FAKULTA INFORMAČNÍCH TECHNOLOGIÍ ČVUT V PRAZE

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

Název:	Sledování a analýza článků v médiích
Student:	Peter Kanoš
Vedoucí:	Ing. Daniel Vašata, Ph.D.
Studijní program:	Informatika
Studijní obor:	Znalostní inženýrství
Katedra:	Katedra teoretické informatiky
Platnost zadání:	Do konce letního semestru 2018/19

Pokyny pro vypracování

1. Vytvořte aplikaci, která bude sbírat data z internetových medií (v českém nebo anglickém jazyce). Sbíraná data jsou především název článku, perex a text, doplněná případně dostupnými metadaty (kategorie, štítek).

Aplikace může být založená na nějakém jiném dostupném řešení. Aplikace musí sledovat vývoj a změny dat (hlavně titulku) u zaindexovaných článků.

2. Proveďte rešerši nástrojů a metod, které umožňují analýzu nasbíraných dat, například měří relevanci titulku a textu, měří podobnost dvou titulků apod.

3. Proveďte analýzu nasbíraných dat a jejich vývoje v čase. Zaměřte se na trendy ve změnách titulků (a jiných dat) a relevanci titulku vzhledem k dalším datům u článku. Pokuste se porovnat stejné zprávy z různých zdrojů.

4. Na základě vašich poznatků učiňte závěry a navrhněte, jakým způsobem by bylo možné analýzu plně algoritmizovat a dále rozšiřovat.

Seznam odborné literatury

Dodá vedoucí práce.

doc. Ing. Jan Janoušek, Ph.D.
vedoucí katedry

doc. RNDr. Ing. Marcel Jiřina, Ph.D.
děkan

V Praze dne 4. února 2018

FAKULTA
INFORMAČNÍCH
TECHNOLOGIÍ
ČVUT V PRAZE

Bakalárska práca

Sledování a analýza článků v médiích

Peter Kanoš

Katedra teoretické informatiky
Vedúci práce: Ing. Daniel Vašata, Ph.D.

14. mája 2018

Poďakovanie

Poďakovanie patrí v prvom rade mojím rodičom, ktorí za mnou celý čas štúdia stáli, bez ktorých podpory a lásky by som to nezvládol. Ďalej by som chcel poďakovať za všetku lásku, podporu a pomoc mojim trom úžasným súrodencom. A v neposlednom rade patrí obrovská vďaka za trpezlivosť, námahu a pomoc pri písaní mojej práce vedúcemu práce Ing. Danielovi Vašatovi Ph.D.

Prehlásenie

Prehlasujem, že som predloženú prácu vypracoval(a) samostatne a že som uviedol(uviedla) všetky informačné zdroje v súlade s Metodickým pokynom o etickej príprave vysokoškolských záverečných prác.

Beriem na vedomie, že sa na moju prácu vzťahujú práva a povinnosti vyplývajúce zo zákona č. 121/2000 Sb., autorského zákona, v znení neskorších predpisov, a skutočnosť, že České vysoké učení technické v Praze má právo na uzavrenie licenčnej zmluvy o použití tejto práce ako školského diela podľa § 60 odst. 1 autorského zákona.

V Prahe 14. mája 2018 .....

Odkaz na túto prácu

Kanoš, Peter. *Sledování a analýza článků v médiích.* Bakalárska práca. Praha: České vysoké učení technické v Praze, Fakulta informačních technologií, 2018.

Abstrakt

Práca sa zaoberá implementáciou aplikácie pre zbieranie článkov a ich verzií v čase z českých spravodajských serverov iDnes.cz a Aktuality.cz. Následne analýzou týchto článkov vykonanou nástrojom Doc2Vec. Analýza týchto článkov je zameraná najmä na zmeny článkov v čase a porovnávanie podobností medzi ich časťami. Zmeny sa týkali najmä titulkov, perexov článkov a textov daných článkov. Skúmané boli najmä závislosti rôznych faktorov ako sú napríklad čas vydania článku, problematika ktorou sa článok zaoberá a podobne. Výsledkom práce samotnej je aplikácia naprísaná v jazyku Python.

Klíčová slova Python, text mining, sémantická analýza, Paragraf vektor

Abstract

This thesis deals with implementation of application for collection of the articles and their version in the time from czech news servers iDnes.cz and Aktuality.cz. The analysis is subsequently done by Doc2Vec. Analysis of these articles is focused on changes during the time and comparison of similarities between their sections. The changes refer to titles of the articles, perexes of

the articles,text of the articles. Examined were mainly relations between differents factors such as time of publication of the article, article's main issues etc. The result of the thesis is an application written in the Python language.

Keywords Python, text mining, semantic analysis, Paragraph vector

Obsah

Zoznam obrázkov

Zoznam tabuliek

Úvod

V dnešnej dobe, keď väčšina najmä mladých ľudí získava informácie o dianí vo svete prednostne z elektronických médií je téma automatického spracovanie textu veľmi aktuálna. Zo všetkých strán sa na nás valí obrovské množstvo odkazov, ktoré vábia svojimi expresívnymi titulkami. Veľmi často sa však stane, že po prekliknutí sa na odkazovaný obsah a jeho prečítaní zostaneme sklamaní niekedy až rozčarovaní z toho, aký obsah sa skrýval za daným titulkom. Tento jav je obecne známy pod pojmom clickBait. Jeden z dôvodov prečo som si vybral túto tému bolo pokúsiť sa pomôcť eliminovať tieto typy článkov na sociálnych sieťach. Pevne verím, že na základe výsledkov práce by sa dalo týmto smerom nadviazať a urobiť tak program, ktorý by pomáhal tieto clickBaity odhaľovať. Ďalší dôvod bol pre mňa záujem o text mining ako taký a jeho využitie v oblasti masmédií. Zaujímalo ma tiež ako často a hlavne v akom rozsahu sa menia články po ich vydaní. Čoho najmä sa dané zmeny týkajú a či sa v týchto zmenách dajú nájsť nejaké súvislosti. V neposlednom rade bolo motiváciou zaoberať sa oblasťou text miningu, ktorá nie je veľmi prebádaná a tou je sémantická analýza textov s krátkym rozsahom, obzvlášť v českom jazyku.

Cieľom práce bolo navrhnúť a implementovať aplikáciu, ktorá bude zbierať dáta z internetových médii v českom jazyku. Následne takto pozbierané dáta spracovať a pretvoriť do jednotnej formy. Hlavnou časťou práce bola potom analýza týchto dát a diskusia nad výslednou analýzou. V analýze som sa zaoberal najmä zmenami článkov v ich perexoch, titulkoch a samotných textoch. Ďalej podobnosťou medzi textom a perexom, perexom a titulkom a titulkom a textom článkov.

Práca samotná je štrukturovaná tak, aby čitateľovi poskytla v prvých kapitolách najmä dobrý úvod do problematiky ako takej. V jednotlivých kapitolách sa postupne budem zaoberať teoretickými základmi v oblasti data miningu a text miningu. V kapitole pojednávajúcej o data miningu vysvetlím elementárne základy a pojmy týkajúce sa danej problematiky. V kapitole

o text miningu sa pokúsim bližšie popísať jednotlivé možnosti spracovania textu. Následne sa dostanem k rešeršnej časti, ktorá sa zaoberá hľadaním vhodného softwaru pre analýzu a porovnanie textu, v ktorej závere vyberiem vhodné riešenie pre daný problém. Nasledujúca kapitola popisuje ako vybrané riešenie funguje. Potom nasleduje kapitola, ktorá pojednáva o web scrapingu a crawlingu za ktorou zhodnotím české spravodajské servery a vyberiem dva, nad ktorými bude zrealizovaná samotná analýza. V praktickej časti rozoberiem implementáciu samotného crawleru a poukážem na štruktúru vybraných stránok. Následne zoznámim čitateľa s výsledkami analýzy a vyvodím závery.

Kapitola 1

Úvod do problematiky

Súčasný rozvoj technologií dovoľuje získavať, a uskladňovať dáta v obrovských množstvách. Všetky tieto dáta v sebe nesú užitočné informácie. Všade kam sa pozrieme zanechávame o sebe určitý druh informácií. Týka sa to najmä internetových obchodov, médií, fór a sociálnych sietí, kde ide o obyčajné veci, ako čo máme vo zvyku nakupovať, akú hudbu počúvame atp. Tieto ale iné druhy dát sú zbierané, uchovávané a prenášané v dátových úložiskách naprieč všetkými oblasťami nášho každodenného života. Pri takomto množstve dát prichádza otázka ako je z nich možné čo najefektívnejšie vyextrahovať to, čo nás zaujíma a previesť to v znalosť. Nájdenie užitočných trendov, vzorov a anomálií v dátach a pretvorenie týchto informácií v znalosti sa preto stalo jedným z hlavných odvetví súčastnej informatiky. Toto odvetvie je všeobecne známe ako data mining.

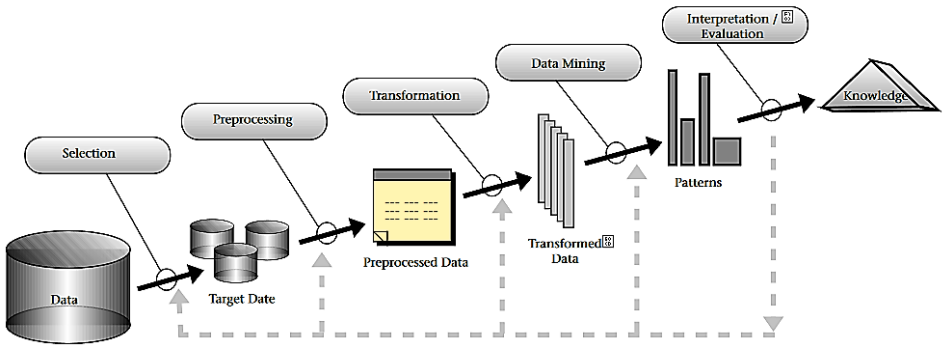
1.1 Data mining

Data mining alebo tiež vyťažovanie znalostí z dát, je proces získavania užitočných vzorov a znalostí z veľkého množstva dát. Data mining sa často definuje dvojako, jednou definíciou je, že data mining je len jedným krokom v procese získavania znalostí z dát. V ostatných prípadoch je data mining označovaný ako celý proces získavania znalostí z dát. Kde získavanie znalostí z dát alebo tiež KDD (Knowledge Discovery from data) zahŕňa nasledujúce kroky.

1. Čistenie dát - Odstránenie chybných a nekonzistentných dát.

2. Integrácia dát - Kombinovanie mnohých datových zdrojov.

3. Selekcia dát - Výber relevantných údajov o sledovaných objektoch z dát uložených v datovom sklade.



Obr. 1.1: Schéma získavania znalostí z databáz [1]

4. Transformácia dát - Dáta sú transformované a upravované do foriem vhodných na vykonávanie súhrnných alebo agregačných operáci. Ide napríklad o diskretizáciu spojitých veličín.

5. Data mining - Samotný proces hľadania súvislostí za pomoci vopred vybraných analytických metód.

6. Hodnotenie vzorov - Identifikácie skutočne zaujímavých vzorov a modelov reprezentujúcich informácie, ktoré sú predmetom nášho záujmu.

7. Interpretácia znalostí - Spracovanie výsledkov použitých metód do tvaru zrozumiteľného pre koncového užívateľa [4].

.

Na obrázku 1.1 je zobrazený proces vyťažovania znalostí z databáz. Tento proces sa od vyťažovania znalostí z dát líši len v prvých dvoch krokoch, pretože sa v ňom predpokladá už vopred zozbieraná množina dát.

Data mining sa zaoberá spracovaním a dolovaním informácií z dát v rôznych formách. Najbežnejšou formou dát v akej je možné dáta spracovať pomocou data miningových metód sú dáta uložené v databázach. Okrem nich je tu aj široké spektrum iných foriem dát ako napríklad multimediálne dáta, webové dáta alebo textové dáta. Na základe týchto typov dát sa ďalej dajú definovať rôzne typy oborov, ktoré sa zaoberajú vyťažovaním informácií z týchto foriem Vzhľadom na to, že práca sa ďalej zaoberá web minigom a text miningom je pre nás dôležité si definovať aspoň tieto dve odvetvia.

1.2 Web mining

Obor informatiky ktorý sa zaoberá dolovaním dát z webu sa nazýva web mining. Web mining je aplikácia data miningových techník k objavovaniu vzorov,

štruktúr a znalostí z webu [4, str. 597]. Z hľadiska zamerania analýzy sa dá web mining ďalej rozdeliť ešte na tri skupiny.

Web content mining zahŕňa analýzu obsahu webu teda napríklad text a multimédiá.

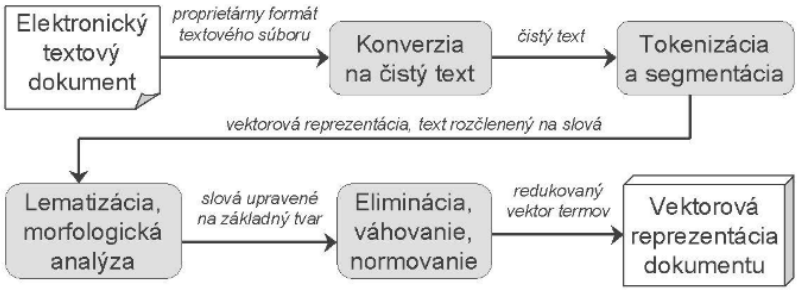
Web structure mining skúma štruktúru webu, z hľadiska stránok ako uzlov a hypertextových odkazov ako ich spojení medzi sebou.

Web usage mining je dolovanie znalostí z hľadiska užívateľských trendov, teda skúma napríklad aké stránky určitá skupina užívateľov navštevuje a na základe toho sa snaží napríklad predikovať ich budúce správanie.

1.3 Text mining

Text mining (slovensky dolovanie v textoch) sa dá definovať ako proces získavania znalostí, ktorý má za cieľ identifikovať a analyzovať užitočné informácie v textoch [5]. Text mining je možné využívať na riešenie širokej škály problémov: vyhľadávanie informacií, automatická kategorizácia, zhlukovanie dokumentov, extrakcia informacií, sumarizácia textov.

Text mining ako taký sa skladá z dvoch základných častí a to predspracovanie vstupného dokumentu a následné získavanie znalostí. Celý proces text miningu sa dá potom jednoducho namapovať na už vyššie spomenutý proces vyťažovania znalostí z databáz. Z tohto celého procesu je najzložitejšou a najviac časovo náročnou časťou predspracovanie textových dát. Správne predspracovanie je základným predpokladom pre úspešné a hlavne správne aplikovanie metód dolovania znalostí. Samotný prevod textu v tejto fáze sa skladá z nasledujúcich krokov.



Obr. 1.2: Schéma predspracovania textových údajov [2]

Prvý krok, teda prevod z elektronického textového dokumentu na čistý text zahŕňa odstránenie všetkých netextových informácií. Netextovými informáciami sú myslené rôzne tabuľky, grafy, obrázky, videá a podobne. V

tejto práci to bolo hlavne odstránenie html a css tagov z príslušných webových stránok. Ďalším krokom v tomto procese je tokenizácia a segmentácia. Segmentáciou sa rozumie rozdelenie čistého textu do základných jednotiek. Základnou jednotkou textu sa myslí každý znak, či skupina znakov oddelená bielym znakom. Tieto sú potom v procese tokenizácie spájané do pojmov, ktoré zodpovedajú nejakému slovníkovému výrazu. Výsledné fragmenty textu sa nazývajú tokeny.

Lematizácia a morfologická analýza

Lematizácia je proces pri ktorom sa slová prevádzajú na ich základný tvar. Toto sa deje najčastejšie odstraňovaním predpôn a prípon. Nasleduje morfologická analýza, v tomto procese sa slová označia morfologickými značkami. Tieto popisujú o aký slovný druh, rod, pád a podobne sa jedná.

Eliminácia, váhovanie, normovanie:
Pri eliminácii prebieha odstránenie takzvaných stopwords. Tieto slová, ako už sám názov napovedá nenesú v sebe veľkú informačnú hodnotu o celkovom obsahu textu a preto sú zväčša pre nás zbytočné. Najčastejšie sú to spojky, predložky, príslovky atp. Váhovanie je proces pri ktorom sa termom priradí hodnota podľa počtu výskytov. Toto váhovanie môže byť dvojúrovňové. Prvý spôsob je na úrovni dokumentov samotných - teda lokálne váhovanie. Tým druhým je globálne, kde sa termy váhujú vzhľadom k celému korpusu. Samotné váhovanie termov môže byť prevedené rôznymi prístupmi.

Možností pre konečnú reprezentáciu dokumentov existuje hneď niekoľko. Asi najjednoduchšou je boolovský model. Boolovský model používa základné hodnoty 0,1 a indikuje iba to, či sa daný term v dokumente vyskytuje alebo nie. Boolovský model je možné nahradiť invertovaným indexom, ktorý v sebe uchováva informácie o tom, v akých dokumentoch sa nachádza ten, ktorý term. Ďalšími modelmi reprezentácie sú pravdepodobnostný model a model distribuovanej sémantiky. Tento model stavia na tom, že slová sa často vyskytujú v tom istom kontexte. Kontextom sa v tomto prípade myslí okolie slova. Najviac používaným modelom podľa [2] je Vector space model teda vektorový model. Vektorový model popisuje všetky dokumenty ako vektory termov, ktoré nesú v sebe váhu toho ktorého termu.

1.3.1 Vector space model

Vector space model bol vyvinutý pre SMART [6] systém na vyťažovanie znalostí. Hlavnou myšlienkou vektorového modelu bolo reprezentovať jednotlivé dokumenty z kolekcie ako body v priestore a tak umožniť výpočet sémantickej podobnosti textov. Body, ktoré sú si bližšie v danom priestore sú si sémanticky podobnejšie a body od seba viac vzdialené v priestore sú významovo odlišnejšie. V práci [7] sú rozlišované tri druhy vektorových mode-

lov na základe typu matice a to term–document, word–context a pair–pattern matica.

Word-Context zakladá na tom, že slová ktoré sa vyskytujú v rovnakých kontextoch majú podobný význam. Slovo je reprezentované vektorom, v ktorom elementy sú odvodené od výskytu daného slova v rôznych kontextoch. Kontext daného slova je daný slovami, frázami, odstavcami, kapitolami atp. Podobné riadkové vektory matice Word-Context teda indikujú podobnosť daných slov.

Term–Document slúži na porovnávanie podobnosti celých dokumentov. Riadkové vektory danej matice korešpondujú s termami a stĺpcové zase s dokumentmi. V tejto matici je dokument reprezentovaný vektorom, ktorý znázorňuje aké slová sa v dokumente nachádzajú. Vzniknutá matica býva spravidla riedka pre mnohonásobne väčší počet slov než dokumentov. Hlavnou nevýhodou takejto reprezentácie dokumentov je strata poradia daných slov.

Pair–Pattern matica, popisuje páry slov a ich vzájomný súvis. Konkrétne sú to teda jej riadkové vektory reprezentujúce páry slov ako murár:dom, pekár:chlieb a podobne, zatiaľ čo stĺpcové vektory tejto matice vyjadrujú ako spolu tieto slová súvisia. Napríklad rezbár:drevo, zlatník:zlato, maliar:plátno majú spoločný vzťah remeselník:materiál.

1.4 Meranie podobnosti

Vzhľadom na to, že prirodzený jazyk je od binárneho kódu vzdialený, automatické vyčíslenie podobnosti dvoch slov či textov sa stáva zaujímavým nie však neriešiteľným problémom. Podľa [8] sa dá na textovú podobnosť pozerať z dvoch uhlov, jedným je sémantická podobnosť a tým druhým lexikálna podobnosť. Lexikálna podobnosť funguje na báze rozdielnosti znakov a výrazov. Napríklad môžeme povedať, že slová kradnúť chradnúť sú si lexikálne podobné. Oproti tomu sémantická podobnosť vyjadruje samotnú podobnosť významu daných slov napríklad môžeme povedať, že ustanoviť a nominovať sú si sémanticky podobné naproti tomu lexikálne sú od seba veľmi odlišné.

Lexikálna podobnosť

Algoritmy pre výpočet lexikálnej podobnosti je možné rozdeliť do dvoch tried a to do výpočtu podobnosti na základe znakov a výpočtu podobnosti na základe výrazov. Výrazom sa v tomto prípade myslí slovo alebo slovné spojenie, ktoré zodpovedajú nejakému slovníkovému výrazu.

Metriky na meranie podobnosti na základe znakov sú:

- najdlhší spoločný podreťazec,

- Levensteinová vzdialenosť,
- Jaro vzdialenosť,
- N-gramová podobnosť.

Metriky na meranie podobnosti na základe výrazov sú:

- Cosínusova podobnosť,
- Centroid based similarity,
- Web Simpson podobnosť,
- Web PMI podobnosť.

Viacej informácií ohľadom vyššie spomenutých podobností nájdete v [8].

Sémantická podobnosť

Sémantická podobnosť sa dá ďalej rozdeliť na Corpus based Similarity (podobnosť na základe korpusu) a Knowledge based similarity (podobnosť na základe znalostí).

Corpus based Similarity sa snaží identifikovať podobnosť medzi slovami na základe informácií získaných z velkých korpusov textov. Tieto korpusy väčšinou obsahujú veľké množstvo dokumentov z rôznych oblastí v v danom jazyku [9].

Knowledge based similarity vyjadruje sémantickú podobnosť dvoch slov na základe informácií získaných zo sémantických sietí ako je napríklad WordNet.

WordNet [10] je lexikálna databáza angličtiny, obsahujúca slová zoskupené do takzvaných synsetov. Synsety sú súbory synonymických slov popisujúci rovnaký koncept, ktoré sú navzájom prepojené pomocou konceptuálne-sémantických vzťahov. Vďaka tejto štruktúre sa stal WordNet užitočným nástrojom na poli spracovania prirodzeného jazyka.

Vyššie spomenuté metódy usporiadajú vektory reprezentujúce slová do vektorového priestoru na základe ich významu. Zostáva nám už len vzdialenosti medzi týmito vektormi zmerať. Medzi najznámejšie metriky na porovnávanie vektorov podľa [2] patria napríklad:

Euklidovská vzdialenosť: $\sum_{i=1}^{N} \sqrt{(a_i - b_i)^2}$

Manhatanská vzdialenosť: $\sum_{i=1}^{N} |a_i - b_i|^2$

V oboch týchto metrikách sa predpokladá rovnaká dĺžka vektorov, medzi ktorými sa určuje podobnosť. Vo vzťahoch d_i a d_j predstavujú vektory dokumentov v ktorých a_x a b_x sú frekvencie výskytov slov v dokumentoch. Konštanta N je počet slov, teda dĺžka vektorov d_i a d_j.

Najviac preferovaná zostáva však podľa [8] Cosínusova podobnosť.

Cosínusova podobnosť [11] sa používa v širokej škále problémov v oblasti data miningu. V oblasti text miningu napríklad vo vyhľadávačoch, kde uživateľský dotaz je prevedený do vektorovej reprezentácie a následne sa nechá porovnať Cosínusovou podobnosťou s vektormi prehľadávaných dokumentov. Pre dva N dimenzionálne vektory d_i a d_j sa ich cosínusova podobnosť medzi nimi vypočíta ako:

$$\text{similarity}(d_i, d_j) = \frac{d_i \cdot d_j}{|d_i||d_j|} = \frac{\sum_{i=1}^{N} a_i \cdot b_i}{\sqrt{\sum_{i=1}^{N} a_i^2}\sqrt{\sum_{i=1}^{N} b_i^2}}$$

Vo vzťahu d_i a d_j predstavujú vektory dokumentov v ktorých a_x a b_x sú prvky týchto vektorov. Konštanta N je dĺžka vektorov d_i a d_j.

Cosínusova podobnosť môže byť v rozsahu od -1 pre opačné vektory (zvierajúce 180 stupňový uhol) po 1 pre vektory rovnobežné (zvierajúce 0 stupňový uhol).

Kapitola 2

Existujúce implementácie textovej analýzy

S veľkým progresom Deep learningu najmä v posledných rokoch, nastal aj obrovský rozmach nových algoritmov pre spracovanie prirodzeného jazyka. Najväčšie pokroky sa udiali najmä v oblastiach rozoznávania reči a automatického prekladu. Spolu s objavovaním nových riešení a algoritmov v oblasti NLP prišlo na trh aj veľké množstvo softwaru pre spracovanie prirodzeného jazyka dostupného či už ako open-source alebo ako platený produkt. Vzhľadom na veľký počet týchto nástrojov som vybral pár, ktoré boli pre mňa najzaujímavejšie a tieto v tejto časti práce stručne predstavím. Zároveň vyberiem v závere tejto kapitoly metódu, vhodnú pre analyzovanie mnou zozbieraných dát. Môj výber následne odôvodním a vybrané riešenie v nasledujúcej kapitole bližšie predstavím.

2.1 NLTK

Natural Language Toolkit [12] je knižnica v jazyku Python, ktorá poskytuje skvelý základ pre spracovanie prirodzeného jazyka. Knižnica je použiteľná napríklad na klasifikáciu, zhlukovanie, stemmatizáciu, tokenizáciu, taggovanie a mnoho ďalších základných procesov predspracovania textu. Má vstavané veľké množstvo korpusov a trénovacích modelov. Nástroj je open-source, jeho veľkými výhodami sú živá komunita a veľmi dobre spracované tutoriály a dokumentácia. Patrí medzi najviac využívané nástroje v oblasti text miningu.

2.2 SEMILAR

Semantic similarity toolkit [13] je nástroj na výpočet sémantickej podobnosti textov napísaný v jazyku Java a je dostupný ako knižnica. Implementované je avšak už aj GUI rozhranie, ktoré je podporované operačnými systémami Linux

a Windows. SEMILAR obsahuje mnoho algoritmov na výpočet sémantickej podobnosti ako napríklad LSA alebo LDA. SEMILAR v sebe tak isto obsahuje sadu nástrojov na predspracovanie textov ale aj komponentu na vizualizáciu dát a mnoho ďalších. SEMILAR má tak isto zabudovaný aj vlastný korpus, ktorý obsahuje 701 párov oanotovaných výrazov. Tiež obsahuje ladiace a testovacie zariadenia pre výber vhodného modelu. Jeho nevýhodou však je, že je stále dostupný len pre anglický jazyk.

2.3 Text Similarity API od Rxnlp

Text Similarity API [14] je software na meranie podobností textov. Text je možné porovnávať na základe troch podporovaných metrík a to Cosinus, Jaccard a Dice. Software je možné využívať zdarma pokiaľ užívateľ splní podmienky, ktorými sú registrácia a obmädzený prístup k nástroju. K programu je možné pristupovať cez klienta implementovaného v jazyku Python. Nevýhodou je, že veľkosť porovnávaného textu môže byť len do objemu 1MB. Vstup môže byť buď plain text alebo text už predspracovaný lemmatizáciou, tokenizáciou odstránením stop slov a pod. Výstup je vo formáte JSON a obsahuje výsledok z porovnania všetkých troch spomínaných metrík plus ich priemer. Okrem iného má firma Rxnlp viacero APIs napríklad na zhlukovanie výrazov, extrakciu tém alebo na prevod HTML stránky do čistého textu. Má aj voľne dostupnú Python knižnicu pod názvom PyRXNLP, ktorá obsahuje algoritmy pre extrakciu tém a zhlukovanie výrazov.

2.4 DKPro Similarity

DKPro Similarity [15] je Java open source framework na výpočet textovej podobnosti. Framework dopĺňa DKPro Core, čo je vlastne celá kolekcia komponent pre spracovanie prirodzeného jazyka postavená na Apache UIMA frameworku. DKPro similarity zahŕňa širokú škálu metrík na meranie podobnosti. Od tých založených na n-gramoch a spoločných subsekvenciách až k latentnej sémantickej analýze. Program obsahuje dva módy pre porovnanie textu. Tým prvým je stand-alone mode, ktorý umožňuje používať komponenty na výpočet podobnosti nezávisle na sebe v akomkoľvek experimentálnom nastavení. Tento mód však neobsahuje prostriedky na iné predspracovanie textu napríklad na lemmatizáciu. Tým druhým je UIMA-coupled Mode, v ktorom je výpočet podobnosti prepojený s nástrojmi na preprocess zo sady postavenej na UIMA frameworku.

2.5 Gensim

Gensim [16] je open source knižnica napísaná v jazyku Python jej autormi sú Radim Řehuřek a Petr Sojka. Gensim je softwarový framework na spracovanie

prirodzeného jazyka, prispôsobený na tématickú inferenciu - teda určovanie tém textov.
Jeho základnou ideou, na rozdiel od ostatných frameworkov v tejto oblasti, je nezávislosť na veľkosti korpusu. Tá vyplýva z nutnosti spracúvať korpusy mnohokrát väčšie než pamäť RAM. Nezávislosť na veľkosti korpusu je zabezpečená streamovaním dokumentov.

Streamovanie dokumentov prebieha v iteratívnom duchu a teda, že sú dokumenty spracúvané postupne a v žiadnom bode spracovania nie je potrebné držať celý korpus v pamäti. Korpus je v tomto prípade reprezentovaný ako sekvencia jednotlivých dokumentov.

Ďalšou dôležitou výhodou frameworku je jednoduchá inštalácia a intuitívne API. Api Gensimu podporuje možnosť práce s dokumentmi reprezentovanými v tvare riedkych matíc.
Je tu ďalej možnosť ukladania natrénovaného modelu na disk a ich následného nahratia z disku. Jeho druhým interfaceom je interface transformácií, ktorý umožňuje transformáciu z jedného vektorového priestoru do iného. Gensim má implementované množstvo vektor space algoritmov ako LDA a LSA. Okrem iného sú v Gensim implementované modely Word2Vec a Doc2Vec. Doc2Vec ponúka ako jediný možnosť porovnávať celé texty na základe ich obsahu. Doc2Vec bol nakoniec hlavným dôvodom, prečo som sa rozhodol použiť tento nástroj. Ďalšími dôvodmi môjho výberu boli dostupnosť a výborné vlastnosti jazyka Python v oblasti analýzy dát

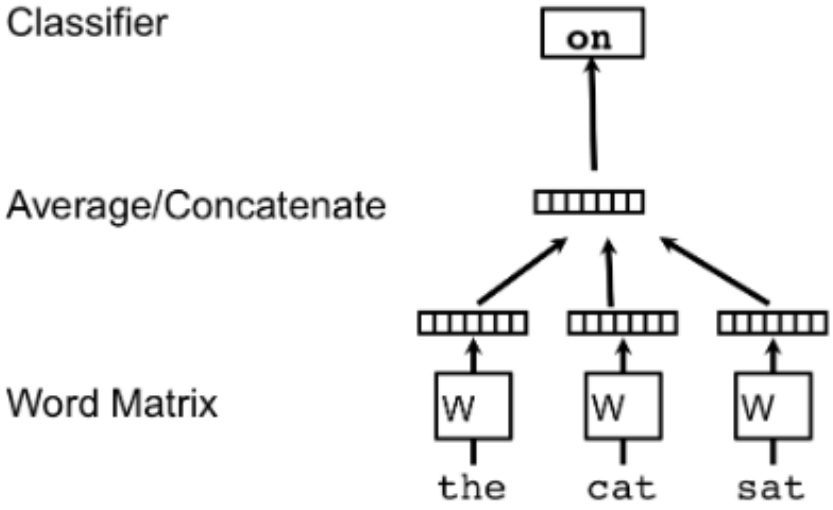
2.5.1 Paragraf vektor

Paragraf vektor je framework, ktorý funguje na princípe umelých neuronových sietí s učením bez učiteľa. Učí sa spojitým distribuovaným vektorovým reprezentaciám pre kusy textov a je súčasťou knižnice Gensim pod názvom Doc2Vec. Hlavnou myšlienkou je vytvoriť algoritmus, ktorý prevedie texty rôznej dĺžky do vektorovej reprezentácie s pevnou, rovnakou dĺžkou. Základnou motiváciou bolo splniť požiadavku mnohých metód strojového učenia, ktoré vektory tohto typu na vstupe očakávajú.

V modeli je vektorová reprezentácia trénovaná na princípe predikcie slov v danom kontexte. Táto technika je inšpirovaná súčasnými prácami v učení vektorových reprezentácií slov za použitia neuronových sietí. Kde každé slovo je reprezentované ako vektor, ktorý je zreťazený alebo zpriemerovaný s inými vektormi slov v danom kontexte a výsledný vektor je použitý na predikciu ďalších slov v tomto kontexte.

Paragraf vektor implementuje dva rôzne algoritmy. Jedným je "Distributed Memory Model of Paragraph Vectors"(model zdieľanej pamäte paragrafových vektorov) skrátene PV-DM a tým druhým je "Distributed Bag of Words version of Paragraph Vector"(model zdieľaného Bag of words) skrátene PV-DBOW. Základný rozdiel medzi týmito algoritmami je v tom, že model

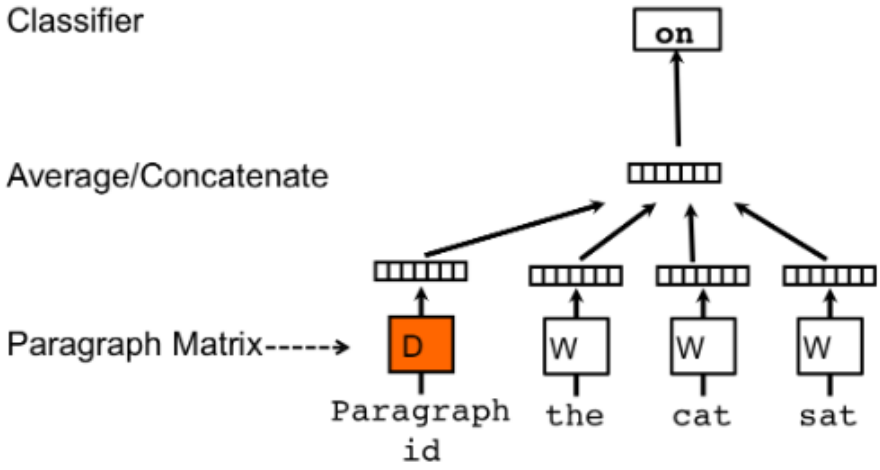
PV-DM berie do uvahy poradie slov vo vetách. Algoritmus funguje podobne ako učenie vektorovej reprezentácie slov teda Word2Vec.



Obr. 2.1: zdroj: [3] Schéma frameworku na učenie vektorovej reprezentácie slov

V tomto modeli vektory slov v kontexte prispievajú do predikcie nasledujúcich slov, naviac však ešte k predikcii prispievajú svojou váhou vektory daných ostavcov. Predikčná úloha sa zvyčajne vykonáva prostredníctvom viactriedneho klasifikátora ako je softmax. V praxi sa uprednostňuje hierarchický softmax kôli časovej náročnosti výpočtu. V paragraf vektor modeli je štruktúra hierarchickej softmax matice v tvare Huffmanovho binárneho stromu, kde krátke kódy sú priradené častým slovám. Fungovanie hierarchického softmaxu viz. [17].

Vo frameworku je každý paragraf namapovaný na jedinečný vektor reprezentovaný stĺpcom v matici D a každé slovo je namapované na vektor reprezentovaný stĺpcom v matici W. Paragrafový vektor a vektory slov sú zreťazené alebo spriemerované na predikciu nasledujúceho slova v kontexte. Paragrafový vektor sa pritom správa ako pamäť, v ktorej je uložené chýbajúce slovo do daného kontextu resp téma textu. Kontexty sú fixnej dĺžky a sú vzorkované z posuvného okna nad odsekom. Paragrafový - odsekový vektor je zdieľaný vo všetkých kontextoch generovaných z toho istého odseku, ale nie vo viacerých odsekoch. Matica slov je však zdieľaná naprieč odsekmi. Pri každom kroku stochastickej aproximácie vyberieme vzorku pevnej dĺžky z náhodného odseku a vypočítame gradient chýb zo siete na obrázku 2.2. Model je klasická neurónová sieť s učením pomocou backpropagation. Váhy v modeli sú ďalej upravované štandardným spôsobom. V predikčnom čase je potrebné vykonať odvoďovací

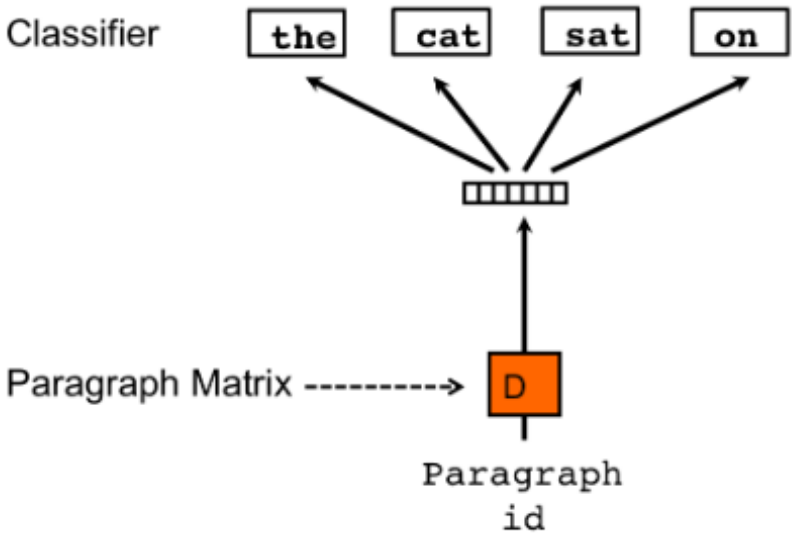


Obr. 2.2: zdroj:[3] Schéma frameworku paragraf vektor s PV-DM algoritmom

krok na vyrátanie paragrafoveho vektoru pre nový paragraf. Toto je znovu uskutočnené klesaním podľa gradientu. V tomto kroku sú ostatné parametre modelu, vektory slov W a softmax váhy zafixované. Po natrénovaní môžu byť paragrafove vektory použité ako funkcie. Tieto funkcie môžeme následne použiť priamo ako vstupy pre metódy logistickej regresie napríklad K-means.

PV-DBOW Paragraf Vektor bez poradia slov. Tento spôsob ignoruje kontext slov na vstupe a necháva model predikovať slová náhodne zvolené na výstupe. V praxi to znamená, že v každej iterácii klesania podľa gradientu, zvolíme textové okno a potom z neho vyberieme náhodné slovo, vytvoríme tak klasifikačnú úlohu ktorá je daná paragrafovým vektorom. V tejto technike sú kladené oveľa menšie nároky na pamäť, kde stačí mať uložené softmaxové váhy. Avšak na základe straty poradia slov je výkonosť tohto algoritmu nižšia než PV-DM. Z výsledkov práce [3] vyplýva, že Paragraf Vektor je konštatne lepším modelom než napríklad Bag of Words a podobné do teraz používane modely. V práci bol dokázaný predpoklad, že algoritmus PV-DM funguje v úlohách klasifikácie alebo analýzy sentimentu omnoho presnejšie než PV-DBOW.
Model Doc2Vec v knižnici Gensim obsahuje veľké množstvo nastaviteľných parametrov. V nasledujúcej časti čitateľovi predstavím tie najdôležitejšie ostatné je možné dohľadať v dokumentácii Doc2Vec.

