FACULTY OF INFORMATION TECHNOLOGY CTU IN PRAGUE

Supervisor's statement of a final thesis

Supervisor:	Ing. Milan Dojčinovski, Ph.D.
Student:	Bc. Oleksandr Husiev
Thesis title:	Framework for Extraction of Wikipedia Articles Content
Branch / specialization:	Web and Software Engineering, specialization Software Engineering
Created on:	23 August 2021

Evaluation criteria

1. Fulfillment of the assignment

[1] assignment fulfilled
[2] assignment fulfilled with minor objections
▸ **[3] assignment fulfilled with major objections**
[4] assignment not fulfilled

The student fulfilled the assignment with few major objections:
- the output generated by the tool is invalid.
- the evaluation part is limited and the actual quality of the results is unknown.
- it is unclear whether the tool has been applied to languages other than English and what is the quality of the results.

2. Main written part

55 /100 (E)

In general, the thesis is organized into relevant chapters and satisfies the minimal requirements. However, there are several issues with the thesis:
- The work is poorly motivated (see Introduction/motivation) section.
- Some parts lack citations.
- There are some unnecessary parts, e.g. the 1.2.1 Web Ontology Language section or 1.2.3 section on SPARQL, which are not central in the thesis but are covered/described.
- The work is poorly positioned wrt the related work (see section 1.5.1). Only one related work has been identified.
- The actual implementation of the tool is poorly described. The biggest challenge of the thesis is the process of parsing and modeling the extracted information. These aspects are however not properly described.
- There are some tests described but the scope and the results from the testing are unclear. E.g. what are the results from the validation step (section 3.3.3) or what means a success rate in section 3.3.4.
- Possible directions for future work are not provided.

3. Non-written part, attachments 60 /100 (D)

The student developed a tool for parsing and extracting information from Wikipedia XML dumps and modeling this information in RDF. The used technology is suitable. There are however there are some several major problems:
- The information/output is incorrectly modelled, e.g. object URIs are modelled as literals while in RDF they have to be modelled as URIrefs. Example:
<http://dbpedia.org/resource/Anarchism?dbpv=2020-11&nif=context> <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#predLang> "http://lexvo.org/id/iso639-3/eng" .
<http://dbpedia.org/resource/Anarchism?dbpv=2020-11&nif=word_5886_5899>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type> "http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#Word" .
- Also the text context is typed as a "nif:Word" while its correct type is "nif:Context", see https://github.com/husieo/wiki-realtime-extractor/blob/master/output/nif_context.nt
- The executed experiments are minimal and at the same time the results from the experiments are unclear.
- One of the goals of the thesis was to apply the tool to English and four other selected languages. It is unclear whether these languages have been processed and the results from the processing are unknown.

4. Evaluation of results, publication outputs and awards 45 /100 (F)

At the current state, the results can not be deployed in practice. The tool requires some improvements so that it can be used by the community.
The student developed a tool which is relatively new as it parses and processes information directly from Wikipedia XML dumps.

5. Activity of the student

[1] excellent activity
[2] very good activity
▸ **[3] average activity**
[4] weaker, but still sufficient activity
[5] insufficient activity

The student had some delays but in general his activity can be summarized as "average activity".

6. Self-reliance of the student

[1] excellent self-reliance
[2] very good self-reliance
▸ **[3] average self-reliance**
[4] weaker, but still sufficient self-reliance
[5] insufficient self-reliance

The student has shown capabilities to develop independent creative work.

The overall evaluation

55 /100 (E)

The main goal of the thesis was to develop a tool for extraction of information from Wikipedia XML dumps. The student has managed to apply the knowledge acquired during the studies and developed a software which parses and extracts information from Wikipedia dumps. There are, however, some major problems with the developed software (e.g. invalid output, unclear quality of results). Moreover the thesis does not well position and describe the work. All the mentioned problems do not have a major impact on the final results. Considering my comments above I recommend mark E.

Instructions

Fulfillment of the assignment

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.

Main written part

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. 52/2021, 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.

Non-written part, attachments

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.

Evaluation of results, publication outputs and awards

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.

Activity of the student

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.

Self-reliance of the student

From your experience with the course of the work on the thesis and its outcome, assess the student's ability to develop independent creative work.

The overall evaluation

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.