FACULTY OF INFORMATION TECHNOLOGY CTU IN PRAGUE

Supervisor's statement of a final thesis

Student:	Bc. Michael Mikuš
Supervisor:	Ing. Tomáš Pajurek
Thesis title:	Utilizing AI/ML methods for measuring data quality
Branch of the study:	Knowledge Engineering

Date: 19. 8. 2020

Evaluation criterion:	*The evaluation scale: 1 to 4.*
1. 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 submitted FT defines the objectives sufficiently and in line with the assignment; whether the objectives are formulated correctly and fulfilled sufficiently. In the comment, specify the points of the assignment that have not been met, assess the severity, impact, and, if appropriate, also the cause of the deficiencies. If the assignment differs substantially from the standards for the FT or if the student has developed the FT beyond the assignment, describe the way it got reflected on the quality of the assignment's fulfilment and the way it affected your final evaluation.

Comments:
Theoretical parts of the assignment were fulfilled in significantly larger extend than required and expected. Parts of the thesis containing description of specific methods and experiments is not so rich but still fulfilling the assignment, including proposition of directions for further improvements.

Evaluation criterion:	*The evaluation scale: 0 to 100 points (grade A to F).*
2. Main written part	*80 (B)*

Criteria description:
Evaluate whether the extent of the FT is adequate to its content and scope: are all the parts of the FT contentful and necessary? Next, consider whether the submitted FT is actually correct – are there factual errors or inaccuracies? Evaluate the logical structure of the FT, the thematic flow between chapters and whether the text is comprehensible to the reader. Assess whether the formal notations in the FT are used correctly. Assess the typographic and language aspects of the FT, follow the Dean's Directive No. 26/2017, Art. 3. Evaluate whether the relevant sources are properly used, quoted and cited. Verify that all quotes are properly distinguished from the results achieved in the FT, thus, that the citation ethics has not been violated and that the citations are complete and in accordance with citation practices and standards. Finally, evaluate whether the software and other copyrighted works have been used in accordance with their license terms.

Comments:
Written part has 107 content pages. It is organized into three chapters starting from more generic topics and ending with experiments with specific methods. There is also an appendix of 3 pages containing interesting design proposal of hypothetical universal system for measuring data quality ("Data Quality Unit").

Theoretical parts of the thesis (chapter 1) are well and neatly written with focus on providing readers with enough context so even the audience with very basic knowledge of the subject matter can consume it. Student wrote this part based on the research of significant number of related books and academic papers. This part also include thorough overview of existing data quality tools and proposes their simple taxonomy.

Second chapter focusing on description of interesting data quality methods, including AI/ML methods is sufficient, but a little bit brief. Namely approaches based on autoencoders and association rules were examined. Given that it is the core of the thesis, it would deserve more effort (additional methods as well as more thorough description).

Last, third chapter of the thesis contains experiment design, description of data sets, results and their interpretation. Experiment design for data quality methods proved to be very challenging and required manual enhancements to the data sets. Very interesting is the comparison of the AI/ML methods to the more traditional counterparts measuring the same data quality dimensions. Results and their interpretation are fine but more experiments with more data sets would beneficial to increase the overall credibility.

Overall, thesis is well organized, written in good English and all sources are properly cited.

Evaluation criterion:	*The evaluation scale: 0 to 100 points (grade A to F).*
3. Non-written part, attachments	*95 (A)*

Criteria description:
Depending on the nature of the FT, comment on the non-written part of the thesis. For example: SW work – the overall quality of the program. Is the technology used (from the development to deployment) suitable and adequate? HW – functional sample. Evaluate the technology and tools used. Research and experimental work – repeatability of the experiment.

Comments:

Non-written part contains mainly Jupyter notebooks with experiments (utilizing widely used libraries Tensorflow, Pandas or Numpy) and their results. The source code is consistent between experiments and is understandable. Raw experiments results are also included.

Evaluation criterion:	*The evaluation scale: 0 to 100 points (grade A to F).*
4. Evaluation of results, publication outputs and awards	*85 (B)*

Criteria description:
Depending on the nature of the thesis, estimate whether the thesis results could be deployed in practice; alternatively, evaluate whether the results of the FT extend the already published/known results or whether they bring in completely new findings.

Comments:

The thesis might serve as a great introductory material to the subject matter. A few of the novel approaches and improvements to existing methods contained in the thesis could be a topic for an academic paper, however, the thesis stops at the proposals only. Attached source codes and conceptual design of general "Data Quality Unit" are definitely useful inputs for engineering

Evaluation criterion:	*The evaluation scale: 1 to 5.*
5. Activity and self-reliance of the student	*5a:* ***1 = excellent activity,*** *2 = very good activity,* *3 = average activity,* *4 = weaker, but still sufficient activity,* *5 = insufficient activity* *5b:* *1 = excellent self-reliance,* *2 = very good self-reliance,* ***3 = average self-reliance,*** *4 = weaker, but still sufficient self-reliance,* *5 = insufficient self-reliance.*

Criteria description:
From your experience with the course of the work on the thesis and its outcome, review the student's activity while working on the thesis, his/her punctuality when meeting the deadlines and whether he/she consulted you as he/she went along and also, whether he/she was well prepared for these consultations (5a). Assess the student's ability to develop independent creative work (5b).

Comments:

Student was active and worked very hard on thesis. However, student struggled to keep focus on researching of methods and conducting experiments (as requested by supervisor) and devoted majority of time to the theoretical part of the thesis.

Evaluation criterion:	*The evaluation scale: 0 to 100 points (grade A to F).*
6. The overall evaluation	*82 (B)*

Criteria description:
Summarize which of the aspects of the FT affected your grading process the most. The overall grade does not need to be an arithmetic mean (or other value) calculated from the evaluation in the previous criteria. Generally, a well-fulfilled assignment is assessed by grade A.

Comments:

Student did a good job. The thesis is pleasant to read and can serve as good introductory material. The description of methods and experiments is sufficient, but would benefit from additional effort. In the thesis, there are several proposals for future research as well as conceptual design that, if properly extended, could be used as basis for engineering data quality systems. However, potential of these ideas is mostly not realized. Experiments with selected AI/ML methods were properly conducted, including interesting comparison with more traditional alternatives.

Signature of the supervisor: