ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

REVIEWER'S OPINION OF FINAL THESIS

I. IDENTIFICATION DATA

Thesis name:	**Using Machine Learning to Detect if Two Products Are the Same**
Author's name:	**Peter Jung**
Type of thesis :	master
Faculty/Institute:	Faculty of Electrical Engineering (FEE)
Department:	Department of Computer Science
Thesis reviewer:	Gust Verbruggen
Reviewer's department:	Department of Computer Science, KU Leuven, Belgium

II. EVALUATION OF INDIVIDUAL CRITERIA

Assignment **challenging**

Evaluation of thesis difficulty of assignment.

Incorporating semantic awareness in text matching is a challenging problem. The presence of noise and large quantities of data makes it even more challenging.

Satisfaction of assignment **fulfilled**

Assess that handed thesis meets assignment. Present points of assignment that fell short or were extended. Try to assess importance, impact or cause of each shortcoming.

A strong solution to the problem of product matching was presented, which is the main problem tackled in this paper.

Method of conception **outstanding**

Assess that student has chosen correct approach or solution methods.

Using of state-of-the-art in NLP for extracting similarity between raw titles in combination with a gradient boosting for combining this feature with other features seems appropriate.

While the experiments were very extensive, here are some minor remarks.

- An experiment on influence of the title match feature would have been interesting.
- Comparison of the presented approach with one of the existing models that were mentioned.
- I like the idea of generating more training examples, but would have liked an experiment on performance gained. Will the model not overfit on the generated examples and miss other mistakes that can happen?

Technical level **A - excellent.**

Assess level of thesis specialty, use of knowledge gained by study and by expert literature, use of sources and data gained by experience.

The thesis combines approaches from many different domains and combines them correctly. Deep learning is a very black-box approach to machine learning, but insight in the methods and their applicability was shown.

A lot of effort was spent in finding the *best tools for the job* for all parts of the pipeline and comparing them.

Formal and language level, scope of thesis **C - good.**

Assess correctness of usage of formal notation. Assess typographical and language arrangement of thesis.

The thesis is very technical, with a lot of focus on methodology and implementation details, yet still understandable. It would be easy to replicate the results.

One thing that is missing is a formal research question—what is the main problem being solved in this thesis? What data is given, and what is the answer that the method tries to solve? A full example of the input data and expected output would have provided a lot more insight.

Selection of sources, citation correctness	**B - very good.**
Present your opinion to student's activity when obtaining and using study materials for thesis creation. Characterize selection of sources. Assess that student used all relevant sources. Verify that all used elements are correctly distinguished from own results and thoughts. Assess that citation ethics has not been breached and that all bibliographic citations are complete and in accordance with citation convention and standards.	
Relevant literature was studied thoroughly for different components of the pipeline. Aside from research papers, other online materials that were consulted are thoroughly cited. Comparison with related work is minimal, however, as many related approaches are mentioned but not benchmarked with the proposed method. Some citations contain typographical errors (i.e., [2]).	

Additional commentary and evaluation
Present your opinion to achieved primary goals of thesis, e.g. level of theoretical results, level and functionality of technical or software conception, publication performance, experimental dexterity etc.
Great engagement in the open source community by submitting pull requests with bugfixes and additional functionality. Experiment are thorough, but little comparison to existing approaches. Great proficiency with software engineering.

III. OVERALL EVALUATION, QUESTIONS FOR DEFENSE, CLASSIFICATION SUGGESTION

Summarize thesis aspects that swayed your final evaluation. Please present apt questions which student should answer during defense.

I evaluate handed thesis with classification grade **B - very good.**

Date: **10.6.2010** Signature: