Review report of a final thesis

Czech Technical University in Prague **Faculty of Information Technology**

Student:	Bc. Sergii Stamenov
Reviewer:	Mgr. Petr Paščenko
Thesis title:	Understanding Documents with Text Mining Methods
Branch of the study:	Knowledge Engineering

Date: 5. 6. 2017

Evaluation criterion:

The evaluation scale: 1 to 5.

1. Difficulty and other comments on the assignment

1 = extremely challenging assignment,
2 = rather difficult assignment,
3 = assignment of average difficulty,
4 = easier, but still sufficient assignment,
5 = insufficient assignment

Criteria description:
Characterize this final thesis in detail and its relationships to previous or current projects. Comment what is difficult about this thesis (in case of a more difficult thesis, you may overlook some shortcomings that you would not in case of an easy assignment, and on the contrary, with an easy assignment those shortcomings should be evaluated more strictly.)

Comments:

The field of algorithms for text processing and text data mining methods is broad and wide and it is indeed not fully covered by standard data mining curriculum. To provide a comprehensive review the student had to fully dig in the field. This is according to my belief the main source of difficulty of the assignment that qualifies is a the master thesis.

The practical part of the thesis contains a proper set of experiments of various text mining algorithms in regard to keyword extraction and document clustering as well as composition of simple text mining system. The experiments are however focused into wide rather than deep, which is the main reason the difficulty of assignment can not be higher than average.

Evaluation criterion:

The evaluation scale: 1 to 4.

2. Fulfilment of the assignment

1 = assignment fulfilled,
2 = assignment fulfilled with minor objections,
3 = assignment fulfilled with major objections,
4 = assignment not fulfilled

Criteria description:
Assess whether the thesis meets the assignment statement. In Comments indicate parts of the assignment that have not been fulfilled, completely or partially, or extensions of the thesis beyond the original assignment. If the assignment was not completely fulfilled, try to assess the importance, impact, and possibly also the reason of the insufficiencies.

Comments:

The thesis fulfilled the assignment in all theoretical and practical tasks, as well as the implementation tasks. The only and marginal exception is the missing experiment on Profinit dataset for the simple reason, that Profinit did not provide any. The experiments on other public datasets are though sufficient for the algorithms evaluation and thus it is not a relevant objection.

Evaluation criterion:

The evaluation scale: 1 to 4.

3. Size of the main written part

1 = meets the criteria,
2 = meets the criteria with minor objections,
3 = meets the criteria with major objections,
4 = does not meet the criteria

Criteria description:
Evaluate the adequacy of the extent of the final thesis, considering its content and the size of the written part, i.e. that all parts of the thesis are rich on information and the text does not contain unnecessary parts.

Comments:

The text of the thesis definitely contain all the necessary parts and it does not contain anything extra to that. All algorithms and methods are briefly described in the theoretical part. Some of them though too briefly. The opponent can not resit a thought however, that the richness of the method description is indirectly proportional to the method complexity and novelty. Example of such is the word embedding - this rather recent approach deserves more than a single paragraph to be properly introduced.

Also the brevity of the thesis tends to slightly increase as the chapters progress - a phenomenon hardly exceptional for final theses of all kinds.

Non of these can be a major objection to the thesis.

Evaluation criterion:

The evaluation scale: 0 to 100 points (grade A to F).

4. Factual and logical level of the thesis

75 (C)

Criteria description:
Assess whether the thesis is correct as to the facts or if there are factual errors and inaccuracies. Evaluate further the logical structure of the thesis, links among the chapters, and the comprehensibility of the text for a reader.

Comments:

The thesis is in general factually correct. There is no reason to doubt the student's understanding to the field. The thesis is, in consort with convention, structured in logic way. All algorithms and methods are briefly described in the theoretical part. The experiments are appropriately documented with tables and charts. The software part of the thesis is documented with schemes and diagrams.

What I miss in the experimental part of the thesis is more elaborative discussion of algorithm parametrization. Machine learning methods are in general highly parameter sensitive. From the given results, it is not clear, whether the published numbers are the result of careful and elaborative parameter selection in order to achieve the best results for every method or it is just a single run outcome. The opponent can hope for the better but must assume the worse. Not even the content of the enclosed CD is much helpful in this.

