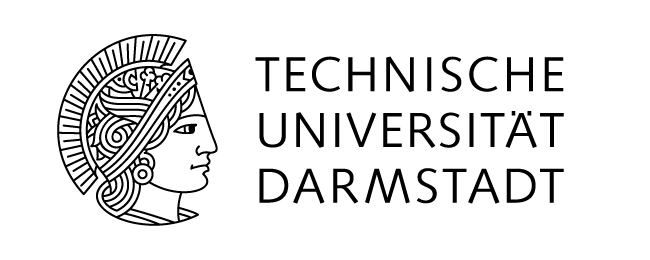
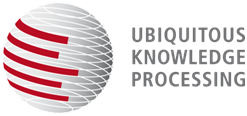
|  |  |
| --- | --- |
|  |  |
|  | **Text Classification Using Concept Maps** |
|  |  |
|  |  |
|  | **Bachelor Thesis Task Description** |
|  |  |





Name: David Marlon Gengenbach  
Matriculation No.: 2510213  
Department: Computer Science (FB 20)  
Supervisor: Prof. Dr. Iryna Gurevych, Prof. Dr Kristian Kersting  
Responsible Staff: Tobias Falke

# Motivation

Concept maps are a powerful tool to structure information and make them accessible. They have found numerous applications in different domains. Consequently, several approaches have been proposed in the natural language processing community to automatically extract them from text, with [1] being the most recent attempt of this kind.

While most work creates concept maps so that they are used directly by humans, motivated by their interpretability and ease of use, it is yet unknown whether they are also useful as a representation of a document’s content that can be used in subsequent tasks. To explore this idea, this thesis focuses on text classification and studies whether graph-based representations are helpful as features in such classifiers.

# Goal

The goal is to determine whether concept maps can be used to effectively classify text documents. To do this, we compare our approach to baseline classifiers which use other types of graphs or no graph-based features at all.

In particular, the following approaches should be compared:

* Baseline 1: Non-graph-based features, e.g. bag-of-word approaches using ngram counts or tf-idf scores
* Baseline 2: Word co-occurence graphs (e.g. as in [5])
* Method: Concept map graphs [1]

First, the concept graphs are extracted from a given corpus of text documents using the method devised by [1]. Then, features are derived from the graph (as well as for baseline 2) with the following methods:

* Graph kernels [2], especially the Weisfeiler-Lehman method [3]
* Graph Convolutional Networks, e.g. [4]

Those feature representations of the documents, together with their labels, are then used to train a classifier, eg. a SVM. Several metrics are then evaluated on the results of the classification and compared to the results with baseline classifiers which do not use the concept-graph based approach.

The thesis will start with the traditional 20 newsgroups classification dataset and might move to other classification tasks and datasets later.

# Prerequisites

* Java/Python programming language
* Graph algorithms, analysis, and storage
* Foundations of natural language processing

# Software and data provided

* Concept graph extraction library (provided by Tobias Falke)
* Datasets are freely available

# Results and Deliverables

* Thesis, describing details and results of all performed experiments as well as the necessary foundations and appropriate conclusions
* Documented code for all performed experiments including necessary preprocessing steps

# Schedule

The following time plan merely provides a coarse schedule. Several stages might overlap or change slightly in the course of the project. Regular meetings with the supervisors are scheduled, in order to discuss work in progress and emerging problems.

|  |  |  |
| --- | --- | --- |
| **Week** | **Task** | **Comment** |
| 0 | Initial meeting; discussion of the project definition |  |
| 0–2 | Familiarization with the project, literature review, setup of the environment and the datasets |  |
| 3-4 | Implementation of baselines, setup of experiments, create first results |  |
| 4-6 | Implementation and running of concept-map based experiments |  |
| 6-10 | Additional experiments, extensions, analysis, defined based on first results |  |
| 6-11 | Writing the thesis |  |
| 11 | Final thesis redaction |  |
| 12 | Submit the thesis |  |
| 12 | Final presentation |  |

# References

[1] Falke, Tobias and Gurevych, Iryna (2017): Concept-Map-based Multi-Document Summarization using Concept Co-Reference Resolution and Global Importance Optimization. (currently under review)

[2] Vishwanathan, S. V. N. et al. (2010): Graph Kernels. In: Journal of Machine Learning Research. Volume 11, pages 1201-1242.

[3] Douglas, B.L. (2011): The Weisfeiler-Lehman Method and Graph Isomorphism Testing. arXiv:1101.5211

[4] Henaff, Mikael et al. (2015): Deep Convolutional Networks on Graph-Structured Data. arXiv:1506.05163

[5] Mihalcea, Rada and Tarau, Paul (2004): TextRank: Bringing Order into Texts. In: Proceedings of EMNLP 2004, pages 404-411, Barcelona, Spain.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
 Place, Date Signature of the Student

□ The project has been introduced in the DKPro-Meeting.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Signature of the responsible co-worker

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Signature of Prof. Dr. Iryna Gurevych