- documents - sada iterovateľných dokumentov
- vector_size - vyjadruje veľkosť vektora generovaného modelom pre každý text



Obr. 2.3: zdroj:[3] Schéma frameworku paragraf vektor s PV-DBOW algoritmom

- window - vyjadruje maximálnu vzdialenosť medzi aktuálnym a predikovaným slovom vo vete
- min_count - ignoruje všetky slová ktorých celkový počet je menší než zadaná hodnota
- epochs - počet iterácií ktorými prejde trénovaný model
- negative - Ak je hodnota väčšia než 0, na trénovanie bude použitý negative sampling, s nastavením na nulovú hodnotu nebude použitý negative sampling
- hs - S hodnotou 1 bude použitý pre trénovanie hierarchycký softmax, s hodnotou 0 a parametrom negative nenulovým bude použitý negative sampling

Samotný model sa dá použiť na vyjadrenie podobnosti jednotlivých slov napríklad:

```
model.wv.similarity('žena', 'dívka')
```

```
0.7820387585106097
```

Podobne dokáže nájsť pomocou metódy doesnt_match slovo, ktoré nepatrí k danému kontextu:

```
model.wv.doesnt_match("oběd snídaně hruška svačina".split())
```

'hruška'

Model funguje dokonca aj na zaujímavé aritmetické operácie ako napríklad:

```
model.most_similar(positive=['žena', 'král'], negative=['muž'])
```

```
[('královna', 0.7777215242385864),
 ('panovnice', 0.7366218566894531),
 ('císařovna', 0.7228655815124512),
 ('choť', 0.7054554224014282),
 ('regentka', 0.6920363903045654),
 ('princezna', 0.68143630027771),
 ('velkovévodkyně', 0.6806958913803101),
 ('alžběta', 0.6774953603744507),
 ('beatrix', 0.6745885014533997),
 ('arsinoé', 0.6711801886558533)]
```

Kapitola 3

Realizácia

3.1 Predstavenie riešenia

Praktická časť práce predstavuje problém, ktorý v sebe zahŕňa problematiku z oblasti web miningu a text miningu. Celá aplikácia bola naimplementovaná v jazyku Python. Základným problémom na ktorý som narazil v mojej práci bolo zbieranie článkov z českých spravodajských serverov. Prvou vecou v tomto procese bolo rozhodnutie sa, z ktorých serverov budem dané články zbierať.

Keďže Novinky.cz a iDNES.cz v návštevnosti medzi českými servermi jasne dominujú [18] , zvolil som pre prácu práve tie. V ďalšej časti práce budem používať len názvy iDNES a Novinky. Bolo to z dôvodu, že tieto dva servery obsahujú dostatočne veľké množstvo na nasledujúcu analýzu.

Ďalším krokom bolo zozbieranie dostatočného objemu linkov k daným článkom, ktoré budú neskôr pravidelne sťahované zo serverov. Pre tento účel som sa rozhodol naimplementovať crawlery v jazyku Python, ktoré prejdú dané servery a pozbierajú z nich relevantné odkazy. "Web crawler je počítačový program, ktorý prechádza web metodickým, automatizovaným spôsobom"[4, str. 65]. Nasledujúcim krokom bolo odstránenie neužitočných informácií z daných stránok a prevod na čistý text. Pre tento účel som naimplementoval vlastný parser, ktorý prevádza jednotlivé html stránky do čistého textu. Na tento účel som využil Python knižnicu Beautifulsoup[19]. Analýza a porovnanie článkov bola realizovaná za pomoci nástroja Gensim konkrétne modelom Doc2Vec.

3.2 Crawler

Crawlery prechádzajú štruktúru každého zo serverov rozličným spôsobom. Pre server Novinky sa zadá do parametrov maximálny počet požadovaných linkov a dátum, do ktorého maximálne sa má ísť do archívu. V prvom kroku sa z hlavnej stránky vyberú linky všetkých rubrík, ktoré v sebe obsahujú archív článkov s výnimkou rubriky horoskop. Následne sa z daných archívov zozbierajú linky

článkov. Ukončovacou podmienkou je buď dosiahnutý zadaný dátum alebo prekročený maximálny počet linkov.

Server iDNES neobsahuje archív článkov a okrem toho celková štruktúra serveru je oveľa menej prehľadná ako u Noviniek. Ukončovacie podmienky sú rovnaké ako v predchádzajúcom prípade, kvôli štruktúre serveru som sa rozhodol pre priechod stránkou zanorovaním. Proces začína na hlavnej stránke serveru v rubrike Zprávy-domácí odkiaľ sa vyzbierajú všetky linky, ktoré stránka obsahuje. Tieto následne prejdú kontrolou obsahu, štruktúry url a dátumu vydania. Takto skontrolované linky sa následne uložia do fronty. Táto fronta slúži na kontrolu stránok na ktorých sa už crawler nachádzal. Z nej sa pri nesledujúcej iterácii vyberie url ešte nenavštívenej stránky, z ktorej sa znovu zozbierajú linky. Tento proces pokračuje až do momentu, kým nemá crawler dostatočný počet vyzbieraných url alebo fronta ešte nenavštívených stránok nezostane prázdna.

Po ukončení procesu oboch crawlerov sa vyzbierané url uložia do csv súborov, ktoré slúžia ako pamäť pri ďalších behoch programu. Tento proces získavania linkov prebieha len pri prvom behu programu, respekntíve v momente keď sa v súborovej štruktúre nenachádza csv súbor so zoznamom článkov. Následne prebieha každú polhodinu kontrola rss kanálov, kde by mali byť aktuálne vydané články pre každý deň. Url na tieto články sa znova uložia do pamäte a vykoná sa kontrola obsahu csv súboru držiaceho celú sadu linkov. V prípade, že sa niektorý z nájdených linkov ešte v csv súbore nenachádza, prebehne zápis tohto linku do súboru.

3.3 Parser

Cieľom parsovania bolo vyextrahovať dôležité informácie pre následnú analýzu. Zbierané informácie sa znovu, v závislosti na servere pre ktorý je parser implementovaný, mierne líšia. Toto je spôsobené rozlišnosťou štruktúry stránok pre dané servery. Novinky totiž neobsahujú vo svojich článkoch kľúčové slová, respektíve témy ktorých sa daný článok týka. Na oboch serveroch je ale zameraný parser na získanie identifikačného čísla článku, titulku, perexu, dátumu publikácie, dátumu aktualizácie a textu.

Identifikačné číslo článku sa u oboch serverov nachádza v url daného článku. Pre jeho vyextrahovanie som teda v oboch prípadoch použil regulárne výrazy. Pomocou knižnice BeautifulSoup a jeho vstavaného lxml parseru na odstránenie html kódu bolo možné mnoho informácií získať pomocou jednoduchého odkazovania sa na tagy. Poprípade je BeautifulSoup schopný orientovať sa aj pomocou metadát nachádzajúcich sa v html kóde. Najväčší problém nastal pri získavaní samotného textu. V prípade oboch serverov sa v textovej časti html kódu nachádzali často obrázky, videá, tabuľky, fotografie a v prípade serveru iDNES aj ankety. Preto boli tieto elementy pred extrakciou textu z štruktúry stránky odstránené. Následne boli vyextrahované informácie

prevedené na jednotný formát.

3.4 Formát a ukladanie

Takto sformátované informácie o článkoch sú ukladané do štruktúry pandas, ktorá sa následne zápisuje do csv súboru. Každý článok daných serverov má vlastný csv súbor. Súbory sú pomenované podľa identifikačného čísla článku a nesú v sebe všetky verzie článku s daným identifikačným číslom.

Na to, aby sa do súboru zapísala nová verzia článku, sa musí táto verzia článku od tej predošlej odlišovať minimálne v jednej zo štyroch častí. A to buď v názve, titulku, témach, alebo perexe. Kontrola v závislosti na témach článku, ktorá bola možná len na servere iDNES, sa nakoniec ukázala ako zbytočná, pretože témy sa u niektorých článkov obmieňajú veľmi často. Toto spôsobovalo neustále zapisovanie nových verzií do csv súborov daného článku. Fakt, že sa témy menia tak často, môže značiť, že témy sú generované automaticky. Pred samotným zápisom sa tieto informácie kontrolujú s predošlou verziou článku. Jeden záznam o verzii článku v sebe ešte nesie dátum a čas stiahnutia danej verzie a url daného článku.

3.5 Analýza

Dáta som zbieral počas troch týždňov zo servera iDNES a počas jedného týždňa zo serveru Novinky. Výsledna množina dát, ktorá bola nakoniec analyzovaná je zobrazená v tabuľke 3.1.

	iDNES	Novinky
Celkový počet	4327	3747
Správy	823	1815

Tabuľka 3.1: Počet článkov v datasete

Zmeny v rámci článkov som zaznamenal len na serveri iDNES. Na serveri Novinky boli zaznamenané celkom len 4 zmeny v článkoch. Z tohto dôvodu som túto množinu pri porovnaní verzií úplne vynechal a používal len dáta zo serveru iDNES.

Možným dôvodom, prečo sa mi zmeny v práci nepodarilo zachytiť môže byť kratší čas, ktorý boli články na serveri kontrolované a zbierané. Je tiež však možné, že na tomto serveri k tak častým zmenám ako na iDNES nedochádza.

Na serveri iDNES k významným zmenám dochádzalo najmä v prípade, že šlo o aktuálnu udalosť pri ktorej sa postupňe informácie dopĺňali, poprípade o neaký seriál článkov na pokračovanie.

Obzvlášť veľkými zmenami tiež prechádzali články vydávané v rubrike šport. Toto bolo spôsobené znovu aktualizovaním výsledkov a odohrávajúcich

sa udalostí.Veľké množstvo zmien nastávalo tiež v prípade opravovania gramatických chýb a slovných preklepov.

	Šport	Správy	Ostatné rubriky
Celkový počet	1315	823	2238
Z toho zmenených	219	178	358
Zmena titulku	102	77	38
Zmena perexu	93	74	69
Zmena textu	203	169	324
Zmena titulku & perexu & textu	71	56	38

Tabuľka 3.2: Zmeny článkov server iDNES

Delenie správ som robil na základe url daných článkov. Množina správ v sekcii Ostatné rubriky obsahuje články týkajúce sa recenzií filmov, techniky a podobne. V sekcii správy sa nachádzajú rubriky domácí a zahraniční správy. Tieto články boli ďalej hlavnými objektmi analýz. Články v sekcii šport v sebe obsahujú všetky články z rubriky hokej, futbal a sport, kde sú na serveri združené všetky ostatné športy.

Pre toto delenie kategórií som sa rozhodol z dôvodu, že články z oblasti športu často obsahovali číselné údaje, vyjadrujúce výsledky odohrávajúcich sa zápasov. Tie spôsobovali skreslenosť meraných podobností medzi ich jednotlivými časťami. Okrem iného, sa v nich často vyskytovali názvy klubov, štadiónov a podobne, ktoré trénovaný model vo svojom slovníku neobsahoval. Kvôli týmto dôvodom som sa nimi viacej pri meraní podobností jednotlivých častí článkov nezaoberal a meral som na nich len podobnosti verzií.

Články z oblasti domácich a zahraničných správ som oddelil od ostatných z dôvodu, že obsahujú vecné informácie z domáceho a zahraničného diania, ktoré čitateľov zaujímajú najviac. A v prípade ich zmiešania s ostatnými kategóriami, by ostatné články spôsobovali skreslenosť výsledkov a nežiadaný šum. Tabuľka 3.2 zobrazuje počet článkov v jednotlivých kategóriách a množstvo zmien, ktoré v daných sekciách nastali.

3.6 Podobnosť

Po získaní dostatočného množstva dát pre nasledujúce meranie podobností bolo potrebné natrénovať daný model na dostatočne veľkom datasete. Obecne zaužívanými trénovacími korpusmi v oblasti spracovania prirodzeného jazyka sa stali korpusy Wikipédie. Wikipédia obsahuje dostatočné množstvo textov v prirodzenom jazyku, ktoré majú okrem iného ešte tú výhodu, že obsahujú informácie o celom spektre tém.

Model na analyzovanie podobností bol natrénovaný na tomto datasete českej wikipédie. Datasety wikipédie sú dostupné online pre širokú škálu jazykov. V rámci Gensim Google skupiny je zverejňované obrovské množstvo

tutoriálov, podľa ktorých som pri trénovaní modelu v práci postupoval. Pri práci som experimentoval s rôznymi nastaveniami parametrov modelu. Všetky vychádzali z nastavení odporúčaných tutoriálom, ktorý sa nachádza priamo v dokumentácii Doc2Vec. Na základe týchto experimentov som nakoniec, pre prácu zvolil algoritmus PV-DM, ktorý bol natrénovaný s nasledujúcimi parametrami:

- vector_size = 100
- window_size = 5
- min_count = 3
- negative = 0
- hs = 1
- train_epoch = 50

V tejto časti sa pokúsim predviesť ako model funguje na meranie podobností medzi časťami článkov. V prvej ukážke predvediem celý proces od spracovania plaintextu cez jeho prevedenie do vektoru až po samotný výpočet podobnosti.

Všetky texty prešli metódou simple_preproces, ktorá je vstavaná v knižnici Gensim. Táto metóda prevedie všetky slová na malé písmená, odstráni interpunkciu z textu a rozdelí text na pole slov. V prípade, že časť textu, najčastejšie titulok, bol kratší než 5 slov, som porovnanie na tomto článku nevykonal. To z dôvodu veľkosti kontextového okna, ktorá bola nastavená na hodnotu 5.

Následne sa takto spracované texty pomocou metódy infer_vector prevedú do vektorovej reprezentácie určenej daným modelom. Cosínusovu podobnosť výsledných vektorov som porovnával vstavanou funkciou spatial.distance.cosine z knižnice SciPy[20]. Ktorá počíta kosínusovu vzdialenosť podľa vzorca:

$$\text{distance}(u, v) = 1 - \frac{u \cdot v}{\|u\|_2 \|v\|_2}$$

Táto funkcia vracia hodnoty z intervalu $< 0, 2 >$. Hodnotu 0 pre totožné vektory, hodnotu 1 pre vektory na seba kolmé a 2 pre vektory opačné.
Podobnosť týchto vektorov som teda počítal ako:

$$\text{similarity}(u, v) = 1 - \mathit{distance}(u, v)$$

V oboch vzťahoch vyššie predstavujú u a v porovnávané vektory. Priebeh celého procesu meranie podobnosti je ilustrovaný na nasledujúcom príklade.

