Call for Master Thesis

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

Words are not enough! Short text classification using words as well as entities

Are you interested in making a big impact with your thesis? Work with us on an innovative approach for short text classification.

Text classification is gaining more attention due to the availability of huge numbers of text data, which includes search snippets, news data as well as text data generated in social networks. Recently, several supervised learning approaches have been proposed for text classification. Most of them use words to generate a feature set and adopt machine learning algorithms. However, if applied to short texts, most of the standard text classification approaches suffer from issues such as data sparsity, and insufficient text length. Moreover, due to the lack of contextual information, short texts are highly ambiguous. As a result, simple text classification approaches based on words only, cannot represent the critical features of short texts properly. Thus, short text classification is much more challenging in comparison to traditional long documents.

In this thesis, to overcome the mentioned shortness and sparsity problem of short text, we will enrich the text representation by leveraging words together with entities represented by the content of the given document. More specifically, the feature set of a given (short) text will be composed of words and entities of the text. To extract the entities present in a text, existing Entity Linking Systems can be applied, such as TagMe[1].

The aim of the thesis is to develop a supervised classification based approach to classify a given short document. In the first step, existing embedding approaches such as Word2Vec[2] will be used in order to map each word and entity to a multidimensional vector space. Next, by utilizing the embeddings the feature set for the subsequent text classification will be generated. Finally, existing Machine Learning and/or Deep Learning methods will be applied to complete the classification task.

This thesis will be supervised by **Prof. Dr. Harald Sack, Information Service Engineering at Institute AIFB, KIT, in collaboration with FIZ Karlsruhe**.

[1] https://tagme.d4science.org/tagme/
[2] https://goo.gl/vGusdA

WIKIPEDIA
The Free Encyclopedia

Which prerequisites should you have?

- Good programming skills in Python or Java
- Interest in Natural Language Processing
- Interest in Semantic Web technologies
- Interest in Deep Learning technologies

Contact person:
Rima Türker
rima.tuerker@kit.edu
rima.tuerker@fiz-karlsruhe.de

aifb

Institute of Applied Informatics and Formal Description Methods
http://www.aifb.kit.edu/

Karlsruhe Institute of Technology