FACULTY OF INFORMATION TECHNOLOGY CTU IN PRAGUE

Supervisor's statement of a final thesis

Student:	BA. Pragalbha Lakshmanan M.A.
Supervisor:	Ing. Milan Dojčinovski, Ph.D.
Thesis title:	Enrichment of the DBpedia NIF dataset
Branch of the study:	Web and Software Engineering

Date: 3. 6. 2019

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:

The goal of the thesis was to pre-process the DBpedia NIF dataset, i.e. execute selected text pre-processing methods.
The student fulfilled the assignment with one major objections: the student has focused only on the English language, while the requirement was to apply several pre-processing methods for several languages. Also, the implementation is not well documented and the solution is not configurable (e.g. executing particular pre-processing task, specifying language, input data path, etc.).

Evaluation criterion:	*The evaluation scale: 0 to 100 points (grade A to F).*
2. Main written part	*50 (E)*

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:
The thesis is well structured and split into relevant chapters.
The English language is satisfactory, however, the style and phrasing could be improved. For example, the student uses "...in variety of formats **like**...", which is very informal, "such as" had to be used instead. The student had to pay more attention on the style of writing and write the thesis in a more formal way.

The Czech version of the abstract seems to a translation of the English abstract using a translation software. Thus, it contains grammatical errors.

The Instruction chapter could be improved - the introductory part and the motivation part could be merged.

In the thesis, the student has included many small and simple listings/code examples, illustrating how to run some command. Such examples (page 27, 28,...) are unnecessary and disturb the reader.
Similarly, it is unnecessary to illustrate in a thesis how to read and list triples in an RDF document (e.g. Figure 3.2).

Instead of providing code examples (they can be added to the appendix) the student could provide pseudocode for the algorithms/steps he has implemented.

No introductory text for the Experiments section, thus, it is unclear what was the goal of the experiments.
In the experiments section, the student refers to "Real positive" and "Prediction positives", but there is no formal definition for these terms.

Other issues:
- Used DBPedia and DBpedia, interchangeably
- The listings are not numbered
- Screenshots instead of proper listings (Figure 3.2, Figure 4.2)
- Sections' title should start with capital letter -> Section 3.6.0.2, Section 3.7
- quite some missing whitespace character after comma

Bibliography and citations
- there are some missing citations, for example, page 17,18 - stemming, sentiment analysis, relationship extraction, sentence splitting...
- the bibliography is not formatted according to the standards

Exactly same statements (sentences) are repeated in several locations, e.g. "The DBpedia project started in 2006..." on page 2 and page 3.

There are several missing pages in the online PDF document - page 8, 13, 20, 21, 22, 24, 33, 55, 57
It seems that the student used font which is not used in the standard template. The font seems to be a bit larger than the standard font in the template.

In summary, there are missing citations, non-standard bibliography listing, informal language, some parts require more explanation, some parts are unnecessary and in the thesis can be found considerable amount of typographic errors.

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

3. Non-written part, attachments *55 (E)*

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:
The main result of the thesis are few Python scripts which implement selected pre-processing tasks.
There are few problems with the implementation:
- it is exclusively developed for English only, while the assignment was to develop solution for several languages
- it is not configurable - i.e it is not possible to select a pre-processing step, language, tool
- it is not modular - currently, the student consider NLTK as tool for execution of different NLP pre-processing tasks. It is unclear, how additional tool/system can be integrated used. It would be nice to have a standard interface which will allow integration of more tools.
- Initial idea was to develop the solution in Apache Spark (due to scalability reasons), however, the student has decided not to use Apache Spark.

Moreover, in the Animalia_(book).ttl-newLINKS.ttl example, the newly generated links have to be encoded as URI references and not literals, which is wrong.

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

4. Evaluation of results, publication outputs and awards	*50 (E)*

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 student provides several examples of the results (processed data) which have some minor issues (e.g. new links encoded as literals and not URIrefs).
These issues block the actual use of the results in practice. It was initially planned to integrate the dataset within the DBpedia core platform.

Moreover, current solution is not scalable, it take much time to process the whole DBpedia NIF dataset.

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:

The student regularly attended the meetings and always came prepared.

Since the student had no background in Semantic Web (RDF) and NLP, it was necessary to regularly explain basic concepts from these topics.
I believe that the lack of knowledge in Semantic Web, NLP (and maybe Python) is the cause for the resulting quality of the thesis.

Evaluation criterion:	*The evaluation scale: 0 to 100 points (grade A to F).*
6. The overall evaluation	*53 (E)*

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:

The ultimate goal of the thesis was to enrich the DBpedia NIF dataset with additional information by performing several text pre-processing tasks and enrich it with additional links.
This had to be implemented for several languages.
The student partially fulfilled the assignment - developed a solution only for English.
Moreover, the solution is not configurable and poorly documented. Therefore, at this stage the results can not be considered in practice (e.g. imported in DBpedia).

As for the written part - the student had to be more attention of the language of writing (style and phrasing), and in general, write the the thesis in more formal way. Also, the bibliography is listed in a non-standard way and there are few missing pages in the uploaded PDF document.

Signature of the supervisor: