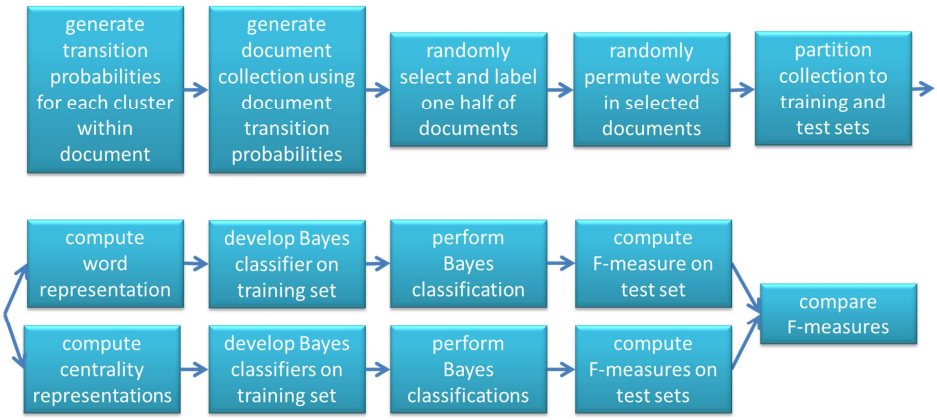


Figure 30: The evaluation of the classification of the permuted documents on the simulated collections.

The classification could be generally performed by other classifiers as well. However, the purpose of the experiment is not to recommend the best algorithm, but to prove that the proposed representations carry the contextual information that can be exploited by classifiers. The Bayesian classifiers are then evaluated using unweighted F-measure that is defined as the harmonic mean of precision and recall (149).

Bayesian classifiers were learned over a training set of documents in each collection and evaluated over a test set. The parameters of the experimental setup are depicted in the following table. Each combination of the parameters was evaluated on ten collections; a different random seed was used for each run to generate collection documents and to split it into training and test sets in the ratio 50:50. The repeated evaluations were used to estimate the statistical significance of the differences between F-measures using the proposed and the standard representations.

parameter	**values**
topic vocabulary size	2, 3, 5, 10
length of left context window	1, 2, 4
number of documents	1000
document length	10, 50, 100, 1000
number of generating clusters	3, 5, 10
number of repetitions	10
total number of collections	1 440
tested representations	*n-gram*, *(n-1)-gram*, *bag-of-topics*, Authority, Betweenness, Closeness, Degree, Eigenvector, Hub, InDegree, OutDegree, PageRank
total number of experiments	17 280

Table 14: The tested values of simulation parameters in the classification of the permuted documents. Bag-of-topics, n-gram and (n-1)-gram representations serve as benchmarks for comparisons with the centrality representations.

The bag-of-topics representation was selected as the standard representation for consequent comparisons. The differences between F-measures of classifiers using the proposed and the standard representations are examined to evaluate the ability of the proposed centrality representations to encode the useful contextual information for the classification.

6.5.5 Classification of permuted documents

The quality of the distinction of permuted documents and original documents that were generated following given n-gram distributions is evaluated by the unweighted F-measure (149) that combines the precision and the recall as their harmonic mean. The F-measure achieved for the proposed representations is compared to the F-measure achieved for the benchmark bag-of-topic representation and their difference is reported. Positive values imply that the proposed representation performs better than the benchmark and vice versa. The benchmark bag-of-topic representation is not affected by the order of topics within a document, hence classifiers built over the benchmark are not able to recognize the permuted documents at all. Therefore we make comparisons with a random prediction in these experiments.

The simulations confirmed our presumption that the centrality selection is critical for the distinction of permuted documents. The performances of the centralities that rely on path lengths within a context network (here Betweenness and Closeness) are significantly better than the performance of the benchmark representation. Betweenness seems to be clearly the best choice, but it will be evident from the next analyses that results for Betweenness vary depending on experimental parameters, hence its selection should be considered carefully.

On the other hand, the centralities that are based on pure sums of weights of incoming and/or outcoming edges of a particular node within a context network (degrees) are not able to outperform the benchmark. It means that they do not offer any information useful for recognizing the permuted documents. However, such results are expected because if a text is not divided into smaller contextual units such as sentences these centralities are proportional to the topic counts in our simulations. These conclusions evidently may not be affected by changing of experimental parameters.

Other presented centralities rely on wider contextual ties within context networks. Except PageRank they influence the results of the classification. However, their contribution is not as great as the contribution of Betweenness or Closeneess. Their performances depend on specific data and experimental parameters; for default values of the experimental parameters nearly any improvement is observed except Authority. Generally, Authority is the most promising centrality from this group working well for most combinations of the experimental parameters.

Representation	Mean F-measure difference	F sig.	F sign scheme
Authority	0.056	<0.01	++
Betweenness	0.312	<0.01	+++
Closeness	0.272	<0.01	+++
Degree	-0.001	0.21	**o**
Eigenvector	0.033	0.09	**o**
Hub	0.039	0.05	**o**
InDegree	0.000	1.00	**o**
OutDegree	0.002	0.17	**o**
PageRank	-0.005	0.27	**o**

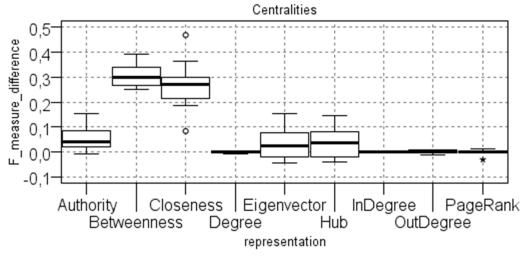


Figure 31: F-measure change by representation, the centrality selection is a critacal issue vocabulary size = 5, n-gram length = 3, document length = 1000, clusters = 3

The distinction of permuted documents from the original ones is affected by the length of the documents. It is more difficult to detect some contextual patterns in short documents, especially the patterns resulting from longer generative n-grams. This effect was manifested on simulated collections as the dependence of the variability of classifiers on the document length. The results are more reliable for longer documents; short documents imply more heterogeneous results.

Represen-tation	Document length	Mean F-measure difference	F sig.	F sign scheme
OutDegree	10	0.009	0.21	**o**
	50	-0.001	0.46	**o**
	100	-0.004	0.13	**o**
	1000	-0.001	0.42	**o**

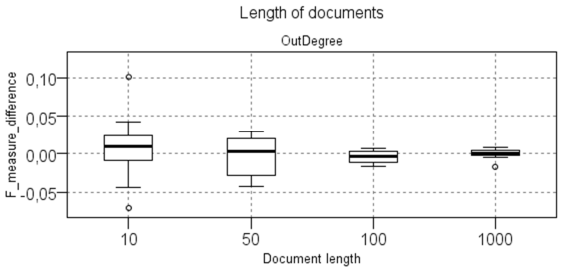


Figure 32: F-measure change by document length in OutDegree representation vocabulary size = 5, n-gram length = 3, clusters = 5

In real data we should not always assume that all documents in a collection come from the same n-gram distribution. Therefore we tested how the number of n-gram distributions used to generate a simulated collection affects the quality of the classification of permuted documents[107].The observed results confirm that the higher number of generative n-grams implies a worse classification performance. It may be caused by two effects: the limited capacity of the proposed representations to encode many different contextual patterns and the restricted capability of tested classifiers to separate more complicated regions that are formed in the input space due to more complicated input patterns. The magnitude of the decrease of the predictive power with the number of generative n-gram distributions differs for the various representations. For example, Closeness is more sensitive to the number of generative distributions than Betweenness.

Represen-tation	Number of clusters	Mean F-measure difference	F sig.	F sign scheme
Betweenness	3	0.312	<0.01	**+++**
	5	0.295	<0.01	**+++**
	10	0.241	<0.01	**+++**

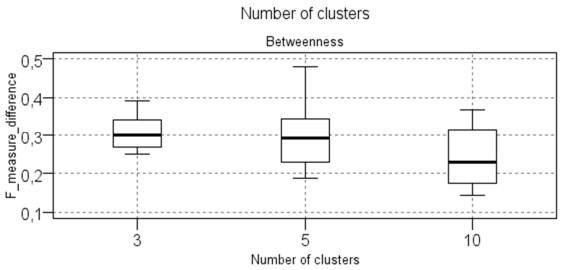


Figure 33: F-measure change by number of clusters in Betweenness representation vocabulary size = 5, n-gram length = 3, document length = 1000

The length of contextual patterns influences the classification performance as well. The longer n-grams are used for the document generation the more difficult the classification is. The effect was observed for all the representations; hence we can conclude that the proposed approach is suitable for capturing of shorter contextual ties. The degradation of classification performance is well observable for Betweeness representation where the long n-grams may completely destroy its advantages. On the other hand, short n-grams imply performance improvements even for the representations that were not so promising on Figure 31.

[107] If a generated document is selected for the permutation his original n-gram distribution may influence its topic frequencies only.

Represen-tation	n-gram length	Mean F-measure difference	F sig.	F sign scheme
Betweenness	2	0.438	<0.01	+++
	3	0.295	<0.01	+++
	5	-0.160	0.11	o

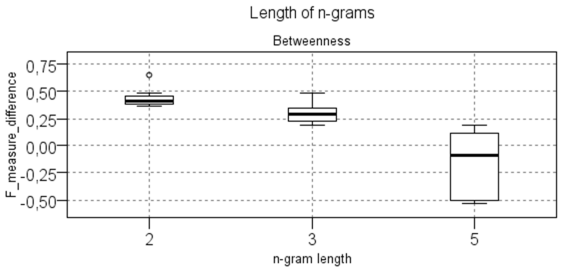


Figure 34: F-measure change by n-gram length in Betweenness representation vocabulary size = 5, document length = 1000, clusters = 5

The usefulness of the proposed representations depends on the vocabulary size as well, but the trend is not often monotonic and is influenced by values of other experimental parameters. In our test range of vocabulary sizes different effects of the vocabulary sizes were observed, but generally, we can conclude for the promising representations that they perform best for moderate sizes of the vocabulary, the best results were usually obtained for the vocabularies of the sizes 3 or 5 topics.

Represen -tation	Vocabulary size	Mean F-measure difference	F sig.	F sign scheme
Between-ness	2	-0.260	0.02	-
	3	0.177	<0.01	++
	5	0.295	<0.01	+++
	10	0.171	<0.01	+++

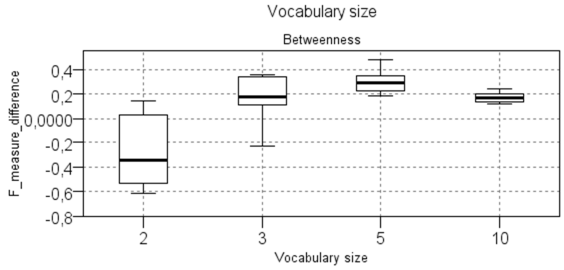


Figure 35: F-measure change by vocabulary size in Betweenness representation n-gram length = 3, document length = 1000, clusters = 5

The above mentioned interaction between the vocabulary size and other experimental parameters is well observable for the length of n-grams. The improvement of the classification for higher vocabulary sizes is degraded for longer contextual ties. The interaction effect even increases for higher vocabulary sizes, hence an observed marginal effect of the vocabulary size may not be monotonic.

Represen-tation	Vocabulary size	n-gram length	Mean F-measure difference	F sig.	Sign scheme
Betweenness	2	2	-0.110	0.15	o
	2	3	-0.260	0.02	-
	2	5	-0.092	0.15	o
	3	2	0.241	<0.01	+++
	3	3	0.177	<0.01	++
	3	5	-0.092	0.34	o
	5	2	0.438	<0.01	+++
	5	3	0.295	<0.01	+++
	5	5	-0.160	0.11	o
	10	2	0.397	<0.01	+++
	10	3	0.171	<0.01	+++
	10	5	-0.109	0.03	-

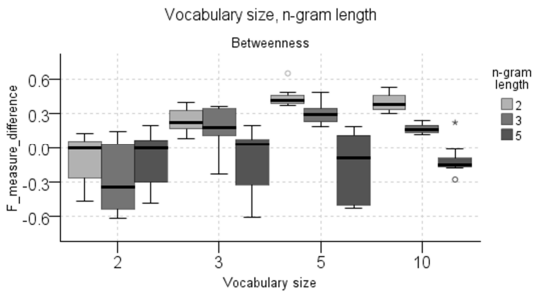


Figure 36: F-measure change by vocabulary size and n-gram length in Betweenness representation number of document length = 1000, clusters = 5

6.6 Performance of contextual representation in text mining tasks

6.6.1 Experimental setup for recognition of translated documents

The recognition of translated documents is a binary classification problem. The objective of this experiment is to evaluate the proposed representations in a task where the context may be more important than in the case of recognitions of main document subjects.

The goal is to distinguish the documents that are products of a machine translator from the documents written by human authors in a supervised way. Both document classes are of the same language, but we assume that machine translators are not still satisfactorily developed to fully comply with morphology and syntax of the target language, hence their outcomes may be recognized on the base of the contextual information that is present in translated documents.

The experiments were conducted on a specifically modified subset of the Czech downloaded collection. The original labels were discarded and the selected documents were translated into English and then back into the Czech language. The documents were labeled by an indicator of the translation to develop a classifier.

The setup of recognition experiments is similar to the setup of other classification experiments (see chapter 6.6.5), only the target is binary. Bayesian classifiers were developed and they were evaluated using unweighted F-measure (149). The parameters of the experimental setup are depicted in the following table. Each combination of the parameters was evaluated 25 times to distinguish better fine contributions of the proposed representations. Different random seeds were used for each run to split the collection into training and test sets in the ratio 70:30, significances of the difference between F-measures using the proposed and the standard bag-of-topics representations are presented as experimental results.

parameter	values
collection	CZ2
topic vocabulary size	5, 10, 20, 50, 100
length of left context window	1, 2, 4, 9
tested representations	*bag-of-topics*, Authority, Betweenness, Closeness, Degree, Eigenvector, Hub, InDegree, OutDegree, PageRank
number of repetitions	25
total number of experiments	5000[108]

Table 15: The values of evaluation parameters in the translated document recognition task. The bag-of-topics representation serves as the benchmark for comparisons with the centrality representations.

6.6.2 Recognition of machine translated documents

Similarly to the general classification experiments, the unweighted F-measure (149) is used as the evaluation metric. The gained F-measure is compared to the F-measure for the benchmark bag-of-topic representation and their differences are reported. The positive differences imply that the proposed representation is better for the binary classification than the benchmark and vice versa.

The experimental results confirm that the embedded contextual information in document vectors can improve the recognition of machine-translated documents. Even though the dependence of the achieved F-measure on the centrality is similar to the previous general classification, problem boxes are slightly shifted up resulting into positive performance of the most representations. Betweenness and Closeness may again worsen the performance. The

[108] The length of context window does not apply to the bag-of topic representation.

results for Betweeness are again rather heterogeneous; Betweennes may not be a wrong selection in some situations, but its suitability for specific data and other representation parameters should be considered carefully. The best and rather save option is Authority, other representations except PageRank may be used as well.

Representation	Mean F-measure difference	Sig.	Sign scheme
Authority	0.023	<0.01	**+++**
Betweenness	-0.183	<0.01	**---**
Closeness	-0.126	<0.01	**---**
Degree	0.013	0.04	**+**
Eigenvector	0.016	0.02	**+**
Hub	0.015	0.02	**+**
InDegree	0.014	<0.01	**++**
OutDegree	0.015	0.03	**+**
PageRank	-0.007	0.21	**o**

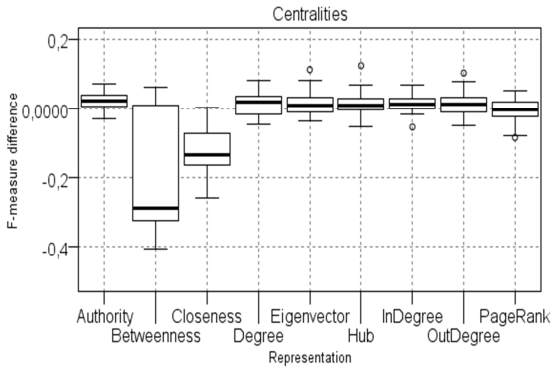


Figure 37: F-measure change by representation
vocabulary size = 5, n-gram length = 3

Unlike in the general classification problem, the n-gram length influences the discrimination of machine-translated documents. Its effect is not observable for all tested representations, but generally we can conclude that shorter and middle sized context windows perform better.

Representation	n-gram length	Mean F-measure difference	Sig.	Sign scheme
Eigenvector	2	0.02	<0.01	**+++**
	3	0.02	0.02	**+**
	5	0.00	0.09	**o**
	10	0.01	0.19	**o**

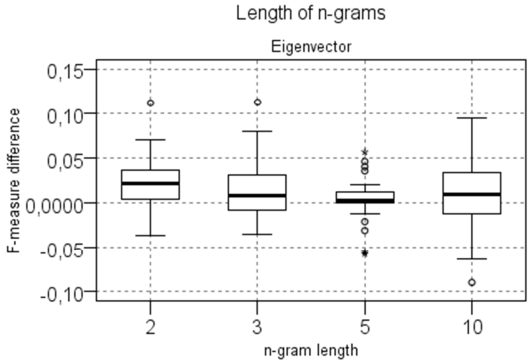


Figure 38: F-measure change by n-gram length in Eigenvector representation
vocabulary size = 5

The effect of the vocabulary size on the discrimination is rather small and again varies for different representations. The only apparent conclusion is that very short vocabularies perform better. Therefore we can conclude that the most of the proposed representations can be recommended for small dictionaries and shorter context windows.

Representation	Vocabulary size	Mean F-measure difference	Sig.	Sign scheme
InDegree	5	0.01	<0.01	++
	10	-0.03	<0.01	--
	20	-0.00	0.40	o
	50	-0.00	0.43	o
	100	-0.01	0.06	o

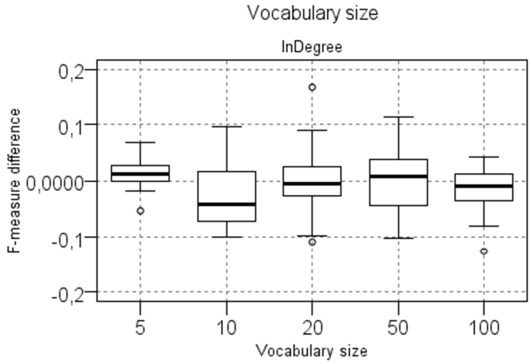


Figure 39: F-measure change by vocabulary size in InDegree representation n-gram length = 3

6.6.3 Information retrieval experimental setup

The goal of the information retrieval task is to select the documents from a collection that are similar to a given query. The query is considered as an additional document. The similarities between each document in the collection and the query are computed and the given number of documents with the highest similarity scores is retrieved.

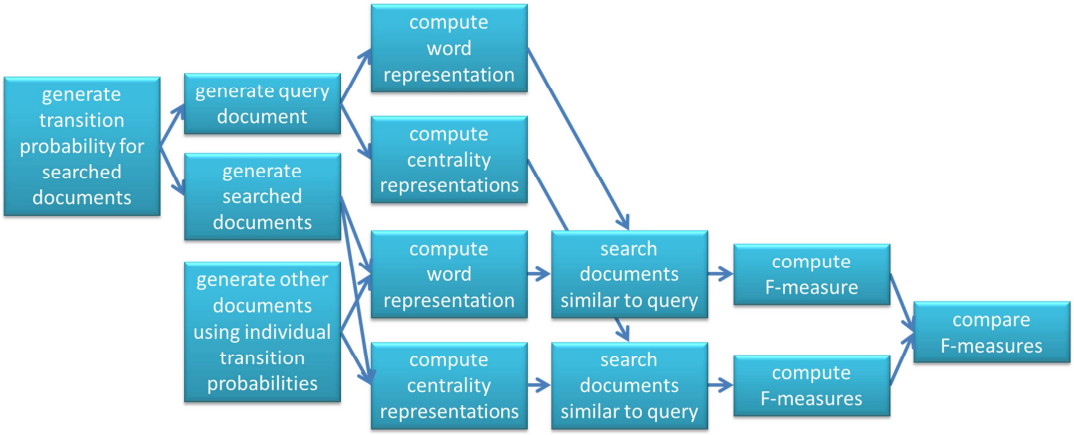


Figure 40: The evaluation of the information retrieval on the simulated documents.

The information retrieval task was examined on collections of generated documents. The generated collections in the experiments consist of 100 documents each; for each collection a query was generated as the special additional document. The given number of documents in the collection was generated using the same n-gram distribution as the query; other documents were generated using different n-gram distributions. The number of the retrieved documents is always the same as the number of the documents with the same n-gram distribution; these documents are considered as the documents to be correctly retrieved.

The similarity between each collection document and the query was measured by the cosine similarity (146). The cosine similarity served as the score to sort the documents; the documents with the highest scores were retrieved. The correctness of the retrieval process is evaluated using the unweighted F-measure (149) that is defined as the harmonic mean of precision and recall.

The numbers of retrieved documents together with other parameters of the experimental setup are depicted in the following table. Each combination of the parameters was evaluated ten

times; a different random seed was used for each run to generate the collection documents and the query. The repeated evaluations were used to estimate the significances of the differences between F-measures using the proposed and the standard representations.

parameter	values
topic vocabulary size	2, 3, 5, 10
length of left context window	1, 2, 4
number of documents	100
document length	10, 50, 100, 1000
number of retrieved documents	5, 10, 50
number of repetitions	10
total number of collections	1 440
tested representations	*n-gram*, *(n-1)-gram*, *bag-of-topics*, Authority, Betweenness, Closeness, Degree, Eigenvector, Hub, InDegree, OutDegree, PageRank
total number of experiments	17 280

Table 16: The tested values of simulation parameters in the information retrieval task. Bag-of-topics, n-gram and (n-1)-gram representations serve as benchmarks for comparisons with the centrality representations.

The bag-of-topics representation served as the standard representation for the comparisons. Differences between F-measures of the information retrieval using the proposed and the standard representations are analyzed to evaluate the appropriateness of the proposed centrality representations for the information retrieval.

6.6.4 Information retrieval

The unweighted F-measure (149) that equally combines the precision and the recall is used to estimate the usefulness of the proposed representations for the selection of the documents that are similar to the query. The obtained F-measure is compared to the F-measure for the benchmark bag-of-topic representation and only their difference is reported. The positive values imply that the proposed representation is better than the benchmark and vice versa.

The simulations confirm that the selection of the centrality is unimportant. The centrality performances are not significantly different from the performance of the benchmark representation. It implies that the way how the contextual information about the topic adjacency is projected to the vector representation does not influence the retrieval process. Hence we can conclude that only the document content is important for search engines and the contextual ties can be neglected.

Representation	Mean F-measure difference	Sig.	Sign scheme
Authority	-0.01	0.50	o
Betweenness	0.00	0.50	o
Closeness	0.00	0.62	o
Degree	0.00	0.75	o
Eigenvector	-0.01	0.50	o
Hub	-0.00	0.68	o
InDegree	0.00	1.00	o
OutDegree	0.02	0.31	o
PageRank	0.01	0.50	o

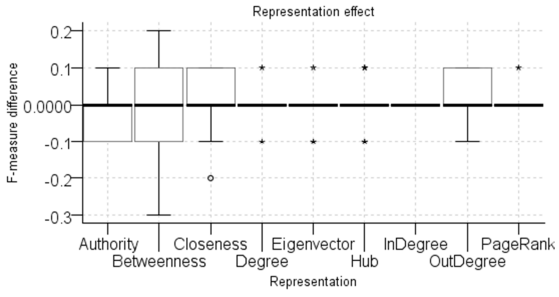


Figure 41: The F-measure change by the representation, the centrality selection is not significant for the standard parameter values
vocabulary size = 5, n-gram length = 3, retrieved documents = 10, document length = 100

The above statement about the context unimportance holds for different vocabulary sizes, n-gram lengths, document lengths and also for the portion of retrieved documents. The

dependencies on experimental parameters were observed for Betweenness only; other representations perform similarly as the benchmark when the values of the parameters are changed. Betweenness is preferable for small vocabularies, hence it is suitable when the documents need to be represented by short vectors. The usefulness of Betweenness is magnified when the small vocabulary is combined with a short context window.

Represen-tation	Vocabulary size	n-gram length	Mean F-measure difference	Sig.	Sign scheme
Between-ness	2	2	0.58	<0.01	+++
	2	3	0.73	<0.01	+++
	2	5	0.86	<0.01	+++
	3	2	0.25	<0.01	++
	3	3	0.37	0.01	+
	3	5	0.81	<0.01	+++
	5	2	-0.34	<0.01	--
	5	3	0.00	0.50	o
	5	5	0.13	0.17	o
	10	2	-0.15	0.01	-
	10	3	0.00	0.57	o
	10	5	-0.02	0.37	o

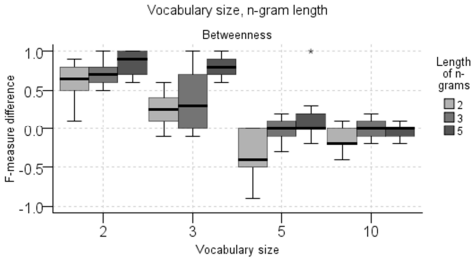


Figure 42: The F-measure change by the vocabulary size and the n-gram length in the Betweenness representation
retrieved documents = 10, document length = 100

6.6.5 Classification experimental setup

The performance of the proposed representation can be also tested in the classification task on downloaded collections because the labels are available for them[109]. The labels were obtained as the names of categories under which the press releases were published. The categories are not of a hierarchical structure. In the German collection the same document can be published under several categories (see the categories in chapter 6.3.1.1.3); for the evaluation purposes such documents were duplicated with the different labels[110]. In other collections the categories are disjunctive; each document is labeled by just one category.

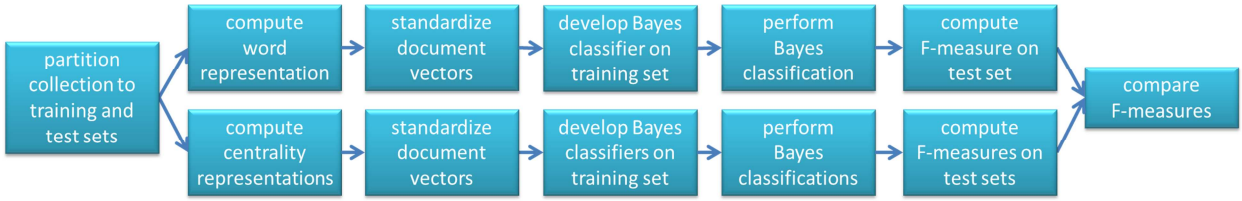


Figure 43: The evaluation of the classification on the downloaded documents.

The classification can be performed using the different supervised algorithms[111] that can influence the evaluation results [112]. However, the purpose of the evaluation of the classification is not to suggest the best algorithm, but to access the suitability of the proposed document representations. Hence the same algorithm was used in all classification experiments to be able to compare the results. Bayesian classifier was selected as the standard classification method. The classifiers are then evaluated using unweighted F-measure that is defined as the harmonic mean of precision and recall (149).

[109] The classification was not performed with the simulated documents because they were not assigned to categories by an independent reader.
[110] This evaluation approach is known as the micro-averaging.
[111] Or even by their ensembles.
[112] The comprehensive analysis of document classification approaches can be found in (Sebastiani & Delle Ricerche, 2002).

Bayesian models were learned over a training set of documents in each collection and evaluated over a test set. The parameters of the experimental setup are depicted in the following table. Each combination of the parameters was evaluated five times; a different random seed was used for each run to split the collection into training and test sets in the ratio 70:30. The repeated evaluations were used to estimate the significance of the differences between F-measures using the proposed and the standard representations.

parameter	values
collection	GE1, EN1, CZ1
topic vocabulary size	5, 10, 20, 50, 100
length of left context window	1, 2, 4, 9
tested representations	*bag-of-topics*, Authority, Betweenness, Closeness, Degree, Eigenvector, Hub, InDegree, OutDegree, PageRank
number of repetitions	5
total number of experiments	2775[113]

Table 17: The values of evaluation parameters in the document classification task. The bag-of-topics representation serves as the benchmark for comparisons with the centrality representations.

The bag-of-topics representation served as the standard representation for the comparisons. Differences between F-measures of classifiers using the proposed and the standard representations are analyzed to evaluate the appropriateness of the proposed centrality representations for the classification.

6.6.6 Classification in downloaded collections

The classification task was explored over three downloaded collections of real documents. The experiments exploit the document categories that are known in advance because the documents were downloaded from the different sections of news servers.

The unweighted F-measure (149) is used as the classification evaluation statistics. The obtained F-measure is compared to the F-measure for the benchmark bag-of-topic representation and only their differences are reported. The positive values imply that the proposed representation is better than the benchmark and vice versa.

The experiments confirm that the presence of the contextual information in document vectors does not improve the quality of classifiers. Seven out of nine centrality representations do not significantly worsen or improve the classifier performance; their F-measure is similar to the benchmark bag-of-topics representation. Betweenness always significantly worsens the classifier performance. Closeness also deteriorates F-measure, but the effect is not as striking as for Betweenness. Hence the experiments lead to the same statement as in the case of the information retrieval on the simulated documents: only the document content is important for the classification and the contextual ties can be neglected.

[113] The length of context window does not apply to the bag-of topic representation.

Represen-tation	Collec-tion	Mean F-measure difference	Sig.	Sign scheme
Authority	CZ1	-0.01	0.15	**o**
	EN1	-0.00	0.15	**o**
	GE1	-0.01	0.31	**o**
Between-ness	CZ1	-0.05	0.03	**-**
	EN1	-0.07	0.03	**-**
	GE1	-0.05	0.03	**-**
Closeness	CZ1	-0.01	0.06	**o**
	EN1	-0.03	0.03	**-**
	GE1	-0.04	0.03	**-**
Degree	CZ1	0.00	0.21	**o**
	EN1	0.00	0.03	**+**
	GE1	0.00	0.50	**o**
Eigen-vector	CZ1	-0.00	0.40	**o**
	EN1	0.00	0.50	**o**
	GE1	-0.01	0.40	**o**
Hub	CZ1	-0.00	0.40	**o**
	EN1	-0.00	0.31	**o**
	GE1	-0.01	0.31	**o**
InDegree	CZ1	0.00	0.06	**o**
	EN1	0.00	0.50	**o**
	GE1	0.00	0.43	**o**
OutDegree	CZ1	0.00	0.50	**o**
	EN1	0.00	0.21	**o**
	GE1	-0.00	0.40	**o**
PageRank	CZ1	0.01	0.03	**+**
	EN1	-0.00	0.40	**o**
	GE1	-0.01	0.15	**o**

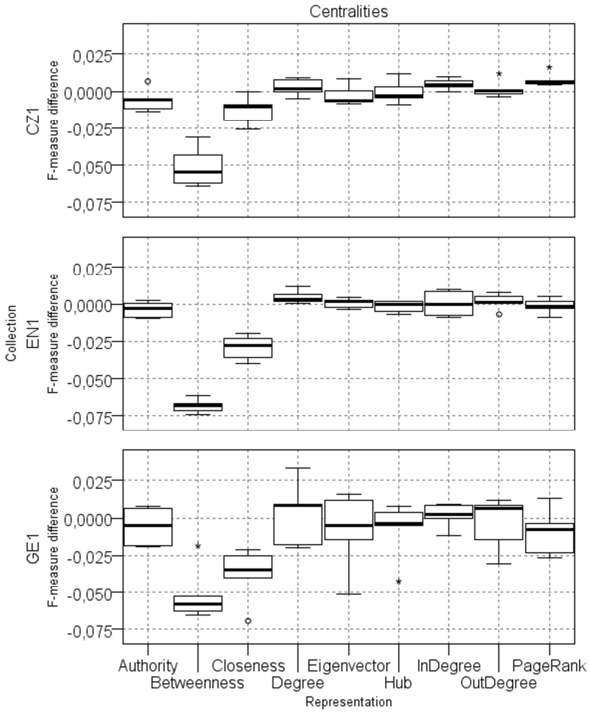


Figure 44: The F-measure change by the representation vocabulary size = 5, n-gram length = 3

The above statement about the context unimportance holds for different n-gram lengths; the size of the context window does not influence the classifier performance for all tested representations as well. The relation between F-measure and the vocabulary size is present for Betweenness and Closenes only. Betweenness performs best for larger vocabularies (the larger number of topics); it approaches to F-measure of bag-of-words representations for very large vocabularies. On the contrary, the large vocabularies negatively influence the performance of Closeness in the classification. The other centralities change their F-measure very little when the size of vocabulary is increased; a rather shallow minimum can be sometimes observed in the interval between 20 and 50 topics.

Represen-tation	Collec-tion	Vocabulary size	Mean F-measure difference	Sig.	Sign scheme
Closeness	CZ1	5	-0.01	0.06	**o**
		10	-0.05	0.03	-
		20	-0.08	0.03	-
		50	-0.09	0.03	-
		100	-0.09	0.03	-
	EN1	5	-0.03	0.03	-
		10	-0.05	0.03	-
		20	-0.07	0.03	-
		50	-0.10	0.03	-
		100	-0.09	0.03	-
	GE1	5	-0.04	0.03	-
		10	-0.06	0.03	-
		20	-0.07	0.03	-
		50	-0.05	0.06	**o**
		100	-0.10	0.03	-

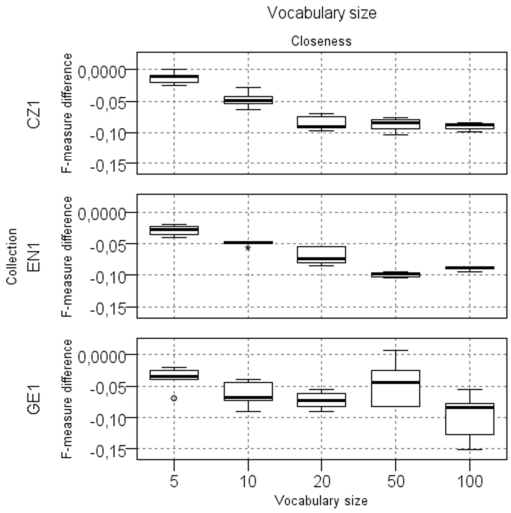


Figure 45: The F-measure change by the vocabulary size in the Closeness representation n-gram length = 3

Represen-tation	Collec-tion	Vocabulary size	Mean F-measure difference	Sig.	Sign scheme
Between-ness	CZ1	5	-0.05	0.03	-
		10	-0.03	0.03	-
		20	-0.03	0.03	-
		50	-0.02	0.03	-
		100	-0.02	0.03	-
	EN1	5	-0.07	0.03	-
		10	-0.05	0.03	-
		20	-0.04	0.03	-
		50	-0.03	0.03	-
		100	-0.02	0.03	-
	GE1	5	-0.05	0.03	-
		10	-0.04	0.03	-
		20	-0.05	0.03	-
		50	-0.04	0.03	-
		100	-0.01	0.21	**o**

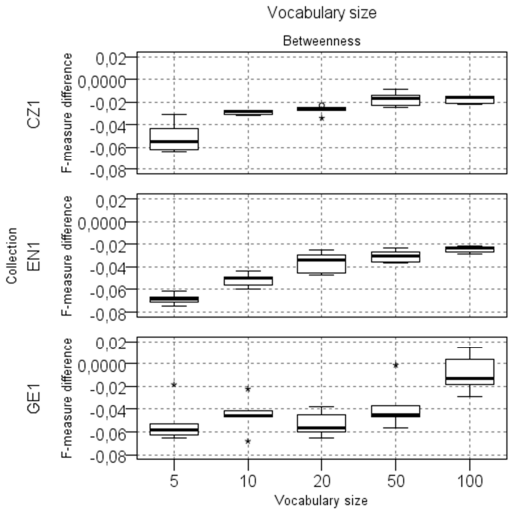


Figure 46: The F-measure change by the vocabulary size in the Betweenness representation n-gram length = 3

The above statements hold for all three collections, but the variability of results is again the largest for the German collection.

Represen-tation	Collec-tion	n-gram length	Mean F-measure difference	Sig.	Sign scheme
PageRank	CZ1	2	0.01	0.03	+
		3	0.01	0.03	+
		5	0.01	0.03	+
		10	0.01	0.21	o
	EN1	2	-0.00	0.21	o
		3	-0.00	0.40	o
		5	0.00	0.40	o
		10	-0.01	0.03	-
	GE1	2	-0.00	0.31	o
		3	-0.01	0.15	o
		5	0.00	0.40	o
		10	-0.01	0.15	o

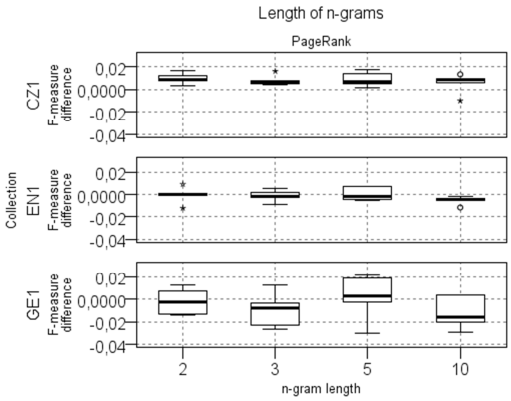


Figure 47: The F-measure change by the n-gram length in the PageRank representation vocabulary size = 5

6.6.7 Clustering experimental setup

The similarity of documents is also exploited in the document clustering[114]. Hence it is worth examining how the proposed document representations perform in the clustering task. However, the clustering can be performed by different algorithms that can influence the results. For evaluation purposes the popular k-means algorithm was selected. It is a common method in the document clustering due to its simplicity and efficiency[115]. Clustering experiments are conducted on both generated and downloaded collections.

The clustering is naturally an unsupervised process, but for simulated documents we can control the hidden n-gram distributions that are used for the document generation. Hence the desired spread out of documents in the input space can be set. In the experiments the documents generated over the same distribution of n-grams are considered to belong to the same cluster. This experimental design enables to evaluate reasonably the clustering results comparing the structure of clusters found by the k-means method with the structure of document groups of the same n-gram generative distribution. The standard evaluation measures used for the supervised evaluation of the unsupervised clustering include the purity, the normalized mutual information (MNI), the rand index (RI) and F-measure. The formulas are given in chapter 6.4.3. The outcomes of our simulation experiments are reported using MNI.

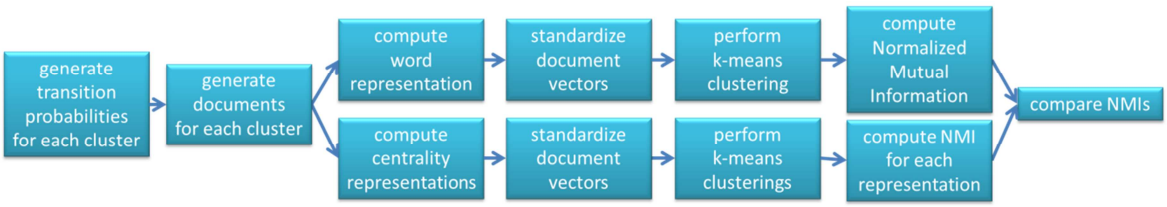


Figure 48: The evaluation of the clustrering on the simulated documents.

Generally, the number of retrieved clusters and the number of actual document groups can differ, but in the performed experiments the number of required clusters is always the same as

[114] The clustering techniques are also used in text mining as an alternative approach to the dimensionality reduction. They help to discover interesting word clusters that characterize word senses or semantic concepts.
[115] Other clustering methods used on the field of text mining include the hierarchical agglomerative clustering, the self-organizing maps or the graph partitioning spectral clustering. The comprehensive analysis of the text clustering approaches can be found in (Aggarwal & Zhai, 2012).

the number of groups of the documents with the same n-gram generative distribution. The sizes of groups of generated documents are either uniform or random. The random group sizes are generated from the uniform distribution and normalized to the number of documents in the collection.

The collection size, the number of clusters and other parameters of the experimental setup for collections of generated documents are depicted in the following table. Each combination of those parameters was evaluated ten times; a different random seed was used for each run to generate collection documents. The repeated evaluations were used to estimate the significances of the differences between MNI scores of the proposed and the standard representations.

parameter	**values**
topic vocabulary size	2, 3, 5, 10
length of left context window	1, 2, 4
number of documents	10, 20, 50, 100
document length	10, 50, 100, 1000
number of clusters	3, 5, 10
size of clusters	uniform, random
number of repetitions	10
total number of collections	11 520
tested representations	*n-gram*, *(n-1)-gram*, *bag-of-topics*, Authority, Betweenness, Closeness, Degree, Eigenvector, Hub, InDegree, OutDegree, PageRank
total number of experiments	138 240

Table 18: The tested values of simulation parameters in the clustering task. Bag-of-topics, n-gram and (n-1)-gram representations serve as the benchmarks for comparisons with the centrality representations.

The bag-of-topics representation served as the standard document representation for the comparisons. The differences between MNI of clustering models using the proposed and the standard representations are analyzed to evaluate the appropriateness of the proposed centrality representations for the clustering of documents.

The setup of clustering experiments with downloaded collections is similar to the setup of the experiments with artificially generated documents. Only one common k-means algorithm was selected for evaluation purposes to compare the performance of the proposed representations[116].

Even though the k-means clustering is an unsupervised process, it can be evaluated using the known document labels that were collected for the downloaded collections for the previous classification experiments. The documents from the German collection that are marked by multiple labels are duplicated; each copy receives its own unique label[117]. The known labels enable to evaluate reasonably the clustering results comparing the structure of clusters found by the k-means method with the structure of category labels. Again the outcomes of the clustering experiments with the downloaded collections are reported using MNI.

The number of retrieved clusters can be changed through the experiments; the number of clusters should not affect the evaluation metrics[118]. However, in the performed experiments the number of required clusters is always the same as the number of document categories in each downloaded collection. Hence the number of clusters is the fixed parameter of each

[116] The comprehensive analysis of document clustering approaches can be found in (Aggarwal & Zhai, 2012).
[117] It is the same micro-averaging approach used also in the classification.
[118] It is not true for the purity; other cited measures are adjusted by the number of clusters.

downloaded collection; it is not changed by the experimental design. The number of categories for each downloaded collection together with their size distributions can be found in chapter 6.3.1.1.

The k-means models were learned over the training set of downloaded documents and evaluated over the test set. The variant parameters of the experimental setup are depicted in the following table. Each combination of the parameters was evaluated five times; a different random seed was used for each run to split each downloaded collection into training and test sets in the ratio 70:30. The repeated evaluations were used to estimate the significance of the differences between MNI scores of the proposed and the standard representations.

parameter	values
collection	GE1, EN1, CZ1
topic vocabulary size	5, 10, 20, 50, 100
length of left context window	1, 2, 4, 9
tested representations	*bag-of-topics*, Authority, Betweenness, Closeness, Degree, Eigenvector, Hub, InDegree, OutDegree, PageRank
number of repetitions	5
total number of experiments	2775[119]

Table 19: The values of evaluation parameters in the clustering task. The bag-of-topics representation serves as the benchmark for comparisons with the centrality representations.

6.6.8 Clustering in generated collections

While the evaluation of the clustering is generally difficult, in our simulations where the cluster assignment is known in advance the evaluation is done in a supervised manner. Among several standard evaluation measures the normalized mutual information (MNI) (154) was selected for the reports. MNI ranges between zero and one, higher values indicate better agreement between the found clusters and the actual clusters. Similarly to other experiments, MNI obtained for the proposed representation is compared to MNI for the benchmark bag-of-topic representation. The percentage difference of MNI is reported, the base value for the percentages is MNI of the bag-of-topic representation. The positive percentages correspond to the increase of MNI and vice versa.

The selection of the centrality is not very important for the clustering; most of them do not perform significantly better or worse than the benchmark representation. The save centrality is again Authority that either improves NMI or does not change it depending on the experimental parameters. Betweenness and Closeness should be used carefully; they can both worsen and improve the results. Their performances depend on the setup of the experiment.

Representation	Mean NMI difference (%)	Sig.	Sign scheme
Authority	21.2	0.04	+
Betweenness	-19.7	0.01	-
Closeness	-4.3	0.11	o
Degree	5.0	0.46	o
Eigenvector	8.3	0.31	o
Hub	2.9	0.42	o
InDegree	-1.5	0.38	o
OutDegree	4.9	0.38	o
PageRank	1.7	0.50	o

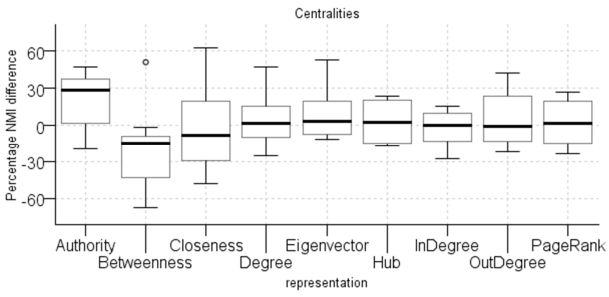


Figure 49: The NMI change by the representation
vocabulary size = 5, n-gram length = 3, documents = 100, document length = 100, clusters = 5

[119] The length of context window does not apply to the bag-of-topic representation.

The dependence of other proposed representations than those for Betweenness and Closeness on the experimental parameters was not observed; their NMI difference is not influenced by the number of clusters, the cluster size distribution, the length of documents, the size of the collection and the length of the context window.

Represen-tation	Document length	Mean NMI difference (%)	Sig.	Sign scheme
Between-ness	10	24.8	0.03	+
	50	-11.4	0.08	o
	100	-19.7	0.01	-
	1000	-56.4	<0.01	---

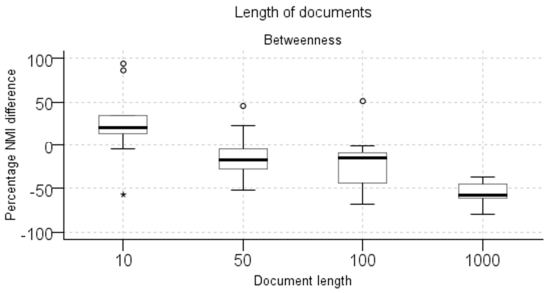


Figure 50: The NMI change by the document length in the Betweenness representation vocabulary size = 5, n-gram length = 3, documents = 100, clusters = 5

Represen-tation	Vocabulary size	Mean NMI difference (%)	Sig.	Sign scheme
Between-ness	2	-100	<0.01	---
	3	-31.3	<0.01	---
	5	-19.7	0.01	-
	10	34	0.05	o

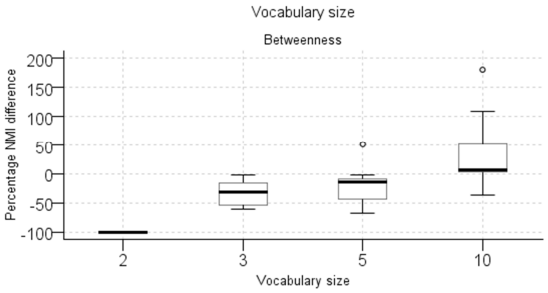


Figure 51: The NMI change by the vocabulary size in the Betweenness representation n-gram length = 3, documents = 100, document length = 100, clusters = 5

On the contrary to the information retrieval, we can state that the contextual ties within documents may not be always neglected; they can improve the clustering. However, the context is not as important as we can expect from the measurement of the preservation of the document diversity; its influence on the clustering is small and can be observed for several combinations of the experimental parameters only. Moreover, the improper parameter setup can significantly worsen the results.