Celý titulok:
Üdělali jste chybu, reaguje Moskva na společný postup zemí vůči Rusku"

Titulok po spracovaní vstavanej metódy simple_preproces:
['udělali', 'jste', 'chybu', 'reaguje', 'moskva', 'na', 'společný', 'postup', 'zemí', 'vůči', 'rusku']

Perex:
Celkem šestnáct zemí Evropské unie, Spojené státy, Kanada a Ukrajina, Norsko a Albánie oznámily, že vyhostí ruské diplomaty v reakci na otravu agenta Sergeje Skripala nervovým plynem. K tomuto kroku se v pondělí uchýlila i Česká republika. Mluvčí Kremlu to označil za chybu a ujistil, že Moskva zareaguje recipročně.

Perex po spracovaní:
['celkem', 'šestnáct', 'zemí', 'evropské', 'unie', 'spojené', 'státy', 'kanada', 'ukrajina', 'norsko', 'albánie', 'oznámily', 'že', 'vyhostí', 'ruské', 'diplomaty', 'reakci', 'na', 'otravu', 'agenta', 'sergeje', 'skripala', 'nervovým', 'plynem', 'tomuto', 'kroku', 'se', 'pondělí', 'uchýlila', 'česká', 'republika', 'mluvčí', 'kremlu', 'to', 'označil', 'za', 'chybu', 'ujistil', 'že', 'moskva', 'zareaguje', 'recipročně']

V ďalšom kroku sa obe polia obsahujúce predspracovaný text prevedú do vektorovej reprezentácie určenej daným modelom. Keďže výsledné vektory majú dimenziu o veľkosti 100 prvkov,uvediem pre ilustráciu aspoň časť jedného z nich.

[-0.17723235, 0.37801775, -0.44836164, -0.74677604, -0.10892361, 0.8912548 , -0.09685361, 0.03941153, -0.35598892, 0.15092833]

Výsledná cosínusova podobnosť tohto titulku a perexu je 0.31873.

Vzhľadom na fakt, že dané texty k sebe naozaj patria, jedná sa o perex a titulok článku zo serveru iDNES, sa môže zdať že výsledná podobnosť je relatívne malá. Dôvodov tejto nízkej podobnosti môže byť mnoho. Jedným z nich je fakt, že slov ktoré sa vyskytujú v oboch textoch je relatívne málo. V tomto prípade je to len rusko, reaguje, moskva a chyba. Okrem toho sa slovo reaguje objavuje v texte perexu v budúcom čase naproti tomu v titulku je tento jav vyjadrený v čase prítomnom čo vyvoláva pocit vzájomného si protirečenia.

Ako uvidíme ďalej v práci, slová, ktoré sú obsiahnuté v oboch porovnávaných dokumentoch tvoria dôležitý faktor v meraní podobností medzi textmi. Často sú to práve citácie textu vyskytujúce sa v titulku, alebo kľúčové slová článku.

Medzi ďalšie ovplyvňujúce faktory patrí ďalej náhodné inicializovanie vektorov pri procese infer_vector. Faktor, ktorý však výsledky podobností textov

ovlyvňuje najviac sú ich dĺžky. V práci poukážem ďalej, na priemerné hodnoty podobností jednotlivých častí článkov, teda to ako sú si napríklad text a perex vzájomne podobné.

Ukáže sa, že tieto priemerné hodnoty sa líšia najmä v závislosti na dĺžkach vzájomne porovnávaných textov. Tento jav je logickým dôsledkom faktu, že viac informácií zaberie viac priestoru. Je nutné spomenúť, že ďalším ovplyvňujúcim faktorom môže byť to, že model vo svojom slovníku neobsahuje slová, ktoré sú obsiahnuté v samotnom texte.

3.7 Porovnanie podobnosti rôznych častí článkov

V tejto časti Vás zoznámim s výsledkami mojich porovnaní daným modelom na zozbieraných dátach. Jednotlivé časti článkov som vždy porovnával z poslednej uloženej verzie. Tieto som ďalej spracoval vyššie demonštrovaným procesom a porovnal.

V tabuľke 3.3 sú tri najvyššie podobnosti medzi textom a perexom, vybrané zo všetkých zozbieraných článkov.

Podobnosť	Titulok	Perex	Server
0.741521	V Hronově se na Noc kostelů otevře pozdně renesanční zvonice	V pátek 25. května se v Hronově v rámci Noci kostelů veřejnosti otevře pozdně renesanční zvonice	Novinky
0.717111	ČEZ prověří možnost vstupu Bulharska do prodeje svých aktiv	Energetická skupina ČEZ prověří možnost vstupu bulharského státu do prodeje svých bulharských aktiv společnosti Inercom. Po skončení několikadenních jednání ČEZ a Inercomu v Praze to sdělila mluvčí ČEZ Alice Horáková. Požadavek podle ní přišel od bulharské firmy	iDNES
0.714638	V Chebu se uskutečnila tradiční Velká cena města v judu	V sobotu 14. dubna se v tělocvičně 5. ZŠ uskutečnila již tradiční Velká cena Chebu v judu.	Novinky

Tabuľka 3.3: Tri najväčšie podobnosti v porovnaní titulok x perex

V prvom uvedenom porovnaní je podobnosť medzi perexom a titulkom zjavná, perex sa od titulku daného článku líši len pridaním informácií, že kostol bude otvorený verejnosti a že udalosť sa uskutoční 25. apríla. Výsledná podobnosť týchto textov teda presne zodpovedá skutočnosti.

Tretie porovnanie zobrazuje presne ten istý typ titulku a perexu, kde perex je prakticky titulok rozvitý o pár ďalších informácií.

Druhý príklad sa od predošlých výrazne líši v dĺžke perexu. Perex okrem informácie, že *"ČEZ preverí možnosť vstupu Bulharska do predaja svojich aktív"* dodáva akej spoločnosti sa to bude týkať, kým a za akých okolností bola informácia zverejnená. Aj napriek týmto javom však model vystihol obsah uvedených textov presne a vyhodnotil, že sú si veľmi podobné.

Pre porovnanie v tabuľke 3.4 uvádzam aj tri najnižšie podobnosti v porovnaní titulku a perexu článkov. V tabuľke 3.4 sú vidieť texty, pri ktorých model správne odhadol rozdielnosť ich obsahov.

V prvom porovnaní titulok informuje čitateľa o tom, že podvodní agenti zarábajú na rodičoch malých hokejistov. Následne v perexe je čitateľ informovaný o tom, že náklady na výchovu detí, ktoré to myslia s hokejom vážne, môžu siahať až do výšky 2 miliónov korún. Teda píše o úplne inej informácií, než aká bola čitateľovi predostrená v titulku. Z tohto dôvodu by som titulok hodnotil ako irelevantný vzhľadom k perexu a hodnotenie modelu je teda správne.

Titulok v druhom porovnaní informuje o tom, že konflikt medzi Slovenským prezidentom Andrejom Kiskom a vtedajším Slovenským premiérom Róbertom Ficom sa stupňuje. Naproti tomu v perexe je čitateľ informovaný o tom, že prezident Slovenskej republiky oznámil, že zahájí rokovania s politickými stranami a chce riešiť politickú krízu. V tomto prípade ide podľa môjho názoru o zavádzajúci titulok a hodnotenie je správne, v perexe nie je vyjadrená žiadna ofenzívna reakcia smerom k premierovi.

Tretie porovnanie zobrazuje expresívny titulok, ktorý v čitateľovi evokuje, že v texte sa dozvie čo sa udialo na predvolebnom mítingu Viktora Orbána. Avšak obsahuje mnoho výrazov, ktoré aj pre čitateľa nepoznajúceho kontext můžu byť nejasné. Perex potom rozvíja titulok tak, aby si čitateľ prečítal aj celý text. Je tu použité veľké množstvo básnických výrazov a prenesených pomenovaní, takže je aj pre človeka ťažké rozhodnúť nakoľko si texty odpovedajú.

V tabuľke 3.5 sú zobrazené výsledky porovnaní medzi jednotlivými časťami článkov. Porovnania boli realizované na počte 3747 článkov a teda na celej množine článkov zozbieranej zo serveru Novinky.

Nasleduje tabuľka 3.6 ktorá znázorňuje výsledky tých istých meraní na množine článkov zo serveru iDNES. Vzhľadom na fakt, že články z rubriky šport sa svojim obsahom od bežných článkov značne líšia, neboli tieto články zahrnuté do porovnávanej množiny. Výslednú množinu porovnávaných textov teda tvorilo 3010 článkov.

Podobnosť	Titulok	Perex	Server
-0.238558	Podvodní agenti vydělávají na rodičích malých hokejistů	Ambiciózní rodič, který chce své dítě jednou vidět jako hráče NHL anebo alespoň v české extralize, je schopen obětovat nemalé částky na jeho výchovu. Od prvních hokejových krůčků až po profesionální smlouvy či start v lize se náklady podle Právem oslovených odborníků šplhají až ke dvěma miliónům korun.	Novinky
-0.215149	Konflikt mezi Kiskou a Ficem se stupňuje	Slovenský prezident Andrej Kiska považuje v současnosti za nejdůležitější vrátit lidem důvěru ve stát, jeho představitele a spravedlnost. „Od středy zahajuji jednání s politickými stranami a jsem připraven vyvést naši zemi z této krize," řekl Kiska v úterý v průběhu pracovní cesty ve městě Tvrdošín	Novinky
-0.214666	Tady je Viktor doma. Na Orbánově mítinku došlo i na Trianon a slzy	Székesfehérvár (od zpravodaje iDNES.cz) - Spíše demonstraci maďarských vlastenců než stranickou akci připomínal poslední předvolební mítink Viktora Orbána. Do Székesfehérváru na něj dorazily tisíce lidí. „Když budou další čtyři roky stejné, poděkujeme za to Bohu," myslí si místní voliči Fideszu.	iDNES

Tabuľka 3.4: Tri najmenšie podobnosti v porovnaní titulok perex

	Podobnosť perexov a titulkov	Podobnosť textov a titulkov	Podobnosť textov a perexov
mean	0.229864	0.218072	0.331661
std	0.139120	0.115160	0.115138
min	-0.238558	-0.148403	-0.095372
25%	0.134868	0.139855	0.255102
50%	0.229882	0.222073	0.337123
75%	0.325751	0.300525	0.411263
max	0.741521	0.609446	0.695885

Tabuľka 3.5: Porovnanie článkov zo serveru Novinky

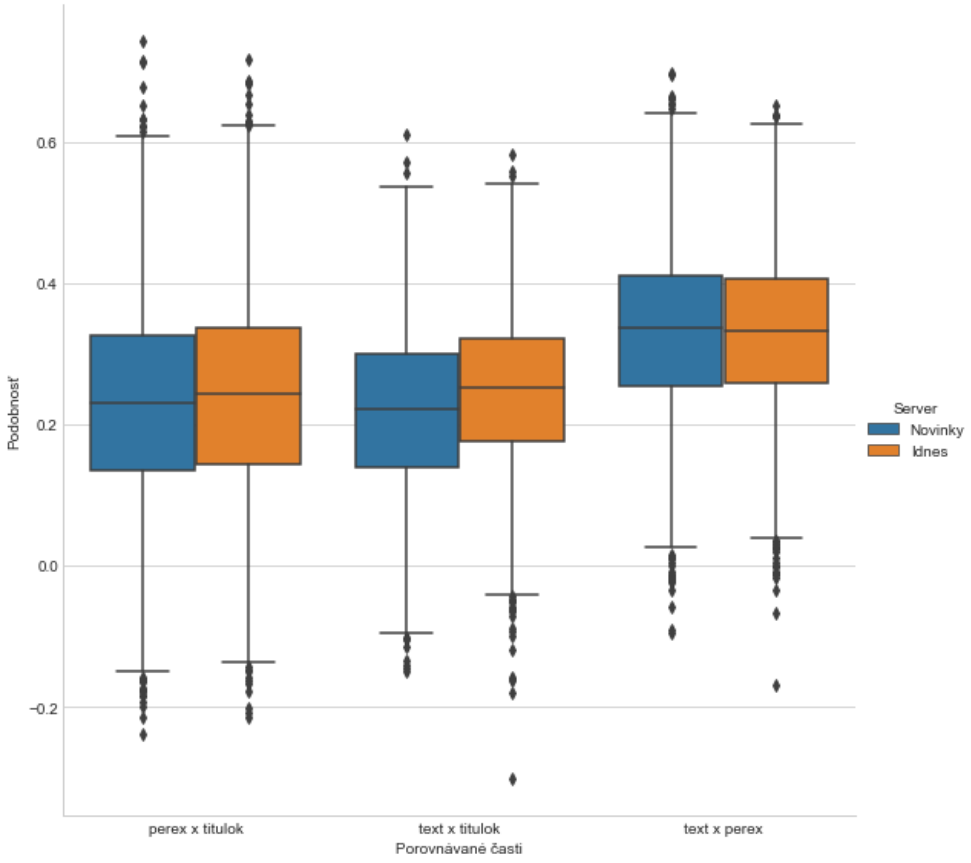
	Podobnosť perexov a titulkov	Podobnosť textov a titulkov	Podobnosť textov a perexov
mean	0.241338	0.247682	0.329487
std	0.139728	0.109136	0.108936
min	-0.214666	-0.301931	-0.169488
25%	0.144584	0.176175	0.258406
50%	0.242413	0.251332	0.332698
75%	0.336087	0.322283	0.405815
max	0.717111	0.582906	0.650257

