**JӒHNICHEN — Exploratory**

<5 figures>

**Exploratory Search through Interactive Visualization of Topic Models**

Jähnichen, P., Oesterling, P., Liebmann, T., Heyer, G., Kuras, C. and Scheuermann, G.

This paper addresses exploratory search in large collections of historical texts. By way of example, we apply our method to a collection of documents comprising dossiers of the former East German Ministry for State Security and classical texts. The bases of our approach are topic models, a class of algorithms that define and infer themes pervading the corpus as probability distributions over the vocabulary. Our topic-centered visual metaphor supports exploring the corpus following an intuitive methodology: First, determine a topic of interest; second, suggest documents that contain the topic with ‘sufficient‘ probability; and third, browse iteratively through related topics and documents. Our main focus lies on providing a suitable bird’s-eye view onto the data to facilitate an in-depth analysis in terms of the topics contained.

# \* \* \*

When dealing with large collections of digitized historical documents, very often only little is known about the quantity, coverage, and relations of contained topics. In order to get an overview of the topics covered, an interactive way to explore the data is needed that goes beyond simple ‘lookup’. The notion of exploratory search has been coined by Marchionini (2006) to cover such cases.

Consider a large corpus of documents. Typically, we know the source and the broader scope of such corpora, but not necessarily the content of individual documents. One classical option to explore this data is based on keyword search. While this approach is useful when the user ‘knows’ what she is looking for, an iterative exploration of the corpus is not possible. Our approach is a structured one. We provide the user with a bird’s-eye view of the data. She then identifies topics of interest and finds the documents related to them. Additionally, documents can also be related to other topics, a connection that helps to reveal new and interesting insights previously unknown. We are also able to identify different contexts, i.e., topics, in which specific terms appear. Especially when working with historical texts, this might help to reveal new aspects of known concepts.

Topic modeling (Blei and Lafferty, 2009) has become one of the main tools for unsupervised text analysis and for gaining insight into corpora via exploratory search and analysis by identifying semantic classes of words in the corpus, coined ‘topics’. An introduction on how topic modeling can help humanists in their research is given in Blei (2012).

However, little effort has been made to develop methods to use the outcome of these models in applications *visually*.

Traditional linguistic approaches such as the vector space model (Salton and Buckley, 1988) translate documents into high-dimensional feature vectors (typically in combination with, e.g., the tf-idf [Jones, 1972] term weighting). Visualizing such high-dimensional structures is a research field on its own. Established approaches include projective techniques, like the Text Map Explorer (Paulovich and Minghim, 2006), Multidimensional Scaling (MDS) (Kruskal and Wish, 1978), Sammon’s mapping (Sammon, 1969), and neuro-computational algorithms like Self Organizing Maps (SOM) (Kohonen, 2001).

Compared to the algorithms used in this paper, the insights obtained from a pure vector space model are rather limited. In fact, we reduce the dimensionality of the space in which documents are defined from the size of the vocabulary to a number of semantic classes (topics) that appear across documents.

# Topic Models

A topic model is a Bayesian hierarchical probabilistic model. It defines an artificial generative process for document generation, describing how the actually observable data (the words in the documents) get into their place. In a topic model, this is controlled by two latent factors: the topics themselves and the documents’ topic proportions. A topic is defined as a probability distribution over the word simplex—i.e., in every topic each word has a certain probability, and the probabilities in each individual topic sum to 1. The set of words with the highest probability is assumed to describe the individual topics thematically. The second factor, the document topic proportions, is again a set of probability distributions (one for each document) defined over the topic simplex. Every topic gets some probability in a document, and their probabilities sum to 1. Simply put, the words that we see in a document are then generated by first finding a topic through the document’s distribution over topics and then finding a word from the chosen topic. Both choices are random draws from their respective distributions. During inference, we reverse this process in order to get approximations for the governing latent factors that best give rise to the observed words—that is, we want to find the setting of the latent factors for which the observed words are highly likely. In most cases, the outcome of such models is presented as a set of sorted word lists for each topic, topic similarity is being neglected, and only hand-picked topics are selected as examples.

# Visualization Approach

Visualization is our method of choice to inspect the knowledge uncovered from the data. However, the model outcome is obviously inappropriate for direct visualization. Without using thresholds, presenting entire probability distributions as sorted lists of words and values is not very handy and quickly results in information excess and cluttered visualizations. Even working with thresholds does not immediately lead to parameter settings that are independent of the input data, e.g., how many words are actually necessary to obtain a reasonably good impression of a topic found by the model. That is, depending on the semantic quality of words and topics, a flexible level of detail is necessary to identify meaningful information in a topic, i.e., in a distribution over words. On the other hand, the amount of information relevant for each element of the topic model is assumed to be rather small. Therefore, the visual implementation of these elements should focus on the pivotal parts of the distributions, while increasingly disregarding irrelevant parts. In the end, the relations between the input documents, the latent topics found by the model, and the actual probabilities of a topic’s keywords are the elements containing interesting insights about the data.

