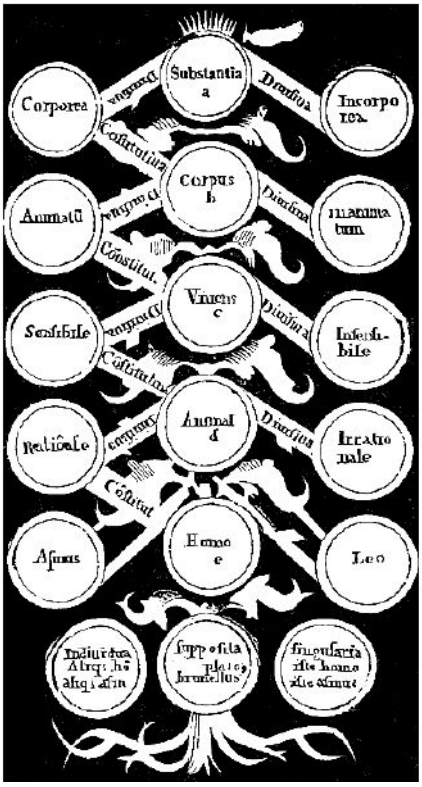
Bachelor/Master Thesis

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

Evaluation of Fine-Grained Text Classification

Work with us on an innovative approach for fine-grained text classification.



Text classification is one of the most effective ways to organize rapidly growing textual data into fine-grained classes, which enables the retrieval and exploration of information. Recently, several supervised learning approaches have achieved promising results for text classification by utilizing pretrained language models and deep learning algorithms [1]. Most of these approaches consider only a small amount of classes. However, many real world tasks have to consider a large number of classes, e.g., the English Wikipedia category systems, used to organize articles, consists of more than 100,000 categories. This poses additional challenges to the classification task, e.g., complex class relationships and class imbalances have to be included.

In this thesis, the additional challenges of fine-grained text classification are investigated to provide an overview of state-of-the-art multi-label text classification approaches and support the current research on fine-grained text classification. As a first step, a set of relevant state-of-the-art feature extraction models is going to be identified based on a literature review. A common deep learning classifier (CNN, LSTM, etc.) is trained on fine-grained classification datasets using the identified features. The Wikipedia category dataset [2] will be used. Finally, the evaluation of each feature extraction model will consider multiple numbers of classes to provide insights about the influence of the class granularity of these features.

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

[1] https://nlpprogress.com/english/text_classification.html
[2] https://arxiv.org/pdf/1503.08581.pdf

WIKIPEDIA
The Free Encyclopedia

Which prerequisites should you have?

- Good programming skills in Python or Java
- Interest in Deep Learning technologies
- Interest in Machine Learning approaches
- Interest in Natural Language Processing

Contact person:
Fabian Hoppe
fabian.hoppe@kit.edu
fabian.hoppe@fiz-karlsruhe.de

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

Karlsruhe Institute of Technology