Tabuľka 3.6: Porovnanie článkov zo serveru iDNES

	Podobnosť perexov a titulkov	Podobnosť textov a titulkov	Podobnosť textov a perexov
mean	0.043545	0.061463	0.087871
std	0.110920	0.106421	0.107680
min	-0.341822	-0.318941	-0.272519
25%	-0.030882	-0.010307	0.015676
50%	0.044395	0.059021	0.085642
75%	0.118047	0.135176	0.158151
max	0.415467	0.378066	0.554405

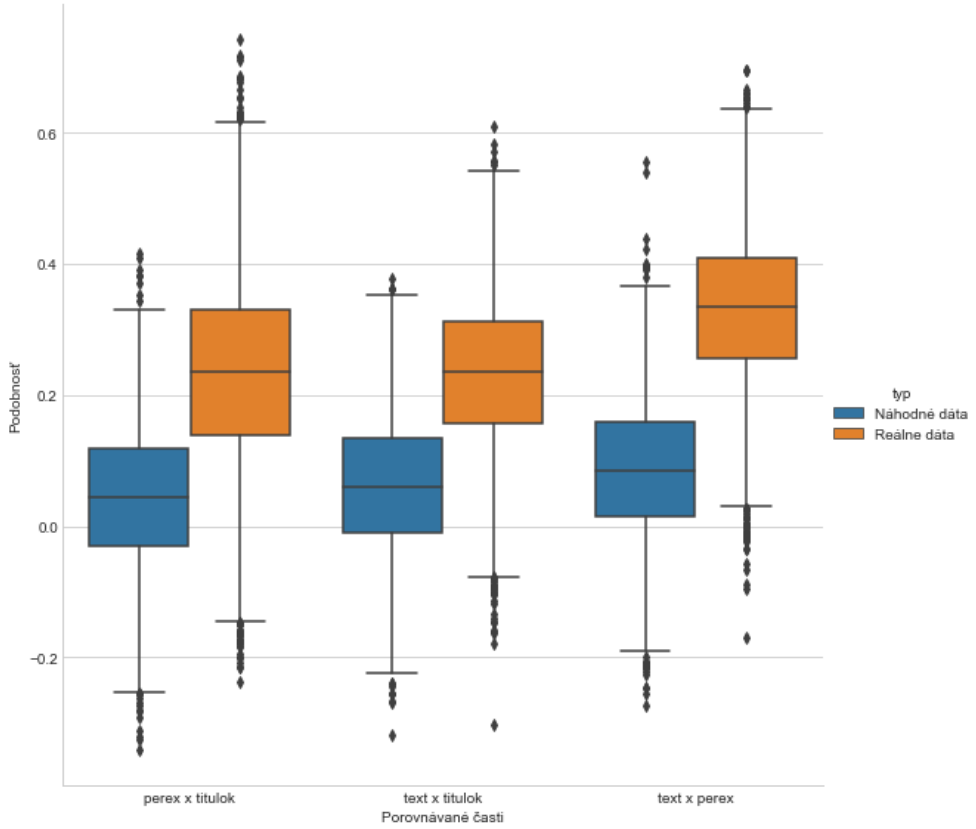
Tabuľka 3.7: Náhodné porovnanie rôznych častí článkov

Môžeme si všimnúť, že výsledné hodnoty porovnaní sa od seba líšia, v závislosti od toho, o porovnanie akých častí textov boli ide. Tento fakt je očividne spôsobený ich rozdielnymi dĺžkami. Kvôli lepšej predstave o tom, kde sa nachádzajú priemerné hodnoty podobností pre časti článkov, ktoré si vzájomne nezodpovedajú, som urobil test merania podobností s náhodnými dátami. Výsledky tohto porovnania sú znázornené v tabuľke 3.7. Toto porovnanie som urobil na 3000 náhodne vybraných častiach článkov. Množinu z ktorej boli tieto perexy, titulky a texty vyberané tvorili všetky články z oboch serverov okrem sekcie šport, ktorá nebola súčasťou ani predošlých porovnaní.



Obr. 3.1: Rozdelenie podobnosti rôznych dvojíc sledovaných častí pre oba servery

Pre lepšiu predstavu, sú v grafe 3.1 znázornené výsledky meraní podobností na týchto dvoch serveroch. Nasleduje graf 3.2 znázorňujúci výsledky meraní podobností s reálnymi dátami oproti tým s náhodnými dátami. V grafe skupina reálnych dát predstavuje spojenie výsledkov meraní podobností na serveri iDNES viz. 3.6 a Novinky viz 3.5. Skupina náhodných dát zobrazuje v grafe výsledky porovnaní z tabuľky 3.7.



Obr. 3.2: Výsledky porovnaní rôznych dvojíc sledovaných častí z rôznych, náhodne vybraných článkov, oproti výsledkom porovnaní rôznych dvojíc sledovaných častí z článkov sebe si odpovedajúcich

Výsledky porovnaní sa v závislosti od serverov líšia len veľmi málo. V prípade priemernej hodnoty podobnosti perexov a titulkov je to napríklad jedna stotina. Väčšie rozdiely sú viditeľné len v prípade podobností textov a titulkov kde priemerná hodnota v meraní na servere iDNES je o 3 stotiny väčšia. Ďalším zaujímavým rozdielom je maximálna hodnota v porovnaní perexov a textov, ktorá je znovu väčšia na servere iDNES o 5 stotín, tento výsledok je jasne vidieť v grafe 3.1. Vzhľadom na náhodnú inicializáciu vektorov, a odchylky meraní modelu, to však môžeme pokladať za zanedbateľné hodnoty.

3.8 Porovnanie verzií článkov

V nasledujúcej časti práce uvediem výsledky porovnaní zmien medzi verziami článkov z hľadiska ich sémantického obsahu. Pred porovnaním bolo nutné

dataset prečistiť od tých článkov, ktoré obsahovali len videá. Poprípade články, ktorých texty, perexy alebo titulky boli kratšie než 5 slov.

Pre demonštráciu ako veľmi zmeny v textoch ovplyvňujú ich vzájomnú podobnosť, predvediem ukážku zmien podobností v čase v závislosti na zmenách upravovaných textov. Kvôli lepšej predstave toho, čo výsledky daných porovnaní predstavujú som znovu vytvoril test na náhodných dátach. Výsledky tohto testu sú znázornené v tabuľke 3.8. V tabuľkách 3.9, 3.10 a 3.11 sú výsledky meraní podobností v závislosti na verzii daného článku. Do meraní som zahrnul len zmeny v rámci druhej a tretej verzii článku, hoci niektoré články v datasete mali aj viac ako 5 verzií. Podobnosti zmenených častí článkov boli vždy merané voči predošlej verzii daného článku. To znamená, že hodnota podobnosti textu druhej verzie, vyjadruje podobnosť voči textu prvej verzie.

Zmeny titulkov:

Prvý titulok článku
Vyhoštění tří Rusů z Česka okomentuje Babišův poradce Svoboda
Čas stiahnutia: 26/03/2018 18:01:37

Druhý titulok článku
Útok na agenta před volbou prezidenta nebyla náhoda, řekl exministr Svoboda
Čas stiahnutia: 27/03/2018 13:15:29
Podobnosť s titulkom z prvej verzie: 0.560227

Tretí titulok článku
Zeman neví, co Rusku říct, myslí si o pátrání po novičoku exministr Svoboda
Čas stiahnutia: 27/03/2018 13:55:28
Podobnosť s titulkom z prvej verzie: 0.248929

Z vyššie uvedenej ukážky si môžme všimnúť typickú zmenu titulku v rámci aktualizovanej udalosti. Tento typ zmien prebiehal často najmä v rubrikách šport a správy. Články zo všetkých kategórií sa menia hlavne v období krátko po ich vydaní keď sú ešte aktuálne. Vzhľadom na tento fakt a na to, že som zbieral v práci články z histórie častokrát aj tri mesiace do zadu, počet článkov, ktoré sa menia, môže byť oveľa vyšší než v tabuľke 3.2. Avšak články, u ktorých zmeny nastávali mali často viac než 2 verzie.

Priemerný počet verzii na článok(u zmenených článkov):

1. Správy - 2,92
2. Šport - 3,15

3. Ostatné kategórie - 2,35

S ohľadom na fakt, že v sekcii ostatné kategórie sa nachádzajú články odborného charakteru - technet, bývanie, bulvár, články pre ženy - sa dá predpokladať, že k zmenám dochádza vo veľkej miere len v prípade gramatických chýb a preklepov. To je spôsobené tým, že informácie, ktoré texty obsahujú nie je potrebné nijak priebežne aktualizovať. Tento jav je viditeľný v grafe 3.3 na priemernej hodnote podobnosti textov nových verzii. A tak isto aj v hodnotách podobností perexov viz 3.5. Naproti tomu priemerná zmena perexu v kategórii správ je značne väčšia. Tento jav je spôsobený faktom, že pri aktualizácii textu aktuálnej udalosti, je potrebné s ním prepísať aj perex. Menšia priemerná hodnota podobnosti s prvou verziou perexu je spôsobená tým, že ide o zmeny na kratšom texte.

	Podobnosť perexov	Podobnosť titulkov	Podobnosť textov
mean	0.068921	0.036887	0.141298
std	0.107373	0.116622	0.121010
min	-0.292708	-0.337614	-0.258820
25%	-0.004401	-0.041923	0.055621
50%	0.070638	0.036467	0.138162
75%	0.140603	0.113239	0.216402
max	0.948861	0.946080	0.597721

Tabuľka 3.8: Náhodné porovnanie rovnakých častí článkov

V grafoch 3.3,3.5,3.4 si môžeme všimnúť, že hodnoty zmien v kategórii šport sa výrazne líšia od ostatných dvoch kategórii. Je to dôsledok už vyššie spomenutých faktov a to, že kategória obsahuje viac článkov, ktoré sa musia upravovať. Je to spôsobené hlavne priebežným aktualizovaním výsledkov z dohratých zápasov. Prejavilo sa to najmä na výsledkoch podobností titulkov daných článkov kde priemerná hodnota podobnosti titulku.

	Podobnosť textov		Podobnosť perexov		Podobnosť titulkov	
Verzie	Druhá	Tretia	Druhá	Tretia	Druhá	Tretia
mean	0.841049	0.907541	0.481333	0.614601	0.733557	0.650320
std	0.197019	0.105580	0.326414	0.320671	0.264807	0.312069
min	0.141694	0.410830	0.022174	0.132695	0.024330	0.016674
25%	0.793544	0.870914	0.185796	0.329198	0.518672	0.356344
50%	0.928378	0.946866	0.428770	0.818183	0.879158	0.884415
75%	0.970217	0.974818	0.838707	0.882653	0.931058	0.912777
max	0.996183	0.997875	0.984411	0.950678	0.970303	0.940426

Tabuľka 3.9: Porovnanie zmien v článkoch kategória Šport

	Podobnosť textov		Podobnosť perexov		Podobnosť titulkov	
Verzie	Druhá	Tretia	Druhá	Tretia	Druhá	Tretia
mean	0.899681	0.917318	0.620557	0.586492	0.706299	0.727411
std	0.133093	0.102866	0.327094	0.245388	0.260522	0.242009
min	0.137695	0.497994	-0.035175	0.199916	0.021880	0.245077
25%	0.873684	0.910562	0.320390	0.435989	0.528400	0.520961
50%	0.952811	0.964406	0.754405	0.532314	0.859111	0.875656
75%	0.974170	0.978544	0.932452	0.779676	0.918636	0.920915
max	0.990690	0.992299	0.984046	0.976865	0.966861	0.953544

Tabuľka 3.10: Porovnanie zmien v článkoch kategória Správy

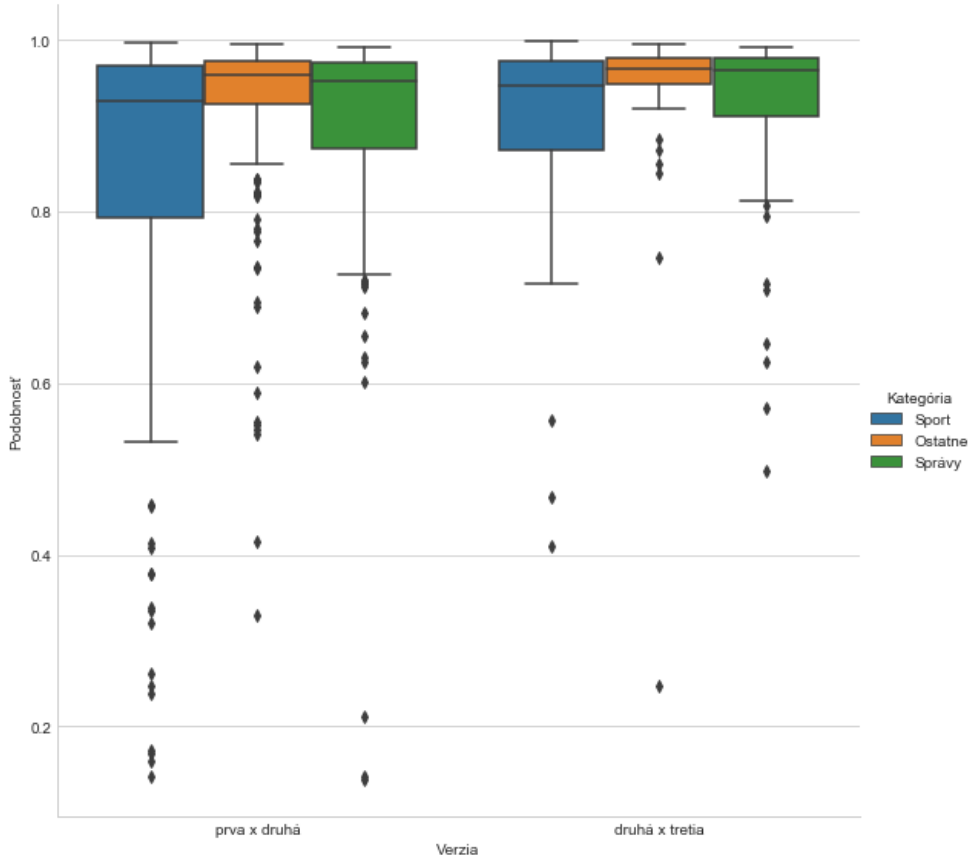
	Podobnosť textov		Podobnosť perexov		Podobnosť titulkov	
Verzie	Druhá	Tretia	Druhá	Tretia	Druhá	Tretia
mean	0.933923	0.947371	0.683889	0.699035	0.688110	0.648402
std	0.083536	0.091065	0.328073	0.217263	0.205078	0.188006
min	0.329971	0.246949	-0.099028	0.213818	0.120979	0.399396
25%	0.924764	0.947757	0.456791	0.574365	0.608567	0.526861
50%	0.959385	0.966079	0.855373	0.697847	0.736141	0.607661
75%	0.976080	0.978520	0.937593	0.861639	0.828655	0.799956
max	0.995929	0.995100	0.980054	0.940728	0.924634	0.878124

Tabuľka 3.11: Porovnanie zmien v článkoch kategória Ostatné

Ďalšími faktormi, ktoré v porovnaniach veľmi ovplyvňovali výsledky boli napríklad zostavy, ktoré nastúpili v danom zápase. Keďže pre model je to len sada neznámych slov, ktorým priradí náhodne váhy.

Ďalšou zaujímavou hodnotou sú zmeny perexov v kategórii správ. Obecne najmenšie zmeny boli zaznamenané v oblasti textov všetkých troch kategórií. tento jav je spôsobený tým, že pokiaľ sa v texte väčšej dĺžky zmení pár slov, na výslednej podobnosti s predchádzajúcou verziou sa táto zmena výrazne neprejaví.

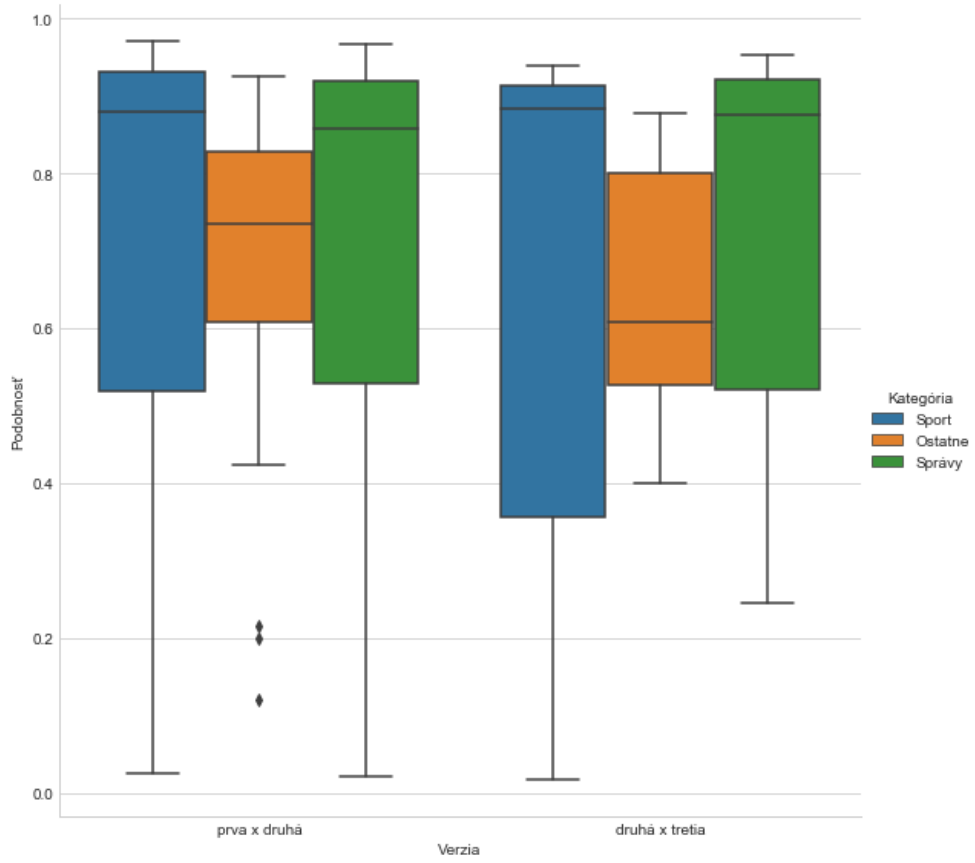
Zaujímavým javom je, že zatiaľ čo zmeny majú v sekcii šport, v závislosti na verzii stály klesajúci charakter, v sekcii ostatné sú v tretej verzii článku vždy väčšie než v tej druhej. Tento jav je možné sledovať aj v sekcii správy u zmien perexov. Možným vysvetlením môže byť fakt, že tretích verzií sa vyskytuje razantne menšie množstvo oproti druhým. Väčšinou tu už ide o samotný prepís obsahu textu než o opravovanie chýb a preto sú tieto väčšie zmeny vo výsledku viac viditeľné.



Obr. 3.3: Zmeny v textoch medzi verziami v závislosti na kategórii

3.9 Porovnanie rovnakých článkov z rôznych zdrojov

V tejto časti porovnám dva články týkajúce sa tej istej témy, ktoré pochádzajú každý z iného zdroja. Pre toto porovnanie som si vybral články týkajúce sa vraždy Slovenského, investigatívneho novinára Jána Kuciaka a jeho priateľky. To z dôvodu, že všetky články informujúce o tejto udalosti vyšli v každom periodiku približne v ten istý čas. Bolo to spôsobené závažnosťou udalosti, preto chcel každý server o nej informovať ako prvý. Celé texty a perexy týchto článkov sa nachádzajú v prílohe A. V tomto texte uvediem len odkiaľ článok pochádza, čas vydania, čas úpravy a ich výsledné podobnosti.



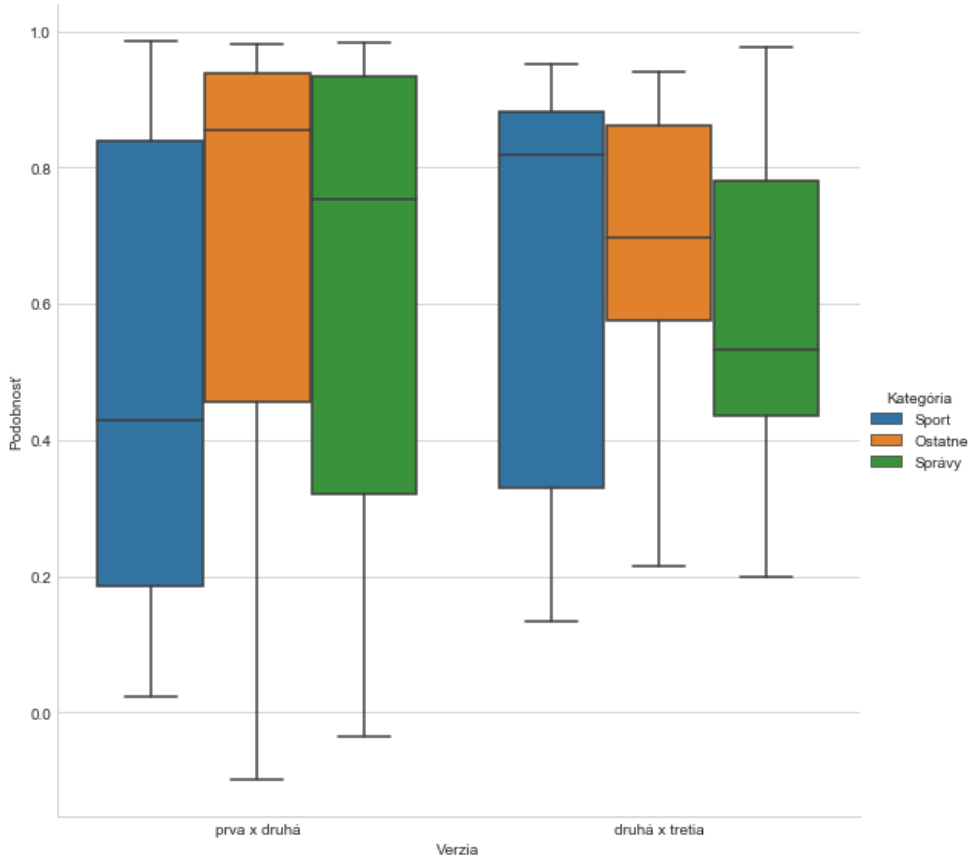
Obr. 3.4: Zmeny v titulkoch medzi verziami v závislosti na kategórii

- Výsledná podobnosť textov: 0.54064
- Výsledná podobnosť perexov: 0.38747
- Výsledná podobnosť titulkov: 0.37234

Článok zo serveru Novinky
Dátum vydania: 26. 2. 2018, 9:16
Upravený: 10:55

Článok zo serveru iDNES
Dátum vydania: 26. 2. 2018, 9:15
Upravený: 15:55

Vyhľadom na fakt, že oba texty boli vydané v ten istý čas, je text pochádzajúci zo serveru iDNES omnoho dlhší. Obsahuje mnoho detailnejších informácii



Obr. 3.5: Zmeny v perexov medzi verziami v závislosti na kategórii

a citovaných vyjadrení než Novinky ako napríklad citovanie vyjadrenia vtedajšieho premiéra Róberta Fica. Rozsah informácií môže byť ovlpyvnený tým, že článok bol upravovaný ešte o štvrtej popoludní zatiaľ čo na novinkách len o jedenástej predpoludním. Aj napriek tomu, výsledná podobnosť textov, titulkov aj perexov prekonáva 99% kvantil porovnania častí rôznych, náhodne vybratých článkov.

Záver

V rámci práce bol vytvorený Python script, ktorý zbiera dáta z českých spravodajských serverov Novinky.cz a iDnes.cz. Pri zbierani dát som sa zameral najmä na zmeny, ktoré v článkoch nastávali.

Výsledná aplikácia umožňuje zozbierať jednorázovo množinu vydaných článkov z oboch serverov. Výsledný počet zozbieraných dát je možné obmädziť časovým intervalom v ktorom sú články vydané, respektíve ich maximálnym počtom. Následne je aplikácia schopná pravidelne kontrolovať zmeny v jednotlivých článkoch a v prípade zmien nové verzie ukladať. Ďalej som v práci spracoval rešerš v ktorej som sa snažil nájsť nástroje a metódy vhodné na následnú analýzu zozbieraného textu. Vzhľadom k charakteru úlohy, ktorou sa práca zaoberá, som sa v nej zameral na sémantickú analýzu textu. Na základe tejto rešeršnej časti, som následne vybral vhodný nástroj pre analýzu zozbieraných dát. Výsledná analýza bola vykonaná za pomoci algoritmu Paragraf vektor, implementovaného v knižnici Gensim. Analýzu som realizoval z hľadiska sémantickej podobnosti textov. Výsledky mojej analýzy poukázali na časté zmeny textov na servere iDNES. Ďalej som otestoval presnosť merania podobností textov za použitia modelu Paragraf vektor.

Výsledky porovnaní ukazujú, že Paragraf vektor správne vyhodnocuje texty s rovnakým obsahom ako podobné. Správnosť fungovania je vidieť napríklad pri výsledkoch, porovnaní textu s perexom v rámci jedného článku oproti výsledkom porovnaní textu s perexom v rámci různych článkov.

Analýzu by bolo možné vylepšiť napríklad presnejším modelom. Vyššiu presnosť modelu by bolo možné dosiahnuť trénovaním na množine dát, ktorá sa podobá viac tej testovacej, napríklad na množine novinových článkov. Ďalšou možnosťou je testovanie nastavení daného modelu respektíve kombinovanie rôznych modelov.

Literatúra

[1] Seerat, B.; Azam, F.: Opinion Mining: Issues and Challenges (A survey). *International Journal of Computer Applications*, ročník 49, 08 2012: s. 42–51.

[2] Ján Paralič et al: *Dolovanie znalostí z textov*. Vydané s podporou Agentúry pre výskum avývoj, na základe zmluvy č.RPEU-0011-06, ISBN 978-80-89284-62-7, 188 s.

[3] Le, Q. V.; Mikolov, T.: Distributed Representations of Sentences and Documents. *CoRR*, ročník abs/1405.4053, 2014, 1405.4053. Dostupné z: http://arxiv.org/abs/1405.4053

[4] Jiawei Han, J. P., Micheline Kamber: *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., třetí vydání, ISBN 0123814790, 9780123814791, 744 s.

[5] Sedláček, P.: Text mining a jeho možnosti (aplikace). [cit. 12.4.2018]. Dostupné z: https://www.fi.muni.cz/usr/jkucera/pv109/2003p/xsedlac5.htm

[6] Salton, G.: *The SMART Retrieval System—Experiments in Automatic Document Processing*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1971.