Evaluation criterion:	*The evaluation scale: 0 to 100 points (grade A to F).*
5. Formal level of the thesis	*85 (B)*

Criteria description:
Assess the correctness of formalisms used in the thesis, the typographical and linguistic aspect s, see Dean's Directive No. 14/2015, Article 3.

Comments:

The thesis is written in English. The language is splendid: clear and properly advanced in descriptions of expert topics, proving the student is familiar with the terminology and jargon. The charts however suffers from several formal flaws. When plotting a diagram of red and blue points, it is better to print it in color, otherwise it has not much sense for the reader. It is generally a good idea to use logarithmic scale for graphs with values non-uniformly distributed from zero to millions, otherwise most of the points just merge in. If on a complex plot, a part of it is marked for special focus with letters (fig. 5.1 and 5.3), this letters should be described in the subtitle or at least somewhere. The fig 5.1 is not mentioned in the text. Typography is nice, thanks to LaTeX template.

Evaluation criterion:	*The evaluation scale: 0 to 100 points (grade A to F).*
6. Bibliography	*90 (A)*

Criteria description:
Evaluate the student's activity in acquisition and use of studying materials in his thesis. Characterize the choice of the sources. Discuss whether the student used all relevant sources, or whether he tried to solve problems that were already solved. Verify that all elements taken from other sources are properly differentiated from his own results and contributions. Comment if there was a possible violation of the citation ethics and if the bibliographical references are complete and in compliance with citation standards.

Comments:

The selection of the thesis bibliography is appropriate in both quality and quantity. The student work with primary sources - mainly academic papers and conference proceedings as well as comprehensive monographs. All algorithms and methods are properly quoted. There is no violation of citation ethics currently known to the opponent.

Evaluation criterion:	*The evaluation scale: 0 to 100 points (grade A to F).*
7. Evaluation of results, publication outputs and awards	*80 (B)*

Criteria description:
Comment on the achieved level of major results of the thesis and indicate whether the main results of the thesis extend published state-of-the-art results and/or bring completely new findings. Assess the quality and functionality of hardware or software solutions. Alternatively, evaluate whether the software or source code that was not created by the student himself was used in accordance with the license terms and copyright. Comment on possible publication output or awards related to the thesis.

Comments:

The main result of the thesis, the evaluation of the three key word extraction algorithms, is negative. The keyword extractors perform poorly compared to human annotators. The key words does not provide good base for document clustering, popularity prediction nor document classification. The thesis does clearly state the negative result and it maturely discuss the reasons and consequences of this fact. What is missing is an attempt for an own solution.

The engineering part of the thesis - the text mining system - is designed and implemented as modular python application. It employs mature composition approaches with standardised interfaces based on REST API and standard libraries.

Evaluation criterion:	*No evaluation scale.*
8. Applicability of the results	

Criteria description:
Indicate the potential of using the results of the thesis in practice.

Comments:

The automated key word extraction and generation is indeed the field of significant practical utility especially in document indexing and search but also in other topics like document similarity and knowledge extraction. The thesis provides valuable comparison of the state of the art methods and a lot of motivation for future research.

Evaluation criterion:	*No evaluation scale.*
9. Questions for the defence	

Criteria description:
Formulate any question(s) that the student should answer to the committee during the defence (use a bullet list).

Questions:

1. What are the most promising research paths you see in the problem of keyword extraction.
2. Could you describe the role of text mining system in a larger general purpose software.

Evaluation criterion:	*The evaluation scale: 0 to 100 points (grade A to F).*
10. The overall evaluation	*85 (B)*

Criteria description:
Summarize the parts of the thesis that had major impact on your evaluation. The overall evaluation **does not** have to be the arithmetic mean or any other formula with the values from the previous evaluation criteria 1 to 9.

Comments:

My final evaluation of the thesis is given by the fact, that it is a properly elaborated review study and a well worked design of light but functional text mining software system.

The only reason I can not propose the A mark is, that student did not make an effort to provide his own approach or his own improvement to current methods.

Signature of the reviewer: