Making topic modeling easy: A programming library in Python

Fotis Jannidis
fotis.jannidis@uni-wuerzburg.de
University of Würzburg, Germany

Steffen Pielström
pielstroem@biozentrum.uni-wuerzburg.de
University of Würzburg, Germany

Christof Schöch
hristof.schoech@uni-wuerzburg.de
University of Würzburg, Germany

Thorsten Vitt
thorsten.vitt@uni-wuerzburg.de
University of Würzburg, Germany

Topic modeling, a method for the semantic analysis of large text collections, has been in the focus of interest in digital literary studies during the recent years. The method uses probabilistic procedures to generate probability distributions for words out of a collection of texts, sorting many single word distributions into distinct semantic groups called 'topics'. These topics constitute groups of semantically related words, and the contribution of each topic to the composition of each text can be quantified mathematically (Blei 2012, Steyvers und Griffiths 2006). In digital literary studies, topic models can be interesting in themselves. For example their dynamic development either during the plot of single literary texts or over multiple texts in a stage of literary history can be analyzed (Jockers 2013, Blevins 2012, Rhody 2012, Schöch to appear), though comparing literary themes and the probabilistic concept of 'topics' described here is obviously not unproblematic. And topic models can also be interesting features for classifying or clustering texts (Blei 2012).
There are currently two state-of-the-art implementations of the relevant algorithms: 'Mallet' (McCallum 2002) and 'Gensim' (Rehurek 2010). But usually more is required than simply running a topic modeling algorithm (Fig. 1):

- Longer texts like novels need to be split into smaller parts (e.g. paragraphs, scenes, or a fixed amount of characters or words).
- NLP based preprocessing is necessary
- To achieve optimal results, texts must be reduced to content words, either by filtering out function words with stopword lists, or by using a part-of-speech tagger to exclude unwanted word classes.
- Similarly, lemmatization and elimination of proper names can be useful.
- After the topics have been generated, results are usually visualized based on the relevant metadata.
- Results need to be evaluated with regard to internal or external criteria rather than just being left to interpretation.

The aim of our work is to provide digital literary scholars with a consistent, extensive and well documented programming library that allows them to do the necessary preprocessing, to generate topic models relying on the existing implementations, and to visualize and evaluate results within a single convenient scripting environment. We want to empower users with little or no previous experience and programming skills to create custom workflows mostly using predefined functions within a familiar environment (see the current stage of development on Github). Hereby, we want to lower the access threshold to topic modeling as a method, facilitating researchers in spending time experimenting with topic modeling and understanding how it generates results, rather than spending it for acquiring advanced technical skills before being able to try topic modeling at all.

The library will be developed for the programming language 'Python' that is well suited for NLP and data analysis tasks, and popular among digital literary scholars already. In addition, development can be partially based on functions from, and experiences made with a previous Python-based implementation of a topic modeling workflow developed by Christof Schöch et al., that can be regarded as a proof of concept.

A convenient tool for NLP analysis during preprocessing does exist in the DARIAH-DKPro-Wrapper (DDW) that covers a wide range of NLP tasks for many different languages and generates annotations in a Python-Pandas compatible format easily usable within our library.

Evaluating the results of topic modeling is not a trivial task but has proven to be rather challenging

(Wallach et al. 2009, Chang et al. 2009). In order to evaluate topics we will provide functions for intrinsic evaluations, for example perplexity, and external evaluations, for example path length in a resource like wordnet. We will also support task based evaluation where topics are used for a classification task and evaluated on the basis of the results.

For the visualization of the results we can build on the tmw library mentioned above and provide plots of topics over time (or other dimensions), distribution of topics over texts using heat maps and others.

Functions will be designed with consistent syntax and in a way that allows users to grasp easily what they do and why, so users can combine them into scripts to implement their own project ideas with minimal learning effort. The ability to customize workflows will be facilitated by a thorough tutorial describing all functions, outputs and potential combinations in detail (see https://www.penflip.com/c.schoech/tmw-tutorial as an example for what we are planning).

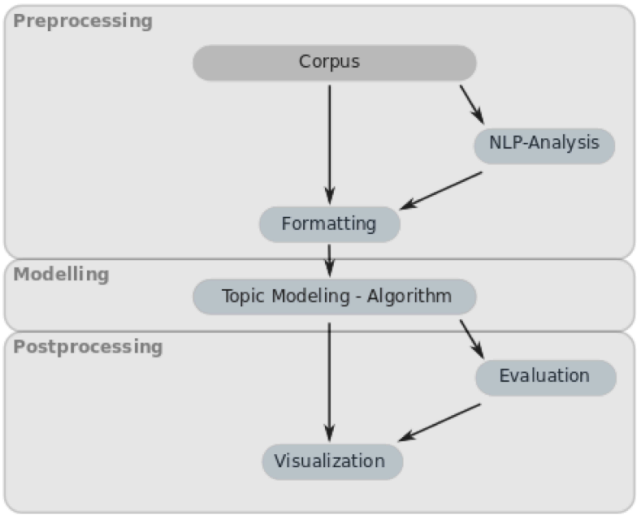

Figure 1: Topic modeling project workflow

Bibliography

Blei, D. M. (2012): „Probabilistic Topic Models". *Communication of the ACM* 55, Nr. 4 (2012): 77–84. doi:10.1145/2133806.2133826.

Blevins, C. (2010): „Topic Modeling Martha Ballard's Diary". *Historying*. http://historying.org/2010/04/01/topic-modeling-martha-ballards-diary/.

Chang, J. (2009): Reading Tea Leaves: How Humans Interpret Topic Models. In: Y. Bengio et al.: Advances in Neural Information Processing Systems 22, 288--296.

Jockers, M. L. (2013). *Macroanalysis - Digital Methods and Literary History*. Champaign, IL: University of Illinois Press.

McCallum, A. K. (2002): *MALLET : A Machine Learning for Language Toolkit*. http://mallet.cs.umass.edu.

Rehurek, R., and Sojka, P. (2010): "Software framework for topic modelling with large corpora." *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*.

Rhody, L. M. (2012): „Topic Modeling and Figurative Language". *Journal of Digital Humanities* 2.1. http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody/

Schöch, C. (to appear): „Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama". *Digital Humanities Quarterly*. http://digitalhumanities.org/dhq. Preprint: https://zenodo.org/record/48356.

Steyvers, M., and Griffiths, T. (2006). „Probabilistic Topic Models". In *Latent Semantic Analysis: A Road to Meaning*, herausgegeben von T. Landauer, D. McNamara, S. Dennis, und W. Kintsch. Laurence Erlbaum.

Wallach, H. M. (2009)et.al.: Evaluation Methods for Topic Models. In: Proceedings of the 26th International Conference on Machine Learning, Montreal, Canada, 2009.