6.6.9 Clustering in downloaded collections

While the clustering in downloaded collections is the unsupervised task, its evaluation is supervised. The known document categories are compared with the identified clusters. The normalized mutual information (MNI) (154) is reported as the evaluation measure. MNI ranges between zero and one, higher values indicate better agreement between the identified clusters and the document categories. MNI obtained for the proposed representation is compared to MNI for the benchmark bag-of-topic representation. Their percentage differences are reported; the base value for the percentages is MNI of the bag-of-topic representation. The positive percentages correspond to the increase of MNI and vice versa.

Eight out of nine proposed representations do not exhibit much better or worse performance in the clustering task; their NMI is about the same as for the benchmark bag-of-topics representation. Closeness is nearly always worse than the bag-of-topics representation with minor exceptions for the German collection.

The size of the context window (the n-gram length) does not influence the clustering performance for all tested representations. It confirms the hypothesis that the contextual dependences present in the text should not be exploited to enhance the quality of the

clustering. The vocabulary size influences the clustering performance only slightly. The actual relation depends on the representation and on the collection. If there is any relation between the vocabulary size and NMI, it is best observable on the German collection.

Represen-tation	Collec-tion	Mean of NMI difference (%)	Sig.	Sign scheme
Authority	CZ1	0.1	0.40	o
	EN1	0.9	0.31	o
	GE1	1.3	0.21	o
Between-ness	CZ1	0.8	0.31	o
	EN1	0.6	0.50	o
	GE1	-2.2	0.21	o
Closeness	CZ1	-16.7	0.03	-
	EN1	-26.9	0.03	-
	GE1	-0.7	0.31	o
Degree	CZ1	0.3	0.15	o
	EN1	3.4	0.03	+
	GE1	2.8	0.06	o
Eigen-vector	CZ1	-0.5	0.15	o
	EN1	2.4	0.09	o
	GE1	3.8	0.03	+
Hub	CZ1	0.2	0.21	o
	EN1	2.5	0.03	+
	GE1	4.0	0.03	+
InDegree	CZ1	-0.2	0.31	o
	EN1	3.7	0.03	+
	GE1	0.7	0.31	o
OutDegree	CZ1	0.5	0.15	o
	EN1	3.9	0.03	+
	GE1	2.7	0.03	+
PageRank	CZ1	-2.4	0.03	-
	EN1	-2.3	0.03	-
	GE1	0.3	0.50	o

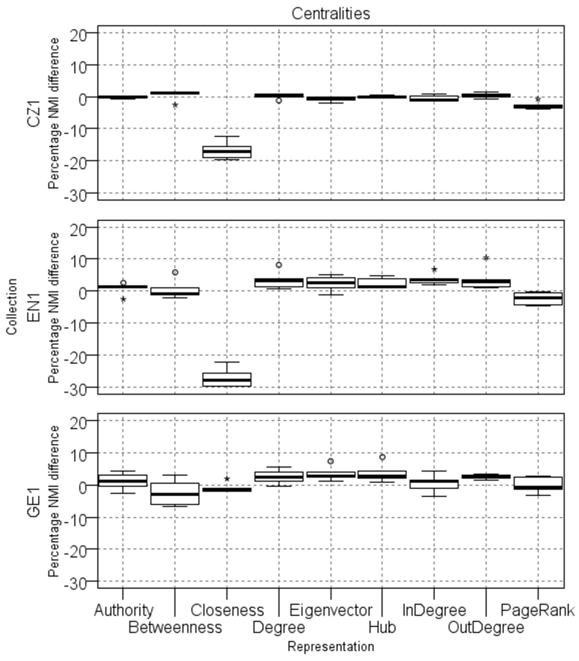


Figure 52: The NMI change by the representation
vocabulary size = 5, n-gram length = 3

Represen-tation	Collec-tion	Vocabulary size	Mean of NMI difference (%)	Sig.	Sign scheme
Hub	CZ1	5	0.2	0.21	o
		10	-0.4	0.15	o
		20	-0.6	0.31	o
		50	1.5	0.50	o
		100	1.2	0.50	o
	EN1	5	2.5	0.03	+
		10	-3.7	0.21	o
		20	1.9	0.31	o
		50	-0.1	0.40	o
		100	-2.7	0.15	o
	GE1	5	4.0	0.03	+
		10	-4.4	0.06	o
		20	-10.4	0.03	-
		50	-11.3	0.15	o
		100	-16.8	0.03	-

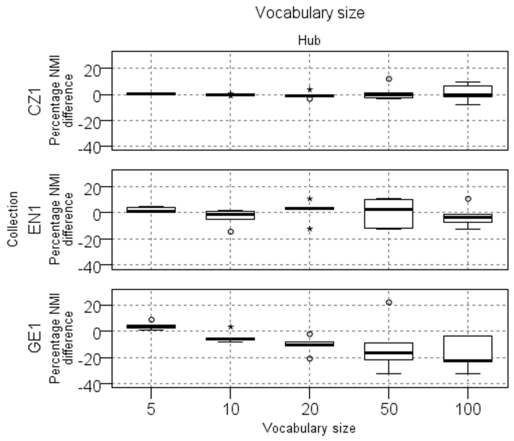


Figure 53: The NMI change by the vocabulary size in the Hub representation
n-gram length = 3

The relation between the vocabulary size and NMI is clearly visible for Closeness. However, this particular relation is not monotonic; NMI takes its minimum value for about twenty topics.

Represen-tation	Collec-tion	Vocabulary size	Mean of NMI difference (%)	Sig.	Sign scheme
Closeness	CZ1	5	-16.7	0.03	-
		10	-31.5	0.03	-
		20	-34.3	0.03	-
		50	-19.2	0.03	-
		100	-10.8	0.03	-
	EN1	5	-26.9	0.03	-
		10	-39.8	0.03	-
		20	-34.7	0.03	-
		50	-25.8	0.03	-
		100	-20.7	0.06	**o**
	GE1	5	-0.7	0.31	**o**
		10	-35.0	0.03	-
		20	-36.7	0.03	-
		50	-31.4	0.03	-
		100	-20.5	0.06	**o**

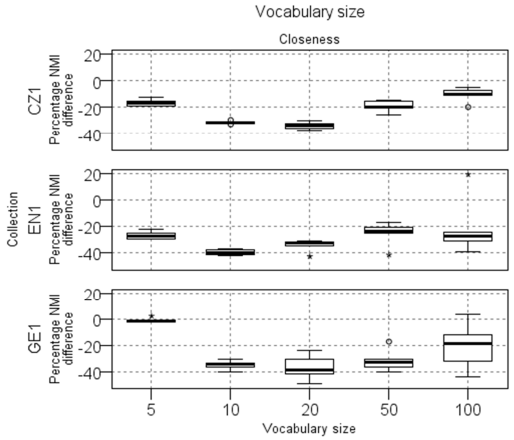


Figure 54: The NMI change by the vocabulary size in the Closeness representation n-gram length = 3

Based on the experimental results, we can state that it is usually not worth including the contextual ties that occur within documents into the final document vectors if they are prepared for the clustering. Moreover, the improper selection of the contextual representation can significantly worsen the results.

7 Summary and conclusions

The new approach to the document representation was proposed in the thesis. It enables to enhance document vectors by the contextual information retrieved from text. The analysis of the comprimation of the context into the proposed representations was introduced. The main effort was made to experimentally evaluate benefits of this contextual enhancement for common text mining tasks. The experiments confirmed that the common text mining tasks such as document clustering are not sensitive to the contextual arrangement of a text. Hence we recommend simplifying of the process of the extraction of useful features from a text by focusing on the document content only.

7.1 Process recapitulation

The main goal of this thesis was to investigate the alternative approaches to structured representations of documents and theoretically as well as practically evaluate the appropriateness of the proposed representations for text mining tasks. The main idea behind the work is to find a way how to transfer useful contextual information from unstructured texts to the numeric document vectors. Following the process of the general knowledge discovery[120], any input data are modified, merged with other sources and reduced to form a two-dimensional modeling matrix of subjects of our interest in rows with their features in columns. Such a matrix serves as an input for the machine learning algorithms that discover useful patterns in data that support decision-making. In the text mining field the subjects are clearly identifiable; they are text documents. However, the features that describe the documents are not exactly known in advance; the essential step in any text mining task is to select or construct the appropriate descriptive features[121]. The extracted features should support tasks such as the information retrieval, the document classification or the document clustering.

The data preparation phase of any text mining project involves the transformation of unstructured texts of variant lengths to some structured vectors of the fixed length. Such transformations necessarily reduce the information that is included in the text. If an important pattern is lost through this process, the subsequent machine learning algorithm cannot exploit it. Hence it is worth paying attention to the selection of the appropriate document representation rather than to applying sophisticated models that try to find something that was filtered out from the input data[122].

In the unstructured texts the information is encoded in a rather complicated way. The natural languages are very rich and each text can be investigated on different levels of the linguistic hierarchy[123]. The common languages like English or Czech enable to analyze a document as a stream of characters, as a sequence of word-forms or as a particularly random structure that describes the ordered system of hidden topics. However, the richness of the written language is not only based on its hierarchical structure; each linguistic level usually operates with a broad number of distinct categories. For example, the morphology level usually offers tens of thousands of wordforms. Additionally, the words or other entities from different linguistic levels do not appear randomly in a text, their appearance depends on a context. The context is usually modeled by language models where the presence of the particular entity is conditioned

[120] The process is formally described by CRISP-DM methodology.
[121] The document features can be combined with the features that come from other originally structured data sources in complex data mining solutions.
[122] GIGO (Garbage In Garbage Out) effect
[123] The linguistic recognizes different language levels like lexicology, morphology, syntax and semantic.

by its preceding entities. The language modeling offers document representations where the features are ordered groups of some linguistic entities called n-grams. Unfortunately, such approach further significantly increases the dimensionality of possible representations; the number of potential n-grams is much higher than the already high number of the entities from which the n-grams are formed [124]. If any n-gram representation is used, the n-grams frequencies are very low even in a large corpora; most vocabulary n-grams are not present in a particular document that implies the extremely spare representation. Hence the dimensionality reduction must be employed to receive the vectors of the reasonable length that can differentiate the documents enough when applying data mining models.

Due to the importance of the context it is worth trying to maintain the information about the linguistic entity order in a reduced representation. Hence there are two may be conflicting requirements for the process of dimensionality reduction: to maintain the document content represented by the presence of the extracted entities in the document and to maintain the document context represented by the adjacency of these entities in the document. The second requirement is often omitted and the documents are represented by frequencies of some selected or extracted entities only; their order is not taken into account[125]. However, the loss of the context information may cause the deterioration in the discrimination of documents; the predictive text mining models may perform badly because some essential information is not present in the input document vectors.

The relations among linguistic entities within a document can be captured by a graph. The entities form graph vertices while their relations are depicted by edges between pairs of vertices. The weight assigned to the particular edge is proportional to the co-occurrence of its vertices in the document within the defined vicinity. Hence each document can be represented by its contextual graph in the form of a social network. Such a context network further enables to derive measures that describe the properties of the network and can be arranged to the vector form.

