Leveraging Topic Models for the Study of Chinese Conceptual History

Gesa Stupperich

Institute of Sinology, University of Heidelberg

Objectives

- Develop an extensible framework for estimation and visualization of topic models of user-defined Chinese source collections.
- Find out how to turn topic models into a productive tool for the study of conceptual history, one that usefully complements manual browsing and close reading.

Introduction

LDA (latent Dirichlet allocation) topic models are generative statistical model that uncover the hidden thematic structure of document collections [1]. They identify patterns of co-occurring words in textual data. These clusters are helpful for detecting semantically related sets of words across a text collection. Conceptual history and related approaches select a linguistic unit of analysis such as a "concept", "discourse" or "language" and study the "knowledge" (views of social reality and behavioral patterns or "models for action") that these linguistic forms help construct. The usefulness of topic models for the study of conceptual history cannot be evaluated quantitatively. It can only be evaluated qualitatively by assessing how helpful the model is in forming hypotheses about general characteristics and tendencies of language use. The presented framework is intended to serve as a tool for this kind of assessment, and as a testing field for appropriate modes of visualization and the impact of different sets of hyper-parameters.

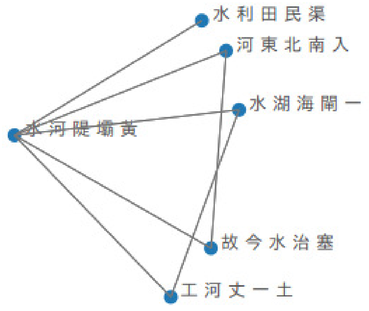

Figure 1: A cluster of semantically related topics

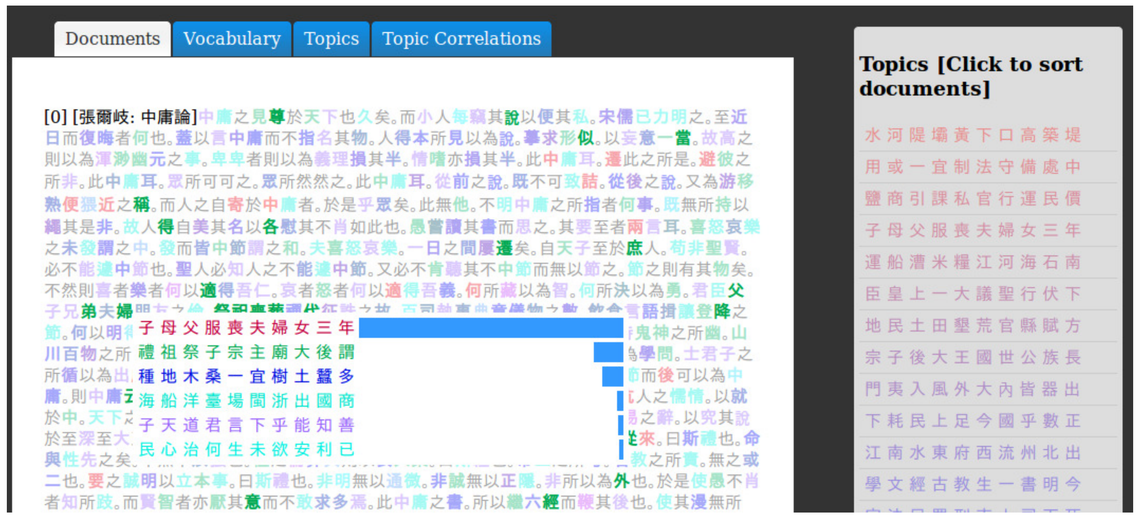

Figure 2: Document view with document-topic proportions and word-topic proportions, and sidebar with topic list

LDA Topic Models

LDA uncovers the hidden semantic structures of document collections by assigning each word in each document a topic that is global to the whole collection of documents. LDA imposes a model that assumes probabilistic relationships between topics, documents and words. These relationships are expressed with probability distributions. The most important assumptions and the corresponding distributions are

1. Any document d consists of a mixture of topics: $p_d(t)$ is the proportion of topic t in d (fig. 2).
2. Any topic t has a certain probability of being realized through word w: $p_t(w)$ is the probability of realizing t through w (fig. 3).

These distributions are inferred by fitting the model to the data with algorithms such as Gibbs sampling or variational inference.

Framework

The presented framework is realized as a Javascript web application that estimates and visualizes topic models for user-defined source collections. It was developed on top of a clone of David Mimno's *jsLDA* [2]. It adapts functionality to visualize mixture proportions from Andrew Goldstone's *dfr-browser* [3].

Conceptual History

Conceptual historians extensively scan their sources to get an overview of the terminology of a discourse, and to arrive at hypotheses concerning general tendencies regarding word use. They then analyze a subset of the sources in detail in order to qualify the hypotheses previously formed. LDA topic models can assist in identifying tendencies during the extensive scanning of source documents.
Exploration of the topic model can assist the conceptual historian in

- tracking topics (discourses) of interest
- identifying key words and their different meanings or kinds of usage (relations to other words)

The historian checks the hypotheses formed during exploration with close reading and against prior knowledge of the source collection. He or she thereby assesses the usefulness of topic models as a tool in historical research:

- Do the topics correspond to actual discursive tendencies that are representative of the whole collection?
- Which hyperparameter settings (number of topics, symmetric or asymmetric priors) achieve the best results in this regard?

Conclusion and Future Work

Topic models offer an alternative perspective on collections of source texts. Collocation statistics (fig. 3, second column) help a lot in making sense of topics, owing to the fact that semantic units of Qing dynasty Chinese often consist of more than one character. It would be worth to implement the "topical word-character model" [4] which already takes this fact into account and assigns topics both on the character and the word level. In order to better account for shifts in meaning *over time*, it makes much sense to incorporate time data into the visualization of topic proportions over time, but also to implement the "dynamic topic model", that incorporates time data into the model [5]. A further worthwhile addition would be non-symmetric alpha priors (on the document-topic distribution) to avoid domination of topics by very common words [6].

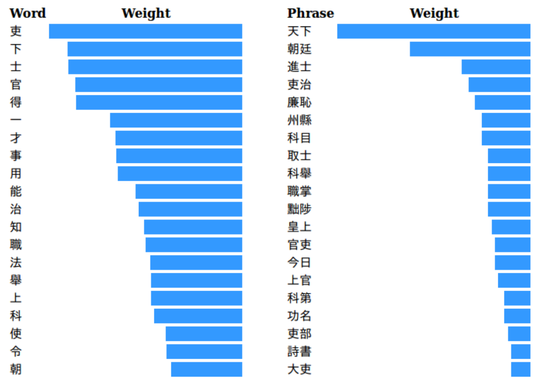

Figure 3: Top 15 topic words and topic phrases

References

1. Blei, David M., Andrew Y. Ng, Michael I. Jordan, John Lafferty. "Latent Dirichlet allocation" in: *Journal of Machine Learning Research*, Number 3, January 2003. 993-1022.
2. Mimno, David. "jsLDA: An implementation of latent Dirichlet allocation in javascript" (https://github.com/mimno/jsLDA), 2016.
3. Goldstone, Andrew. "dfr-browser: Take a MALLET to disciplinary history" v0.8a (agoldst.github.io/dfr-browser), 2016.
4. Hu, Wei, et al. "Modeling Chinese documents with topical word-character models" in: *Proceedings of the 22nd ICL*, 2008.
5. Blei, David M., and John D. Lafferty. "Dynamic topic models" in: *Proceedings of the 23rd ICML*, 2006.
6. Wallach, Hanna, David Mimno and Andrew McCallum. "Rethinking LDA: Why Priors Matter", NIPS, 2009, Vancouver, BC.