FACULTY OF INFORMATION TECHNOLOGY CTU IN PRAGUE

Review report of a final thesis

Reviewer:	Ing. Milan Dojčinovski, Ph.D.
Student:	Bc. Tomáš Lenoch
Thesis title:	Automated extraction of personal profiles from a university domain using web scraping and NLP methods
Branch / specialization:	Knowledge Engineering
Created on:	6 June 2023

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 points from the assignment with few minor objections.
point 1) the research on web scraping techniques and NLP methods for information extraction does not clearly present the state-of-the-art methods and does not identify the gaps and issues.
point 6) the software has not been properly tested on a real-world university website(s).

2. Main written part

55 /100 (E)

The thesis is logically organized into theoretical and practical parts. While it covers most of the relevant aspects, there are some major problems with the written part:

- The problem definition does not clearly present the problems and the challenges. As part of the problem definition there are parts, such as the data model and the hierarchical tree, which do not fit well early in the thesis but in the design section.
- The related work section does not identify the gaps and problems of the related work.
- While the second part presents the implementation, the overall information extraction process is unclear.
- The selection of technologies is not well justified.
- Overall, the thesis is difficult to read and follow.
- there is missing reference for the trained model (en_core_web_trf)
- it is unclear how the system deals with situations when on a single page there are different types of entities, for example, a person and an organization information.
- Chapter 7 Deployment is non-informative.

Overall, the written part requires improvement and re-organisation of the content and improvement of the explanation of the key contributions.

3. Non-written part, attachments 65 /100 (D)

The student implemented a fully functional system. There are, however, some issues and unclarities.

- It is unclear why the concept of "semantic trees" has been introduced. Why not extract information directly from the DOM tree?
- It is unclear how the system performs extraction from dynamically (AJAX) created content.
- It is unclear what data has been used for the evaluation. What specifically are the four (A, B, C, D) university datasets? Why has the system not been evaluated on a real-world dataset, e.g. against the fit.cvut.cz website?

In summary, 1) some implementation decisions are not well justified (e.g. semantic trees), the evaluation is unclear and also some parts of the extraction process are unclear.

4. Evaluation of results, publication outputs and awards 75 /100 (C)

While the system is functional, its performance is unclear. Moreover, deployment in practice would require further improvements (proper testing and evaluation, user authentication,...). The scalability of the system is also unclear. Finally, the novelty of the system is not well justified.

The overall evaluation 60 /100 (D)

The student implemented a functional prototype of a system for crawling and extraction of information from university websites. The written part has some major flaws while the development has also some flaws with respect to the design and implementation decisions and the evaluation.
Overall, the student has managed to apply the knowledge acquired during the studies and developed a functional prototype system.

Questions for the defense

Q1: Why the concept of "semantic trees" has been introduced? Why not extract information directly from the DOM tree?
Q2: Explain how the system performs extraction of information from dynamically (AJAX) created content?
Q3: What data has been used for the evaluation? Was it a real-world dataset (website)?

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.

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.