The centralities of nodes in the document's context network were selected to represent the document. They encode both the context and the content of the document and the centralities of all vertices together form the centrality vector of the document. The centrality vector is the final proposed representation of the document; it is the product of the specific dimensionality reduction approach. Many centrality measures have been proposed in the field of Social Network Analysis (SNA). They describe the importance of a node regarding its connections to other nodes and/or they can take into account the importance of the adjacent nodes. Another group of centralities exploits possible paths between the network nodes and their distances. In the thesis nine centralities were selected for testing. They reflect the prestige of the nodes and can be divided into three groups:

- centralities that preferably describe the document content (Degree, InDegree, OutDegree),
- statuses of the nodes that are based on document context patterns (Eigenvector, Authority, Hub, PageRank),
- centralities that rely on the proximities of the nodes that alternatively describe the document content (Closeness, Betweenness).

The proposed representations of documents can be combined with other common approaches to the dimensionality reduction. The centrality vectors can be further simplified by the

[124] For example if a vocabulary of lemmas consists of fifty thousand words, the number of possible 3-grams equals $1.25*10^{14}$.
[125] For example the common bag-of-words representation.

selection of their important components or they can be projected to a low-dimensional space to form new derived attributes. On the other hand, any dimensional reduction can also be performed before the context networks are extracted from documents. The new extracted attributes will form the network vertices for which the centralities are computed.

The context network approach itself is language independent; it exploits the selected features extracted from a text and their relations. However, the features that constitute the vocabulary are often linguistic entities or their derivatives that can be extracted using some specific language dependent resources. In the thesis the attention is paid to the higher level features that are derived without special vocabularies. Specifically Latent Dirichlet Allocation (LDA) is exploited to produce a relatively small set of latent topics hidden behind the words of a document. The words are substituted by the topics before context networks are built. Hence the final dimensionality depends on the number of extracted topics. This parameter must be set in advance. The only language dependent step used in the presented experiments is the stemming. The stemming was included to speed up the substitution by LDA. It is not a resource consuming procedure; the stemming is usually a simple ruled-based algorithm.

The whole process of the extraction of the proposed document representation can be adjusted by three parameters: the number of extracted topics, the length of context window and the centrality measure. The number of topics extracted by LDA implies the dimensionality of final vectors. The topic extraction especially reduces the document content. The context window length parameter fixes the maximal distance in the text on which the interaction among the topics is taken into account. This parameter influences especially the reduction of the document context. And finally, the selected centrality measure affects how the evidences about the context and about the content are combined into document vectors.

7.2 Theory recapitulation and findings

In the theoretical part of the thesis the reduction of information that is caused by the proposed transformation of documents to vectors is evaluated. The goal is to explore how the above mentioned parameters influence the information maintained in the document vectors.

Regarding the common n-gram language model, the document can be represented by the matrix of transitions between (n-1)-grams[126]. Hence each document can be regarded as a product of the random process that is described by the transition probabilities. The document transition probabilities are unobservable, but they can be estimated from the observed transition frequencies. The probabilities respectively the observed frequencies are regarded as the full description of a document. The variability of the transitions causes the diversity of observed documents. The diversity of the documents is exploited by predictive models in all text mining tasks. Therefore the theory is focused on the investigation how the variability of transitions among the (n-1)-grams is affected by the proposed vector representation of documents.

The transitions among the (n-1)-grams are determined by the probabilities of n-grams. We assume that the observed n-gram frequencies within a document come from a multinomial distribution. Hence the frequencies of the observed transitions among the (n-1)-grams come from the same multinomial distribution, but they form a transition matrix. Some elements of the transition matrix are of zero probability[127]. Regarding these zero probabilities of the impossible transitions, the process can be considered as Markov chain. For further transformations it is worth considering only the matrix of transitions among (n-1)-grams and

[126] They can be transformed to transitions between (n-1)-grams and 1-grams without any loss of information.
[127] Two (n-1)-grams have to share first and last n-2 items to enable the transition between them.

unigrams. The transitions form a random matrix variable. However, the random matrix is treated as a random vector because it was built from the random vector of n-gram frequencies. Additionally, its multinomial distribution is worth approximating by the normal distribution to be able to estimate the distribution of the proposed document vectors.

The context network is described by the frequencies of unigram pairs that occur within the context window. These frequencies also form a random matrix variable. The context network matrix can be derived from the original (n-1)-grams transition matrix by multiplying a special matrix whose form depends on the length of the context window and the length of n-gram. Some of the selected network centralities can be also expressed using linear matrix operations. These linear transformations enable to estimate the distributions of proposed representations.

To be able to compare the distribution of the (n-1)-grams transitions with the distribution of the proposed vector representation, we try to estimate their dependence and the amount of information that is lost during the transformations from transitions to vectors. The random variables that are investigated are the vector variables, hence the usual correlation-like approach cannot be used[128]. For example, if we want to estimate R-squared measure that should tell us the proportion of original variability that is still present in the proposed vectors, we have to generalize the notion of covariance and partial covariance to apply it to random vectors. It results in comparisons of determinants of variance and partial variance matrices. More precisely, the generalization of R-square called Wilks' lambda is proportional to the ratio of these determinants.

The alternative approach how to evaluate the loss of the information during the derivation of the proposed representation is to estimate the mutual information between the original and the proposed representation. The mutual information between two normal vectors again depends on the ratio of determinants of covariance and partial covariance matrices, more precisely, it is proportional to the logarithm of this ratio. However, it is proved in the thesis that the determinant of the partial covariance matrix of linearly dependent vectors equals zero. The consequence of this finding is that the mutual information diverges and the generalized R-square measure is always one. In the case of non-linearly dependent vectors the measures can be theoretically used. However, the relation between the original representation and the centralities that cannot be declared using a matrix multiplication is so complicated that the estimation of their distributions is not tractable.

Therefore the main conclusion of the presented theory is to focus on experiments because we cannot reliably theoretically estimate how the diversity of documents is maintained in the proposed representations.

7.3 Simulations

The usefulness of the proposed representation is based on the assumption that linguistic entities which occur in a text do not appear independently, but they influence each other within a context window of the constant length. However, this assumption may not be always met. The usage of the fixed context window is naive; the linguistic entities such as words that interact may occur anywhere in the document. On the other hand, closer words interact more often than the words which are far away, hence the restriction of the contextual influence to the fixed length window may be justified.

[128] A vector random normal variable is described by a vector of means and a square covariance matrix instead of just two parameters in the case of a scalar random normal variable.

The main purpose of the investigation of the preservation of the document diversity on simulated documents is to estimate how the context information is changed in the proposed representation if the documents fully meet our context assumption. It is rather simple to generate the documents where the presence of particular linguistics entity depends on the presence of the fixed number of the previous entities.

The simulated documents are also further used to examine the performance of the proposed representations in standard text mining tasks: the classification, the information retrieval and the clustering. The preservation of the context information in the document vectors does not guarantee that such vectors improve the results of these tasks. Furthermore, the pertinence of the proposed representations may depend on specific document properties and on the parameters that influence the proposed document vectors. Both the document types and the representational parameters can be adjusted in the simulation experiments. Hence many simulation runs for various setups were executed to examine how different parameters influence the performance of the proposed representations. Namely the tested document properties include the collection size and the document length. The other general tested parameters include the vocabulary size, the context window length and the centrality measure. Additionally, in the clustering experiments the number of clusters and the distribution of the cluster sizes were changed through the simulations to evaluate their effects. The effect of the number of generative n-gram distributions was examined in the classification simulations. And in the information retrieval experiments the number of relevant documents was varied.

Each combination of the experimental parameters was evaluated ten times with different simulated documents to be able to test statistically their effects. The SSTERSS measure was used to estimate generally how the diversity of documents is maintained in the proposed representation. The standard F-measure served as an assessment metric in the classification of the documents with permuted topics. The evaluation of the clustering exploits the simulation origin of documents in which we know to which cluster each document belongs. Similarly, in the information retrieval simulations the documents that should be retrieved are labeled. Hence the normalized mutual information and F-measure serve as the evaluation measures for the clustering and the information retrieval respectively. The results for the proposed representations are always compared with results for the standard bag-of-topic [129] representation that serves as the benchmark and the differences between the evaluation measures are statistically tested to prove or to decline the usefulness of the proposed document vectors.

It was shown that the experimental parameters influence the results in the evaluation of the context preservation using SSTRESS measure. The centrality selection is the most important option. The simple centralities like Degree do not preserve the document diversity better than the benchmark representation, but the advanced centralities like Authority that exploit complex relations in context networks are useful. The usage of centralities like Betweenness, that depend on the lengths of paths through a context network, must be considered carefully regarding other parameters. Their performance can be the best and also the worst depending on the other properties of documents like the vocabulary and the context size. These parameters influence the performance of other centralities as well. It is rather difficult to generalize the observed relations; they are centrality dependent.

The presence of the contextual information in the proposed document vectors is exploitable by classifiers if the goal is to recognize the documents that follow different contextual patterns. It was shown that the proposed representations enable to distinct between the

[129] The bag-of-topic representation does not rely on the order of topics in text; hence the usefulness of context information for text mining tasks can be investigated.

documents that follow given n-gram distributions and the documents in which words were randomly permuted. The representations that rely on centralities like Betweenness that depend on path lengths through a context network are clearly preferred in this task. However, their performance is the most variable one hence their usage must be carefully considered regarding other parameters. On the other hand simple centralities like Degree, that depend only on plain counts of edges in a context network, do not improve performances of the classifiers. In the experiments, it was also shown that if documents include extensive contextual patterns, these patterns are simplified in the proposed representations. The simplification was observed as a reduced performance of classifiers that recognize documents generated using complicated contextual rules. Additionaly the number of underlining n-gram distributions, the length of n-grams or the vocabulary size influence the rate of compression of the contextual information into the proposed document vectors which can be further exploited for the classification of permuted documents. Moreover, the effects of these experimental parameters are not independent. For example, a larger vocabulary enables to encode richer contextual information because the vocabulary size implies the dimensionality of the document vectors, but longer n-grams are difficult to encode to a fixed number of dimensions, hence we observe a complex dependence among the classification accuracy, the dictionary size and the n-gram length. The way how the experimental parameters interact differs for different centralities.

The expediency of the proposed representations in the information retrieval task was not proved on the simulated documents. The performance of all tested centralities is comparable to the performance of the benchmark representation. Moreover, the experimental parameters do not significantly influence the results. The only exception is the Betweenness representation. It performs significantly better than the benchmark for small vocabularies, hence it should be the preferable representation in the cases where the documents need to be represented by short vectors.

The expediency of the proposed representations for the clustering task was not proved on the simulated documents as well. The difference between NMI for the proposed and the benchmark representations are not usually significant. The only promising centrality is Authority which performs better than the benchmark for some combinations of the experimental parameters. The centralities like Betweenness and Closeness, that rely on the path distances within the context network, are again the tricky ones. They can perform very well and also very badly depending on the properties of documents and the experimental parameters. They are especially useful for collections that contain short documents, but they should not be selected when the short context window is combined with the small vocabulary size.

7.4 Experiments with real documents

The contextual dependencies are very important in a written text; they enable the reader to fully understand the described facts. On the other hand, we proved in the experiments with simulated documents that the context information is not so important for the general document discrimination in text mining tasks, the context information is exploitable only in specific tasks. The main findings on the collections of real documents are consistent with the conclusions from the simulations.

Three main collections[130] that were used in the experiments differ in many parameters: the language, the collection size, the average length of documents and the number of known

[130] The fourth collection CZ2 was obtained as a subset of the main collection CZ1 and specifically modified for the recognition of machine translated documents.

categories to which the documents are assigned. While the properties of the downloaded collections are fixed, the parameters that influence the tested representation were changed in the experiments to investigate their importance. Namely the effects of the vocabulary size, the context window length and the centrality measure were evaluated. The vocabulary size refers to the number of extracted topics that substitute the words in documents.

Similarly to the experiments with simulated documents, the preservation of the document diversity within a collection was examined firstly. Secondly, the performance of the proposed representations in standard text mining tasks was evaluated. They include the classification and the clustering. In both tasks the known document assignment to categories was exploited in the evaluation. In the last experiment the proposed contextual representations were exploited for the recognition of machine translated documents.

Each combination of the free experimental parameters was evaluated five times[131] in the first three experiments and 25 times in the last experiment. The different random seeds for partitioning of documents to train and test sets were used in the experiment repetitions to be able to perform statistical tests[132]. The SSTERSS measure was used again to estimate the preservation of the diversity of documents in the proposed representations. The normalized mutual information and F-measure served as the evaluation measures for the clustering and the classification[133] respectively. The results for the proposed representations are again compared with the results for the benchmark bag-of-topic representation; the differences are statistically tested to show how useful the proposed representations are.

It was proved that the most proposed representations are able to capture the document diversity significantly better than the benchmark representation. Only Closeness and PageRank perform generally worse. On the other hand, Betweenness is often the best centrality, but its performance is rather variable depending on the specific data and the experimental setup. This observation is consistent with the simulation experiments; the performance of the centralities that exploit path lengths in a context network can be very good and also very bad depending on the collection properties and the experimental parameters. The usefulness of all centralities depends especially on the length of the context that is taken into account when context networks are constructed. The differences of SSTRESS between the proposed and the benchmark representations are large for shorter context windows and these differences diminish when the context window grows. The vocabulary size influences the diversity preservation as well, larger vocabulary size worsens SSTRESS. The performance differences were also observed among the collections. The proposed representations usually performed best on the English collection and the worst results were often observed on the German collection. The German collection is also the most sensitive to the parameter adjustments. It may be the consequence of language specific grammars. Syntax rules in English are stricter than in German or Czech. Additionally, the German collection consists of shorter documents and is the smallest main collection, hence the contextual patterns are harder to detect.

The usefulness of the proposed representation for the general document classification was not proved on three main collections. The performance of the most tested centralities is comparable to the performance of the benchmark representation. Closeness and Betweenness perform even worse than the benchmark. The length of the context window does not influence the classifier performance, but the weak effect of the vocabulary size was detected. This vocabulary size effect is centrality dependent. The differences among the collections were

[131] The smaller number of runs was selected because the experiments were computationally intensive.
[132] The exact tests were used for such small number of runs.
[133] Including the recognition of machine translated documents.

observed as well. Again the most sensitive collection to the parameter adjustments was the German collection.

The conclusions from the clustering experiments are very similar to the conclusions from the classification. The usefulness of the proposed representations for the clustering task was not proved on the main collections. The differences between NMI for the proposed and the benchmark representations are not usually significant. Only Closeness is nearly always worse than the benchmark. The length of the context window does not influence the clustering performance for all tested representations and the effect of the vocabulary size is not usually so large. This relation is better observable on the German collection only, the larger vocabulary implies the worse performance, but there are several exceptions from this rule; the dependence may not be monotonic for some centralities.

The usefulness of the proposed document representation was proved in a specific text mininig task of the recognition of machine translated documents. Machine translators still produce texts that sometimes follow incorrect contextual patterns, hence our approach may produce better results than a context-free representation. Indeed it was experimentally proved that many of the proposed representations can improve this classification. The representations based on centralities that rely on path lengths in a context network should not be taken into account[134] for the translated document recognition, but other tested centralities offer slightly better results than the benchmark bag-of-topic representation. The most promising representation is Authority. The proposed centrality based representations are preferable especially for shorter context windows and smaller vocabularies, hence they are adequate in situations when a low-dimensional contextual representation of documents is desirable.

7.5 Overall conclusions and recommendations

The contextual relations among linguistic entities such as observable words or hidden topics may be quite complicated, but the context is very important for the perception and understanding natural languages. If we need to encode the contextual dependences which occur within a text into numeric vectors that are necessary for the bulk processing of documents, a simplified contextual model has to be taken into account. The number of possible relations between linguistic entities grows rapidly with the vocabulary size, hence any context encoding schema must also include a dimensionality reduction method.

It is rather difficult to exactly estimate the effect of the proposed context-driven dimensional reduction when the centralities of context networks generate the document vectors because they often yield to non-trivial matrix transformations. Therefore many experiments were conducted. The experiments should prove or reject two hypotheses:

- The proposed representations encode and preserve the contextual relations together with the information about the document content.
- The proposed representations are useful for the processing of documents in text mining tasks.

The first hypothesis was proved; the proposed vectors can serve as the carriers of context inter-document relations among linguistic entities such as hidden topics. The proposed representations particularly maintain the contextual diversity of documents. The reduction of context information that is caused by the projection of complex network structures to vectors differs for different experimental parameters. Our representations are especially useful in the situations where the low dimensional document vectors are required which is the common

[134] Their results are also rather variable.

situation when the documents should be processed by data mining models. More precisely, the maintenance of the contextual relations by the proposed representations is the most evident for a small number of topics and a short contextual window.

Some proposed encoding schemas are capable to preserve the context well, some are more content sensitive. A very important parameter is the selected centrality of the context network. The centralities that exploit the lengths of paths through the context network are the most promising carriers of the contextual information. The tested centrality Betweenness is capable to preserve 50-70% of the contextual diversity of a real document collection if the context is considered only within the identified sentences.

The second hypothesis about the usefulness of the encoding of the context relations for common text mining predictive modeling was not generally proved; the contextual information encoded into the proposed document representation is exploited by data mining models only in the special tasks when the document content is unimportant. We showed that the proposed representations can be successfully used for the recognition of grammatically incorrect documents where the order of words or topics was randomly mixed. Again the centralities that depend on the path lengths through the context network like Betweenness perform best and short context windows are preferable.

On the other hand, the performance of predictive models in common text mining tasks (the information retrieval, the document classification and the document clustering) is not significantly affected by adding any contextual information into input document vectors regardless to the fact that the proposed representations maintain the contextual documents diversity well. In these tasks the models exploit preferably content dependent attributes rather than contextual patterns. Hence the proposed representations are useful in the specific context dependent tasks only where documents share the same content and they are distinguishable by their context only.

Therefore we can conclude that the encoded content is not generally the most important part of the document representation in the text mining field. Therefore text miners should preferably focus on the selection of the features that properly describe the document content and they can usually neglect the ordering of linguistic entities within documents. It makes the common text mining tasks different from NLP tasks (e.g. machine translation) where the usage of contextual information is critical. Due to these findings the input data for the text mining tasks can be prepared faster; one cannot compromise between the accuracy of models and the speed of extraction of any context related features. The main effort in the data preparation phase of the common text mining projects when unstructured texts are transformed to structured vectors should be paid to the extraction of the reasonable number of content features that describe each document as a whole; with a few task specific exceptions a vector representation that serves as an input for a common text mining task can ignore the order in which the features appear in documents.

References

Aggarwal, C.C. & Zhai, C., 2012. A Survey of Text Clustering Algorithms. In Aggarwal, C.C. & Zhai, C. *Mining Text Data*. Springer. pp.77-128.

Aldous, D.J., 1983. Exchangeability and related topics. In Aldous, D.J. *École d'Été de Probabilités de Saint-Flour XIII*. Springer. pp.1-198.

Anick, P.G. & Vaithyanathan, S., 1997. Exploiting clustering and phrases for context-based information retrieval. In *ACM SIGIR conference on Research and development in information retrieval*. New York, 1997. ACM.

Bengio, Y., Durcharme, R., Vincent, P. & Janvin, C., 2003. A neural probabilistic language model. *The Journal of Machine Learning Research*, pp.1137-55.

Berry, M., Dumais, S. & O'Brien, G., 1995. Using Linear Algebra for Intelligent Information Retrieval. *SIAM Rev.*, pp.573-95.

Bizer, C. et al., 2009. DBpedia - A crystallization point for the Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web*, September. pp.154-65.

Blei, D., Ng, A. & Jordan, M., 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, pp.993-1022.

Borg, I. & Groenen, P.J.F., 2005. *Modern multidimensional scaling: theory and applications*. Springer.

Brill, E., 1992. A simple rule-based part of speech tagger. In *Proceedings of the third conference on Applied natural language processing.*, 1992.

Cavnar, W. & Trenkle, J., 1994. N-gram-based text categorization. In *In Proceedings of SDAIR-94.*, 1994.

Charniak, E., 1997. Statistical Techniques for Natural Language Parsing. *AI Magazine*, pp.33-43.

Chomsky, N., 1956. Three models for the description of language. *IRE Transactions on Information Theory*, pp.113-24.

Cox, T.F. & Cox, M.A.A., 2000. *Multidimensional Scaling*. Chapman and Hall.

Dash, M. & Liu, H., 1997. Feature Selection for Clustering. In *Pacific-Asia Connference on Knowledge Discovery and Data Mining.*, 1997.

de Finetti, B., 1974. *Theory of Probability*. John Wiley & Sons.

Deerwester, S. et al., 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, pp.391-407.

Dempster, P., Laird, N.M. & Rubin, D.B., 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. In *Journal of the Royal Statistical Society.*, 1977.

Dey, L. & Haque, M.S.K., 2008. Opinion mining from noisy text data. In *Proceedings of the second workshop on Analytics for noisy unstructured text data*. New York, 2008. ACM.

Dhillon, I.S. & Dharmendra, M., 2001. Concept Decompositions for Large Sparse Text Data Using Clustering. *Machine Learning*, pp.143-75.

Ding, C., He, X. & Simon, H.D., 2005. On the equivalence of nonnegative matrix factorization and spectral clustering. In *SIAM International Conference on Data Mining.*, 2005.

Earley, J.C., 1970. An efficient context-free parsing algorithm. *Communications of the ACM*, pp.94-102.

Eckart, C. & Young, G., 1936. The approximation of one matrix by another of lower rank. *Psychometrika*, pp.211-18.

Feldman, R. & Sanger, J., 2007. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*.

Forsythe, G.E., Malcolm, M.A. & Moler, C.B., 1977. *Computer Methods for Mathematical Computations*. Prentice Hall.

Freeman, L.C., 1977. A set of measures of centrality based on betweenness. *Sociometry*, pp.35-41.

Gabrilovich, E. & Markovitch, S., 2007. Computing Semantic Relatedness using Wikipedia-Based Explicit Semantic Analysis. In *Proceedings of the 20th international joint conference on Artifical intelligence*., 2007. Morgan Kaufmann Publishers.

Gelman, A., Carlin, J.B., Stern, H.S. & Rubin, D.B., 2009. *Bayesian Data Analysis*. Chapman & Hall/CRC.

Golub, G.H. & Van Loan, C.F., 1996. *Matrix Computations*. JHU Press.

Grefenstette, E. & Pulman, S., 2010. *Analysing Document Similarity Measures*.

Grefenstette, G. & Tapanainen, P., 1994. *What is a word, What is a sentence? Problems of Tokenization*.

Guyon, I. & Elisseeff, A., 2003. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, March. pp.1157-82.

Háva, O., Skrbek, M. & Kordík, P., 2012. Document classification with supervised latent feature selection. In *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics*., 2012. ACM.

Hayes, P.J. & Weinstein, S.P., 1990. Construe-TIS: A System for Content-Based Indexing of a Database of News Stories. In *Proceedings of the The Second Conference on Innovative*., 1990. AAAI Press.

Hofmann, T., 1999. Probabilistic Latent Semantic Indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. New York, 1999. ACM.

Ide, N. & Véronis, J., 1198. Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art. *Computational Linguistics*, pp.1-40.

Jardine, N. & van Rijsbergen, J., 1971. The use of hierarchical clustering in information retrieval. *Information Storage and Retrieval*, pp.217-40.

Kass, R.E. & Steffey, D., 1989. Approximate Bayesian inference in conditionally independent hierarchical models. *Journal of the American Statistical Association*, pp.717-26.

Kleinberg, J.M., 1999. Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, pp.604-32.

Kruskal, J.B., 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, March. pp.1-27.

Landauer, T., Foltz, P. & Laham, D., 1998. An Introduction to Latent Semantic Analysis. *Discourse Processes*, pp.259-84.

Lee, J.A. & Verleysen, M., 2007. *Nonlinear Dimensionality Reduction*. Springer.

Leontief, W.W., 1941. *The Structure of American Economy, 1919-1929*. Harvard University Press.

Lewis, D.D., 1992. Feature Selection and Feature Extraction for Text Categorization. In *Proceedings of Speech and Natural Language Workshop*., 1992. Morgan Kaufmann.

Liu, H., Christiansen, T., Baumgartner, W.A. & Verspoor, K., 2012. BioLemmatizer: a lemmatization tool for morphological processing of biomedical text. *Journal of biomedical semantics*.

Liu, T., Liu, S., Chen, Z. & Ma, W.-Y., 2003. An Evaluation on Feature Selection for Text Clustering. In *International Conference on Machine Learning*., 2003.

Lovins, J.B., 1968. Development of a Stemming Algorithm. *Mechanical Translation and Computational Linguistics*, pp.22-31.

Luhn, H.P., 1957. A Statistical Approach to Mechanized Encoding and Searching of Literary Information. *IBM Journal of Research and Development*, pp.309-17.

Luhn, H.P., 1958. *Auto-encoding of Documents for Information Retrieval Systems*. IBM Research Center.

Mani, I., 2001. *Automatic Summarization*. John Benjamins Publishing.

Markov, A.A., 1913. An Example of Statistical Investigation of the Text Eugene Onegin Concerning the Connection of Samples in Chains. *Izvistia Imperatorskio Akademii Nauk*, pp.153-62.

Mikolov, T. et al., 2011. Empirical Evaluation and Combination of Advanced Language Modeling Techniques. In *Proceeding of INTERSPEECH 2011*., 2011.

Page, L., Brin, S., Motwani, R. & Winograd, T., 1999. *The PageRank Citation Ranking: Bringing Order to the Web*. Stanford InfoLab.

Paukkeri, M.-S. et al., 2011. Effect of Dimensionality Reduction on Different Distance Measures in Document Clustering. In *Proceedings of 18th International Conference on Neural Information Processing*., 2011.

Porter, M.F., 1980. An Algorithm for Suffix Stripping. *Program*, pp.130-37.

Porter, M., 2000. *Snowball*. [Online] Available at: http://snowball.tartarus.org/.

Qin, T. et al., 2005. A Study of Relevance Propagation for Web Search. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*., 2005. ACM.

Ratnaparkhi, A., 1996. A Maximum Entropy Model for Part-Of-Speech Tagging. In *Proceedings of the Empirical Methods in Natural Language Processing*., 1996.

Rijsbergen, C.J.v., 1975. *Information Retrieval*. Butterworths.

Robertson, S.E. & Zaragoza, H., 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval*, pp.333-89.

Saayes, Y., Inza, I. & Larranaga, P., 2007. A review of feature selection techniques in bioinformatics. *Bioinformatics*, pp.2507-17.

Sabidussi, G., 1966. The centrality index of a graph. *Psychometrika*, pp.581-603.

Salton, G., 1983. *An Introdution to Modern Information Retrieval*. Mc Graw Hill.

Salton, G. & Buckley, C., 1988. Term-weighting Approaches in Automatic Text Retrieval. *Information Processing & Management*, pp.513-23.

Salton, G.M., Wong, A. & Yang, C.-S., 1975. A vector space model for automatic indexing. *Communications of the ACM*, pp.613 - 620.

Schwenk, H., 2007. Continuous space language models. *Computer Speech & Language*, pp.492–518.

Sebastiani, F. & Delle Ricerche, C.N., 2002. Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, pp.1-47.

Seeley, J.R., 1949. The net of reciprocal influence: A problem in treating sociometric data. *The Canadian Journal of Psychology*.

SPSS Inc., 2008. Mastering New Chalenges in Text Analytics.

Sruthi, K. & Reddy, V., 2013. Document Clustering on Various Similarity Measures. *International Journal of Advanced Research in*, pp.1269-73.

Takane, Y., Young, F.W. & de Leeuw, J., 1977. Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features. *Psychometrika*, March. pp.7-67.

Tomita, M., 1986. *Efficient Parsing for Natural Language*. Kluwer Academic Publishers.

Torgerson, W.S., 1952. Multidimensional scaling: I. Theory and method. *PSychometrika*, December. pp.401-19.

Tsoumakas, G. & Katakis, I., 2007. Multi-Label Classification: An Overview. *International Journal of Data Warehousing and Mining*, pp.1-13.

Verbeek, J.J., 2000. Supervised Feature Extraction for Text Categorization. In *Tenth Belgian-Dutch Conference on Machine Learning*., 2000.

Vinh, X.N., Epps, J. & Bailey, J., 2010. Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance. *The Journal of Machine Learning Research*, pp.2837-54.

Wilbur, J.W. & Sirotkin, K., 1992. The automatic identification of stopwords. *Journal of Information Science*, pp.45-55.

Xu, J. & Croft, B.W., 1998. Corpus-Based Stemming using Cooccurrence of Word Variants. *ACM Transactions on Information Systems*, pp.61-81.

Xu, W. & Gong, Y., 2003. Document clustering by concept factorization. In *ACM Special Interest Group on Information Retrieval*., 2003. ACM.

Xu, W., Liu, X. & Gong, Y., 2003. Document Clustering based on non-negative matrix factorization. In *ACM Special Interest Group on Information Retrieval*., 2003. ACM.

Yang, Y., 1995. Noice Reduction in a Statistical Approach to Text Categorization. In *ACM Special Interest Group on Information Retrieval*., 1995. ACM.

Yang, Y. & Pedersen, J.O., 1997. A Comparative Study on Feature Selection in Text Categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning*., 1997. Morgan Kaufmann.

Yang, Y. & Pedersen, J.O., 1997. A Comparative Study on Feature Selection in Text Categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning*., 1997. Morgan Kaufmann Publishers.

Yihong, G. & Xin, L., 2001. Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*., 2001. ACM.

Appendix A: Variable notation overview

$V = \{w_1, w_{2,} \ldots, w_N\}$	Vocabulary, a set of terms of the length $\lvert V\rvert=N$
$d = w_{(1)} w_{(2)} w_{(3)} \ldots w_{(L_d)}$	Document, a sequence of vocabulary terms of the length L_d
$D = \{d_1, d_2, \ldots, d_M\}$	Collection, a set of M documents
$\mathbf{d}^T = (v_1, v_{2,} \ldots, v_N)$	Document vector, a structured representation in the vocabulary space
$\mathbf{D} = (\mathbf{d}_1^T; \mathbf{d}_2^T; \ldots; \mathbf{d}_M^T)$	Document-term $M \times N$ matrix, document vectors are in rows
$e = w_{(1)} w_{(2)} \ldots w_{(n)}$	n-gram, a subsequence of n vocabulary terms
$E = \{e_1, e_2, \ldots, e_S\}$	n-gram vocabulary, a set of all possible n-grams of equal length that can be constructed using the vocabulary terms, $\lvert E\rvert=S=N^n$
$f = w_{(1)} w_{(2)} \ldots w_{(n-1)}$	$(n-1)$-gram, a subsequence of $(n-1)$ vocabulary terms
$F = \{f_1, f_2, \ldots, f_R\}$	$(n-1)$-gram vocabulary, a set of all possible $(n-1)$ grams of equal length that can be constructed using the vocabulary terms, $\lvert F\rvert=R=N^{(n-1)}$
$\mathbf{p}_r^T = (p_{r1}, p_{r2,} \ldots, p_{rN})$	Vector of conditional probabilities of transitions from $(n-1)$-gram f_r to vocabulary terms, $p_{rn} = p(f_r \rightarrow w_n), \sum_{j=1}^{N} p_{rj} = 1$
$\mathbf{P} = (\mathbf{p}_1^T; \mathbf{p}_2^T; \ldots; \mathbf{p}_R^T)$	$R \times N$ matrix of conditional probabilities of transitions from $(n-1)$-grams to vocabulary terms, probability vectors are in rows, $p_{rn} = p(f_r \rightarrow w_n), \sum_{j=1}^{N} p_{rj} = 1$
$\mathbf{Q}$	$R \times N$ matrix of probabilities of n-grams $q_{rn} = p(f_r w_n) =, \sum_{i=1}^{R}\sum_{j=1}^{N} q_{ij} = 1$
$\mathbf{T}$	$R \times N$ matrix of counts of transitions from $(n-1)$-grams to vocabulary terms in a document, $\sum_{i=1}^{R}\sum_{j=1}^{N} t_{ij} = L$
$\mathbf{G}$	$N \times N$ square matrix of co-occurrence frequencies of vocabulary terms in a context window of the length K in a document, $\sum_{i=1}^{N}\sum_{j=1}^{N} g_{ij} = LK$
$G = \{V, \mathbf{G}\}$	Weighted context network of a document, vertices V represent vocabulary terms and weights $\mathbf{G}$ are equal to co-occurrence frequencies of the terms in a context window
$\overline{G} = \{V, \overline{\mathbf{G}}\}$	Weighted context network of a document, vertices V represent vocabulary terms and weights $\overline{\mathbf{G}}$ are equal to inversed co-occurrence frequencies of the terms in a context window
$\mathbf{c}^T(G) = (c_1, c_2, \ldots, c_N)$	Vector of network centralities, the following centralities are taken into account: Authority, Betweenness, Closeness, Degree, Eigenvector, Hub, InDegree, OutDegree, PageRang

Table 20: The list of the variables used in formulas.

Appendix B: Publications of author

publication	author's share	thesis related	impacted journal
Háva, O., Skrbek, M., Kordík, P., 2010. Fast supervised feature extraction from structured representation of text data. In *Proceedings of the 7th EUROSIM Congress on Modelling and Simulation*, 2010. Vydavatelství ČVUT.	100%	yes	no
Háva, O., Skrbek, M., Kordík, P., 2012. Supervised two-step feature extraction for structured representation of text data. *Simulation Modelling Practice and Theory*, 2012. ELSEVIER.	100%	yes	yes
Háva, O., Skrbek, M., Kordík, P., 2012. Document Classification with Supervised Latent Feature Selection. In *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics*, 2012. ACM.	100%	yes	no
Háva, O., Skrbek, M., Kordík, P., 2012. Contextual latent semantic networks used for document classification. In *Proceedings of the 4th International Conference on Knowledge Discovery and Information Retrieval*, 2012. SciTePress.	100%	yes	no
Háva, O., Skrbek, M., Kordík, P., 2013. Vector Representation of Context Networks of Latent Topics. In *Proceedings of the World Congress on Engineering 2013*, 2013. IAENG.	100%	yes	no
Alcnauer, J., Háva, O., 2012. New emerging data mining approaches in marketing and education. In *Proceedings of the 3th International Scientific Conference Management 2012*, 2012. Bookman.	50%	no	no
Háva O., 2007: *Data mining v praxi:* On-line ohodnocení rizika podvodu v neživotním pojišťění, *IT Systems 3/2007*	100%	no	no
Háva O., 2007: Marketingové dataminingové úlohy v telekomunikacích, *IT Systems 6/2007*	100%	no	no
Háva O., 2007: On-line optimalizace marketingových kampaní, *IT Systems 10/2007*	100%	no	no
Háva O., 2007: Data mining okolo nás, *Professional Computing 10/2007*	100%	no	no
Háva O., 2007: Poznejte svá data, *Professional Computing 11/2007*	100%	no	no
Brom O., Háva O., 2007: Příprava dat na modelování, *Professional Computing 12/2007*	50%	no	no
Háva O., 2007: Role modelů v data miningovém projektu, *Professional Computing 13/2007*	100%	no	no
Háva O., 2007: Vyhodnocení data miningového řešení, *Professional Computing 14/2007*	100%	no	no
Háva O., 2008: Integrace data miningových modelů do firemních procesů, *Professional Computing 1/2008*	100%	no	no
Háva O., 2008: RFM skórování, *IT Systems Special 2008*	100%	no	no

Table 21: The list of author's publications.