[7] Turney, P. D.; Pantel, P.: From Frequency to Meaning: Vector Space Models of Semantics. *J. Artif. Int. Res.*, ročník 37, č. 1, Leden 2010: s. 141–188, ISSN 1076-9757. Dostupné z: http://dl.acm.org/citation.cfm?id=1861751.1861756

[8] Pradhan, N.; Gyanchandani, M.; Wadhvani, R.: Article: A Review on Text Similarity Technique used in IR and its Application. *International Journal of Computer Applications*, ročník 120, č. 9, June 2015: s. 29–34, full text available.

[9] Mihalcea, R.; Corley, C.; Strapparava, C.: Corpus-based and Knowledge-based Measures of Text Semantic Similarity. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*, AAAI'06, AAAI Press, 2006, ISBN 978-1-57735-281-5, s. 775–780.

[10] Miller, G. A.: WordNet: A Lexical Database for English. *Commun. ACM*, ročník 38, č. 11, Listopad 1995: s. 39–41, ISSN 0001-0782, doi:10.1145/219717.219748. Dostupné z: http://doi.acm.org/10.1145/219717.219748

[11] Li, B.; Han, L.: Distance Weighted Cosine Similarity Measure for Text Classification. In *Proceedings of the 14th International Conference on Intelligent Data Engineering and Automated Learning — IDEAL 2013 - Volume 8206*, IDEAL 2013, New York, NY, USA: Springer-Verlag New York, Inc., 2013, ISBN 978-3-642-41277-6, s. 611–618, doi:10.1007/978-3-642-41278-3_74. Dostupné z: http://dx.doi.org/10.1007/978-3-642-41278-3_74

[12] Bird, E. L., Steven; Klein, E.: Natural Language Processing with Python. 2009, [cit. 15.4.2018]. Dostupné z: https://www.nltk.org/

[13] Rus V. et al: SEMILAR: The Semantic Similarity Toolkit. 2013, [cit. 15.4.2018]. Dostupné z: http://deeptutor2.memphis.edu/Semilar-Web/public/downloads/ACL-13.SEMILAR.DEMO.pdf

[14] Ganesan, K.; Zhai, C.; Han, J.: Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd international conference on computational linguistics*, Association for Computational Linguistics, 2010, s. 340–348.

[15] Bär, D.; Zesch, T.; Gurevych, I.: DKPro Similarity: An Open Source Framework for Text Similarity. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Sofia, Bulgaria: Association for Computational Linguistics, August 2013, s. 121–126. Dostupné z: http://www.aclweb.org/anthology/P13-4021

[16] Řehůřek, R.; Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, Valletta, Malta: ELRA, Květen 2010, s. 45–50, http://is.muni.cz/publication/884893/en.

[17] Rong, X.: word2vec Parameter Learning Explained. *CoRR*, ročník abs/1411.2738, 2014, 1411.2738. Dostupné z: http://arxiv.org/abs/1411.2738

[18] Zpravodajské weby v meziročním srovnání: Roste přístup z mobilů, polepšilo si Aktuálně.cz, pohoršily Parlamentní listy. *mediahub.cz [online]*, 2016, [cit. 3.4.2018]. Dostupné z: https://mediahub.cz/tema-tydne/

900705-zpravodajske-weby-v-mezirocnim-srovnani-roste-pristup-z-mobilu-polepsilo-si-aktualne-cz-pohorsily-parlamentni-listy

[19] Leonard Richardson: Beautiful Soup. [cit. 11.3.2018]. Dostupné z: www.crummy.com/software/BeautifulSoup/

[20] Jones, E.; Oliphant, T.; Peterson, P.; aj.: SciPy: Open source scientific tools for Python. 2001–, [cit. 1.4.2018]. Dostupné z: http://www.scipy.org/

Dodatok **A**

Porovnanie rovnakých článkov z rôznych zdrojov

Článok zo serveru Novinky

Titulok

Na Slovensku byl zavražděn investigativní novinář s přítelkyní

Perex

V obci Veľká Mača nedaleko Galanty na západě Slovenska zastřelili o víkendu Jána Kuciaka, investigativního novináře serveru Aktuality.sk, který psal o daňových podvodech. Zavražděna byla i novinářova partnerka. V pondělí to potvrdilo slovenské ministerstvo vnitra.

Text

Těla muže a ženy ve věku 27 let našla policie v noci na pondělí v rodinném domě v obci Veľká Mača. Policii upozornili příbuzní, kterým se už několik dnů neohlásili. Přivolaní záchranáři už pouze konstatovali smrt. „Zabezpečili jsme všechny potřebné složky tak, aby byly přítomné na místě, aby byly vykonané prvotní úkony a důsledná obhlídka místa činu. Případ vyšetřujeme jako dvojnásobnou vraždu," řekl Denníku N ředitel Národní kriminální agentury (NAKA) Peter Hraško. Během dne má o vraždě informovat policejní prezident Tibor Gašpar. Policie uzavřela místo, kde se stala vražda. Zdroj: TV Markíza Server Denník N upozornil, že letos 9. února napsal Kuciak o podezřelých převodech v komplexu Five Star Residence. V obchodech měl figurovat podnikatel Marián Kočner. Ten měl údajně Kuciakovi vyhrožovat, že bude „hledat špínu" na jeho rodinu a informace zveřejní. „Začnu se speciálně věnovat vám, vaší osobě, vaší matce, vašemu otci a vašim sourozencům," měl říct Kočner Kuciakovi. Kuciak psal například o tom, že firma, v níž měl Kočner podíl a

postavila komplex Five Star Residence, pravděpodobně zanikne bez splacení dluhu. Budova je spojena rovněž s informací o neoprávněné dvoumiliónové vratce DPH, která měla vzniknout převodem sedmi bytů ve zmíněném komplexu za 12 miliónů eur.V roce 2017 se zavražděný novinář věnoval také dotacím, které získávaly firmy kolem podnikatele blízkého straně Směr-sociální demokracie Miroslava Bödöra. Novinářsky pokrýval rovněž podnikatelské aktivity ministra vnitra Roberta Kaliňáka a exministra financí Jána Počiatka (oba Směr).Po výhrůžkách podal Kuciak na Kočnera trestní oznámení. Dva nezvěstní Před deseti lety na Slovensku zmizel investigativní novinář Pavol Rýpal, který se věnoval mafii. Od roku 2015 je nezvěstný reportér slovenského deníku Hospodárske noviny Miroslav Pejko.

Článok zo serveru iDNES

Titulok

Na Slovensku zastřelili investigativního novináře. Zřejmě kvůli jeho práci

Perex

Na Slovensku byl minulý týden zavražděn investigativní novinář Ján Kuciak. Pracoval pro zpravodajský portál Aktuality.sk. Zemřela i jeho partnerka. Oba byli zastřeleni, policie čin vyšetřuje jako úkladnou vraždu. Kuciakovi v minulosti vyhrožovali. Policejní prezident čin odsoudil jako „bezprecedentní útok na novináře, jaký Slovensko ještě nezažilo“

Text

„Smrt Jána K. pravděpodobně souvisí s jeho novinářskou činností. Je to bezprecedentní útok na novináře a takto závažnému zločinu Slovensko ještě nečelilo. Ostře odsuzuji tento zločin a slibuji, že nasadíme maximální kapacity, abychom ho vyšetřili a poskytli odpovědi na otázky, které s ním souvisí,“ uvedl policejní prezident Tibor Gašpar. Kuciak i jeho partnerka Martina K. byli podle prvního ohledání zastřeleni - Kuciak do hrudi, jeho přítelkyně do hlavy - a zemřeli někdy mezi čtvrtkem a nedělí. Vražedná zbraň se dosud nenašla.Vyšetřujeme to jako dvojnásobnou vraždu,“ řekl Denníku N ředitel Národní kriminální agentury (NAKA) při ministerstvu vnitra Peter Hraško. Později doplnil, že agentura už zahájila „trestní stíhání pro obzvlášť závažný zločin úkladné vraždy“.Smrt svého redaktora potvrdil ráno i server Aktuality.sk . Policii zalarmovala rodina Kuciaka i jeho partnerku podle Denníku N zavraždili v jejich domě ve Veľkej Mači. Dvojice byla zasnoubena. Policii zalarmovala rodina, které se mladý pár už týden neozval. Policisté na místo dorazili v neděli ve 22:30 a oknem viděli, že někdo leží na zemi. Sedmadvacetiletý novinář se věnoval případům lidí z okolí slovenské vládní strany Smer, kteří jsou podezřelí z daňových podvodů. Poslední

text napsal o kontroverzním podnikateli a někdejším kandidátovi na primátora Bratislavy Mariánu Kočnerovi a jeho zvláštních finančních převodech například v luxusním bytovém komplexu Five Star Residence.Kočner měl Kuciakovi poté, co se jej zeptal na daňové podvody, začít vyhrožovat. Řekl mu, že bude „hledat špínu na jeho rodinu a zveřejní ji". „Začnu se věnovat speciálně vám, vaší osobě, vaší matce, vašemu otci a vašim sourozencům," pohrozil údajně Kočner novináři. Kuciak loni na sociální síti podle médií napsal, že na Kočnera podal trestní oznámení a že ani po 44 dnech nebyl případ přidělen konkrétnímu policistovi. Novinářka Zuzana Petková ze serveru Trend.sk , která s Kuciakem na řadě článků spolupracovala, si však nemyslí, že by měl Kočner s vraždou souvislost. „Absolutně nespojuji Mariána Kočnera s tím, co se přes víkend stalo. Myslím si, že vzhledem k veřejné potyčce s novinářem by žádnou ze svých konkrétních výhrůžek reálně nesplnil. Jen chci říct, že Janko ‚lezl na nervy' mnohým vlivným lidem a grázlům," uvedla.Kočner v pondělí televizi Joj řekl, že novináři nevyhrožoval, o jeho trestním oznámení dosud nevěděl a že se s ním osobně potkal pouze na jedné tiskové konferenci. Tvrdil také, že novináři prostřednictvím e-mailů v minulosti opakovaně odpovídal na jeho dotazy. Novináři vyzvali policii: Víte, o čem píšeme, vyšetřujte to Šéfredaktor serveru Peter Bardy se proti zastrašování reportérů ohradil. „A to kýmkoliv. Chápeme, že naše práce může překážet. A to zejména těm, kteří mají co skrývat. Těm, kterým překáží, když jim vidíte pod prsty. Těm, kteří se snaží obohacovat na veřejném majetku nebo dělají věci na hranici zákona nebo za touto hranicí," napsal loni v září v komentáři .Slovenští novináři na tiskové konferenci několikrát vyzvali policii, aby pečlivěji vyšetřovala daňové podvody a další trestnou činnost, o které píší. „Nasadíte maximální kapacity na vyšetření smrti Kuciaka, ale proč jste nenasadili maximální kapacity na to, abyste té smrti zabránili?" zeptala se jedna z novinářek. Gašpar však jakékoliv pochybení policie odmítl. Žádný novinář v historii Slovenska si podle něj na vyhrožování nestěžoval. „Pokud máte pocit, že pracujete na něčem, co vás ohrožuje, pojďme do komunikace," vyzval přítomné novináře. Od těch však zaznělo, že policii nedůvěřují. Odmítli také, že by měli policii vysvětlovat, o čem píší. „Vždyť vy to víte, píšeme to každý den už několik let. Proč s tím nic neděláte?" pustili se novináři do Gašpara. Novináři s ochrankou Policejní prezident také oznámil, že vybraní slovenští novináři dostanou policejní ochranku a také takzvaný „panic button", aby v případě nebezpečí mohli policii zalarmovat. Pokud se potvrdí, že vražda slovenského novináře Kuciaka souvisela s jeho prací, šlo by o útok na svobodu tisku a demokracii, uvedl premiér Robert Fico. Vraždu novináře odsoudil i slovenský ministr vnitra Robert Kaliňák. „Jsem nešťastný a zdrcený z toho, že byl zavražděn novinář Ján Kuciak a jeho snoubenka. Ruším celý svůj plánovaný program a vracím se do Bratislavy. Ministerstvo vnitra a státní bezpečnostní složky udělají vše pro to, aby byli pachatelé tohoto otřesného činu co nejdřív odhaleni," napsal Kaliňák na svém facebookovém profilu .Na Slovensku dosud nebyl známý případ vraždy novináře. Slovenská média upozornila, že deset let je nezvěstný jiný investigativní novinář Pavol Rýpal, který se dlou-

hodobě věnoval tématům spojovaným s mafií. V roce 2015 policie zase žádala veřejnost o pomoc při pátraní pro redaktorovi ekonomického listu Hospodárske noviny Miroslavu Pejkovi.

Dodatok B

Zoznam použitých skratiek

GUI Graphical user interface

XML Extensible markup language

LSA Latent Semantic Analysis

LDA Latent Dirichlet Allocation

NLP Natural-language processing

Dodatok C

Obsah priloženého CD

- `readme.txt` ..... stručný popis obsahu CD
- `src`
  - `impl` ..... zdrojové kódy implementácie
    - `database` ..... zložka s balíkom stiahnutých článkov
    - `script` ..... zložka obsahujúca scripty na crawlovanie, parsovanie a sťahovanie článkov
  - `thesis` ..... zdrojová forma práce vo formáte LaTeX
- `text` ..... text práce
  - `thesis.pdf` ..... text práce vo formáte PDF
  - `thesis.ps` ..... text práce vo formáte PS