Reimagining Topic Models with Feature Engineering

Anna Moskvina
Institut für Informationswissenschaft und Sprachtechnologie,

Universität Hildesheim

moskvina@uni-hildesheim.de

This study was carried out in the framework of the BMBF-funded project Rez@Kultur, where the contents of online book reviews is being analyzed. We propose to apply differently trained topic models on data aggregated from the online platform Amazon (McAuley, Leskovec 2013) in order to create a more thorough description of review content that is at the same time understandable to humans.

One of the main concerns when working with topic models is the difficulty humans have in interpreting the results, especially when the topic models are applied to an inhomogenous corpus the text subject of which varies from medical articles to historical novels.

The research conducted by F. Martin and M. Johnson showed that just training a topic model on a corpus consisting only of nouns significantly improves the outcome as well as the interpretability by humans (word intrusion evaluation, introduced by J. Chang). Though A. Schofield and D. Mimno later indicated that though it is assumed that introducing stemmers to topic models would also result in better topics, the effect was negligible and the procedure was time-consuming. However both of the studies involved only English language, which is by far less morphosyntactically complex than German.

We proposed to go a step further by combining bigrams/trigrams with part of speech information. Instead of using documents comprising of a sequence of unlemmatized words as input for training the models, we represent a document as a sequence of syntactically relevant pairs: combinations of lemmatized nouns and verbs as well as combination of lemmatized adjectives and nouns. Additionally we ran experiments with applying the algorithm on sequencies of bigrams and trigrams.

We trained multiple models on 27.857 German reviews, by using the Gensim implementation (see R. Rehurek, 2010) of the Latent Dirichlet Allocation algorithm.

References: • Chang J., et al (2009): Reading Tea Leaves: How Humans Interpret Topic Models. In *Advances in neural information processing systems.* 288-296. • Lau J., Newman D., Baldwin T. (2014): Machine Reading Tea Leaves: Automatically Evaluating Toppic Coherence and Topic Model Quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014)*, 530-539. Gothenburg, Sweden. • Martin F., Johnson M. (2015): More Efficient Topic Modelling though a noun only approach. In *Proceedings of the Australian Language Technology Association Workshop 2015*. • McAuley J., Leskovec J. (2013): Hidden factors and hidden topics: understanding rating dimensions with review text. *RecSys*.• Rehurek R., Sojka P. (2010): Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*.• Schofield A., Mimno D. (2016): Comparing Apples to Apple: The Effects of Stemmers on Topic Models. In *Transactions of the Association for Computational Linguistics* 4. 287-300.