Department of Economics and Management
Institute of Economics (ECON)
Chair in Economic Policy
Prof. Dr. Ingrid Ott

Karlsruhe, 15. August 2018

Topic models – a look under the hood

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. The term topic models refers to a group of (probabilistic) algorithms to find latent (i.e. hidden or unobservable) semantic structures in texts. While the idea, how topic models work, is usually easy to grasp, to understand the underlying mechanisms can be challenging.

The aim of the proposed thesis is to gain a better understanding of how topic models work. The idea is to start with a simple model (e.g. the mixture of unigrams model) and to work up to more sophisticated models (e.g. Latent Dirichlet Allocation (LDA)).

Possible tasks:

- Understand and formulate the underlying statistical problem in a topic model.
- Explain and be able to implement approximate solution methods (usually analytical solutions are not tractable).
- Thereby understand relevant methods like the Expectation-maximization algorithm or Gibbs sampling.
- Different models and solution methods can be tested on patent data provided by the chair.

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.