Karlsruhe, 15. August 2018

The evaluation of topic models

Bachelor / Master Thesis (in German or English)

Topic models[1] are an increasingly popular text-mining tool. They rely on methods of unsupervised machine learning and natural language processing to identify topics within a body of text documents. As with many models of unsupervised machine learning, the evaluation of a trained topic model can be challenging. At the same time, the evaluation is crucial to improve the interpretability of the identified topics and to find optimal values for the models parameters. Especially to find an adequate number of topics, which has to be chosen before training, evaluation metrics are essential.

The aim of the proposed thesis is to give an overview over commonly used evaluation metrics in the context of topic modelling (esp. perplexity and coherence). Furthermore, alternative methods to find an optimal choice for the number of topics can be explored. This includes e.g. hierarchical topic modelling and (hierarchical) clustering.

Possible tasks:

- Explain common metrics in detail and discuss their advantages and disadvantages.
- Patent data provided by the chair can be used to compare the results of different metrics.
- Explore hierarchical topic modelling and (hierarchical) clustering of text data.
- New evaluation methods specific to the domain of patent data (availability of structured and unstructured data) can be developed.

What you should bring:

- Structured way of thinking and working
- An interest in statistics and machine learning
- Basic programming skills in Python

Are you interested?

- Contact: david.baelz@kit.edu

[1] See for example: https://www.coursera.org/lecture/text-mining.