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:	March 9, 2022

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 assignment has been fulfilled with major objections:
- The XML dump structure is not well researched and presented.
- The experiments/tests are not well presented and standard RDF validation procedures (e.g. SHACL) are not well integrated.
- The written part has a number of issues (listed below).

2. Main written part

55 /100 (E)

The second revision of the thesis is improved, but still there are a number of issues which are listed below.

- The work is poorly motivated (i.e. the Introduction section)
- The thesis is poorly positioned w.r.t. the related work (Section 1.5). The student discusses only the DBpedia extraction framework.
- The main contribution of the thesis (section 2.6.4) is better described, compared to the previous version of the thesis. However, there is still room for improvement and better presentation.
- The XML dump structure of Wikipedia is not clearly presented.
- The student improved the testing. However, there are two issues: 1) the presentation of the testing approaches could be better elaborated and 2) the SHACL based testing is unclear how it was executed.
- The conclusion section does not provide a clear summary of the work and the main findings.

Few factual errors and inaccuracies:
- page 3: "The Semantic Web is a Web of Data, an extension to the Web that links the related data." -> this is an imprecise and definition of the Semantic Web
- page 4: "This collection of interconnected datasets can also be referred to as Linked Data." -> Linked Data is a dataset publishing mechanism.
- page 4: "Specifically, the way URI is now serialized either in RDF/Extensible Markup Language (XML) or via N3(also known as Turtle, or N-triples)." -> URIs are not serialized but RDF graphs can be serialized in these serialization formats
- page 16: the two listed non-functional requirements are not well defined.

Non-cited paragraphs taken "as-is" from other sources:

- page 3: "Although Semantic Web is a term that has often been criticized as con- fusing, opaque, and academic, it does nonetheless capture two of the most critical aspects of these technologies:
* Semantic: The meaning of the data is not only explicitly represented and richly expressive, but it also "travels" along with the data itself;
* Web: Individual pieces of data are linked together into a network of information, just as documents are linked together on the World Wide Web."
-> equal text found at https://cambridgesemantics.com/blog/semantic-university/intro-semantic-web/many-names-semantic-web/

- page 9: following paragraph is equal with the paragraphs (section 4.1) from the paper https://svn.aksw.org/papers/2013/ISWC_NIF/public.pdf "Integrating NLP Using Linked Data", ISWC 2013, Hellmann et al.
"Internationalization Tag Set The Internationalization Tag Set (ITS) Ver- sion 2.0 is a W3C working draft, which is in the final phase of becoming a W3C recommendation. Among other things, ITS standardizes HTML and XML attributes which can be leveraged by the localization industry (espe- cially language service providers) to annotate HTML and XML nodes with processing information for their data value chain.
An example of three attributes in an HTML document is given here"

- page 9: another paragraph equal with the content in the paper mentioned above (Integrating NLP Using Linked Data):
"Ontologies of Linguistic Annotation The Ontologies of Linguistic Annotation (OLiA) provide stable identifiers for morpho-syntactical annotation tag sets, so that NLP applications can use these identifiers as an interface for interoperability. OLiA provides Annotation Models (AMs) for fine-grained identifiers of NLP tag sets. The individuals of these annotation models are then linked via rdf:type to coarse-grained classes from a Reference Model (RM), which provides the interface for applications."

- page 10: paragraph with text equal to the text in the article "DBpedia: A Nucleus for a Web of Open Data" https://link.springer.com/content/pdf/10.1007%2F978-3-540-76298-0_52.pdf on page 723.
"The most effective way of spurring synergistic research along these direc- tions is to provide a rich corpus of diverse data. This would enable researchers to develop, compare, and evaluate different extraction, reasoning, and uncer- tainty management techniques, and to deploy operational systems on the Web. The DBpedia project has derived such a data corpus from the Wikipedia en- cyclopedia."

Formatting/phrasing issues:
- page 4: "On 1.1..." -> It is unclear if it points to Figure or Listing. Proper formatting should

be "In Figure 1.1..." Similar issues occur further in the text.
- page 7: "The NLP Interchange Format (NIF) is an RDF/OWL-based format that aims to achieve between NLP tools..." achieve what?
- page 23: unknown citation -> "configuration - package that contains necessary Spring configuration, further described in ??." Similar issues are found in other locations as well.

3. Non-written part, attachments 70 /100 (C)

- The framework has been improved since the previous version. The results (RDF output) are now valid according to the RDF syntax.
- The student improved the testing however there are still some issues. SHACL is one of the standard RDF validation standards which has not been well integrated. Some SHACL shapes have been developed but the SHACL testing has not been integrated in the framework. It seems that the SHACL shapes/test have been executed manually using an online validation tool and not using an appropriate tool such as RDF Unit.

4. Evaluation of results, publication outputs and awards 68 /100 (D)

The developed framework is in the current state stable but it would require some more effort so that it can be deployed in practice. The framework works on top of the XML dumps and this makes it unique. However, some parts require improvement (e.g. SHACL integration).

5. Activity of the student

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

I would expect more active communication with the student during the work on the second revision of the thesis. Only one meeting has been requested and organized for consultation of the work. Nevertheless, the student at the meeting was prepared for this one meeting.

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

I assess the student's ability to develop independent work as "average self-reliance".

The overall evaluation 59 /100 (E)

In the second revision of the thesis the student has improved both, the software implementation and the thesis. However, with respect to the thesis, there are still

number of issues: phrasing/formatting, clarity, factual errors and inaccuracies, non-cited paragraphs and paragraphs taken "as-is" from other sources. As for the implementation, the student improved the testing, but the standard RDF validation procedures are still not well integrated (i.e. SHACL).
However, the student managed to apply the knowledge acquired during the studies and manage to develop a functional software. While the above-mentioned issues are present, still they are not very significant for the final results of the thesis. Considering my comments above I recommend grade 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.