We present an interactive visual analysis tool to find and display these relations. Our approach is closely related to that of Chaney and Blei (2012), who also attempt to directly visualize the output of topic models. However, their approach generates a set of static websites that can be browsed to explore the dataset, providing a lightweight but largely text-based application to solve the visualization tasks. There exists an overview over the different topics, but each topic is described by the three most probable words only. There is no possibility to further investigate a topic *and* to keep track of the others. As a framework for our visualization we chose OpenWalnut.1 This framework primarily aims at medical imaging but is highly modular and can be easily reused to fit our requirements.

We define the following distinct exploration tasks that a topic model is fit to execute:

1. **Examining a topic.** Examining a single topic is difficult because it is a probability distribution. In the visualization, this information helps the user to quickly identify key words and their relative importance for a topic.

2. **Overview of the Topics.** This task visually summarizes the set of latent topics found by the topic model.

3. **Finding Different Semantics of Polysems and Homonyms.** One advantage of topic models is the automatic disambiguation of semantic meanings of words into topics.

4. **Identifying Documents Covering a Topic.** This task is at the core of exploratory analysis: Having identified a topic of interest, explore a set of documents covering it.

5. **Finding Related Topics of a Document.** Inspect topics related to a document, or, in a transitive way, documents related to these topics.

We present visual implementations for these tasks to provide the user with interactive means to browse through relations between documents, topics, and words. This way, the user uncovers expected or (more interestingly) unexpected facts that eventually lead to interesting documents. Using smooth level-of-detail transitions and by interacting with topic distribution charts, the user freely navigates through the data by concatenating single exploration tasks—following focus-and-context concepts and an intuitive methodology: overview first, details on demand.

# Experiments

We report use cases of fitting topic models to two different datasets. One of the datasets is the series of publications ‘The GDR through the Eyes of the Stasi: The Secret Reports to the SED Leadership’ (German: ‘Die DDR im Blick der Stasi. Die geheimen Berichte an die SED-Führung’) that reveal the Stasi’s2 specific view of the GDR, containing references to real and perceived oppositional conduct as well as to economic and supply problems.

The second dataset is the ECCO-TCP,3 a set of classical literature and non-fiction texts.

In Figures 1 through 5 we pictorially describe how the defined exploration tasks are met using our visualization. Note that we omitted the application’s control panel in Figures 1–3 for spatial reasons.

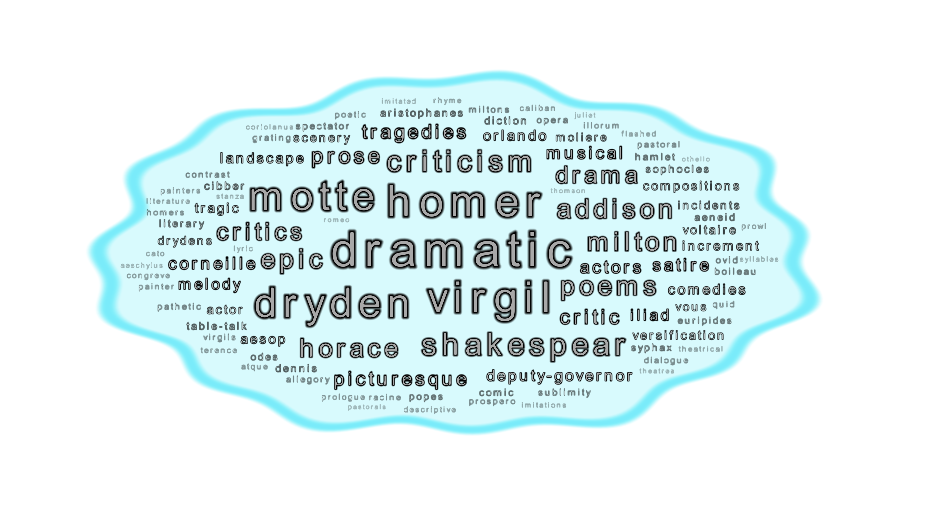


Figure 1. Visually solving task 1, the examination of a specific topic. (left) A topic covering several classical drama authors and their work from the ECCO-TCP dataset. (right) A topic covering Communist Party propaganda in connection with the youth from the Stasi files dataset.

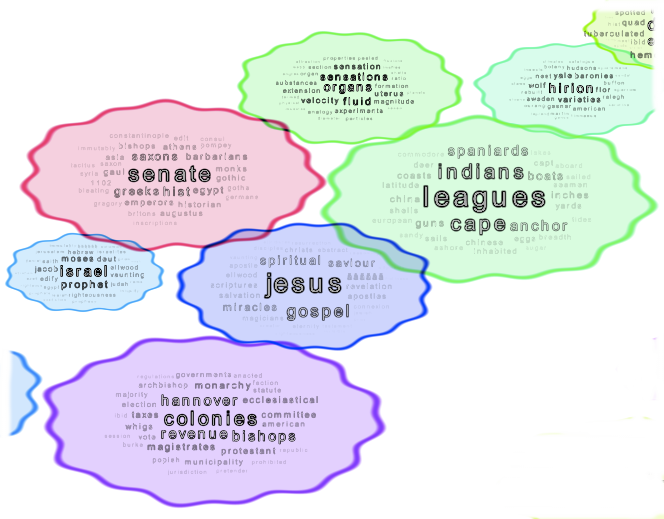


Figure 2. Visually solving task 2, giving an overview over the topics. (left) Example topics from the ECCO-TCP dataset. (right) Examples from the Stasi files data set.

Figure 3. Visually solving task 3, resolving polysems and homonyms. (left) Different usages of the term ‘greeks’ in the ECCO-TCP dataset. Usages in connection with religion, literature, science, and politics are apparent. (right) Usages of the term ‘untersuchungen’ (Eng.: investigations) from the Stasi files dataset. The term is used in connection with the ever suspicious Stasi (topics dealing with addresses, name, Berlin, ‘wohnhaft’ etc. but also with investigations of accidents in factories (‘veb’ meaning ‘Peoples-owned enterprise’) and the state railroad (‘reichsbahn’).

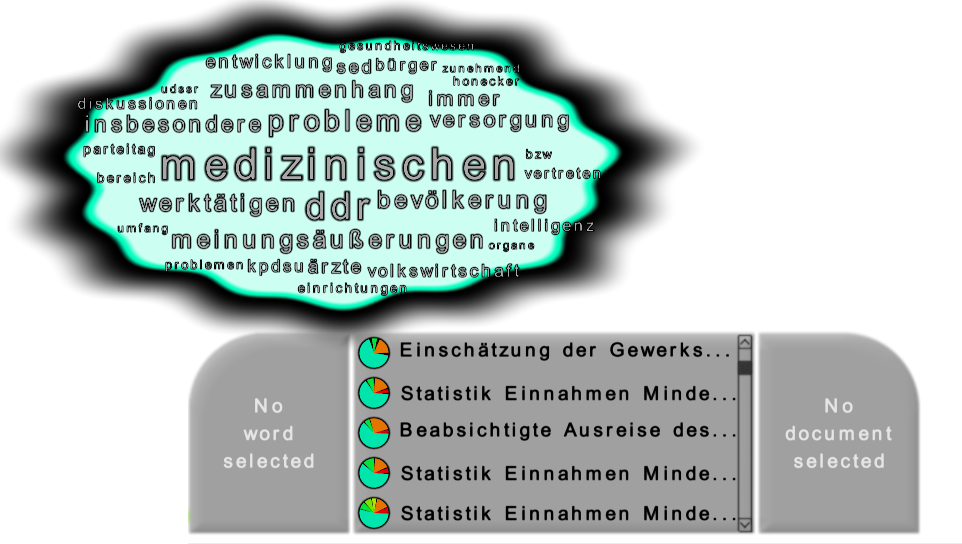
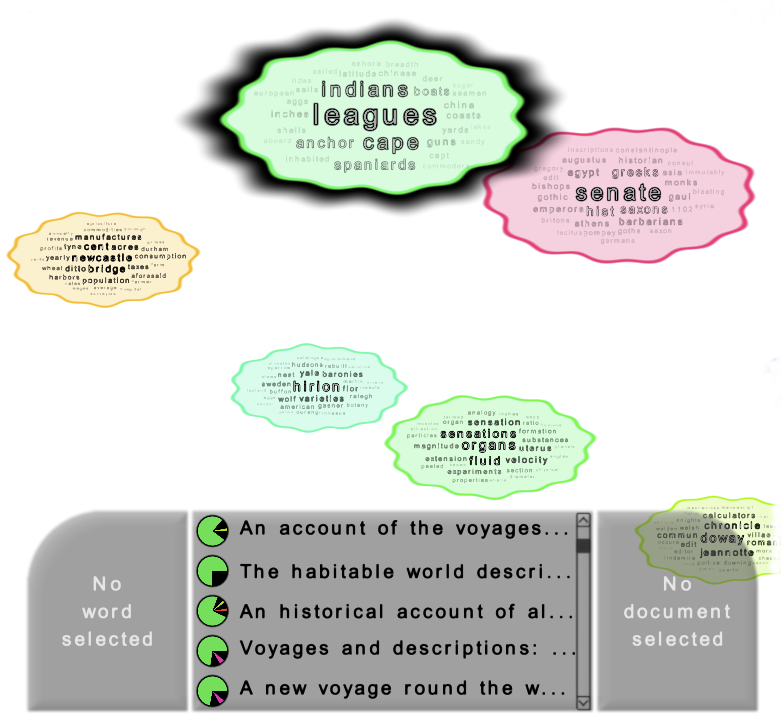
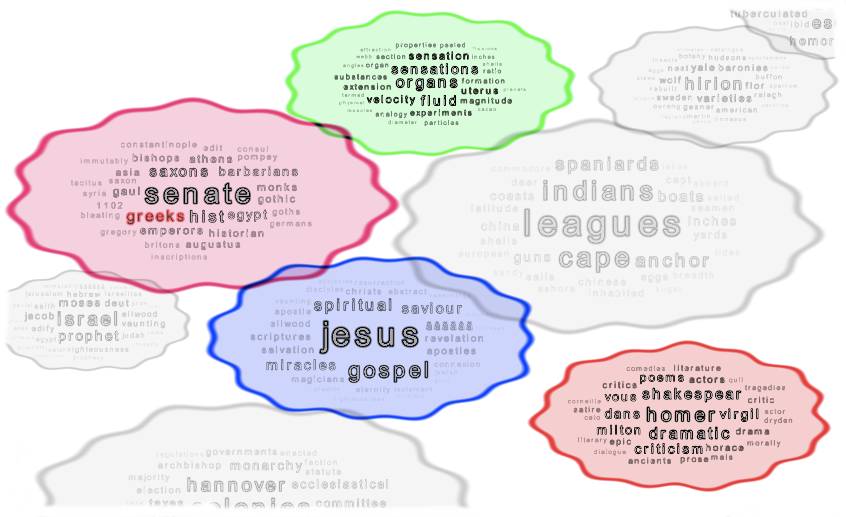


Figure 4. Solving task 4, finding documents related to topics. (left) Documents from the ECCO-TCP dataset that are related to the topic about ‘indians‘, ‘cape‘, ‘anchor‘, ‘spaniards‘, etc.; the documents are mainly travel reports and reports from English colonies. (right) Documents from the Stasi dataset related to the topic about problems in medical care in the former GDR. The documents are mainly statistical reports, one about the unions’ evaluation to the problem and the planned departure of a physician.

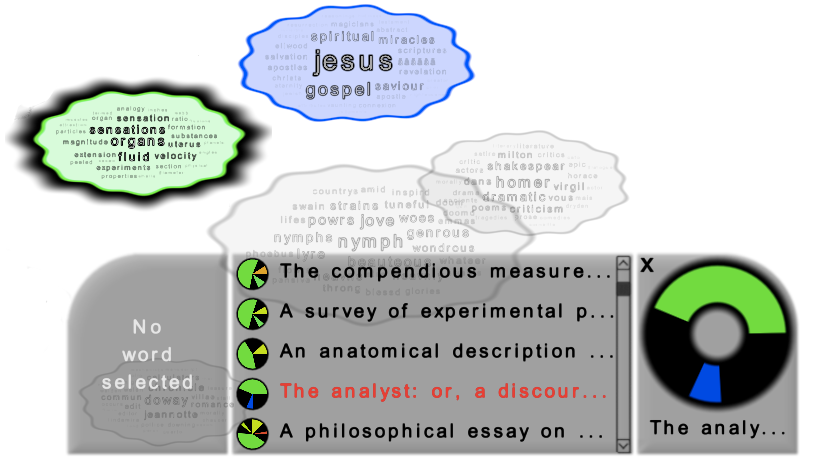


Figure 5. Solving task 5, finding topics related to a document. The selected document reads ‘The analyst: or, a discourse addressed to an infidel mathematician. Wherein it is examined whether the object, . . . and inferences of the modern analysis are more distinctly conceived, or more evidently deduced, than religious mysteries . . . By author of The minute philosopher‘ and has been found by first selecting the ‘science‘ topic (experiments, velocity, fluid, organs, magnitude, etc.). A second topic contained by this document has typical words like Jesus, saviour, spiritual, salvation, etc., as would be expected considering the document’s title.

# Notes

1. http://www.openwalnut.org.

2. Literally, Stasi abbreviates ‘STAatsSIcherheit‘, i.e. State Security.

3. The Eighteenth Century Collection Online Text Creation Partnership, available at http://quod.lib.umich.edu/e/ecco/.

# References

**Blei, D. M.** (2012). Topic Modeling and Digital Humanities. *Journal of Digital Humanities,* **2** (1).

**Blei, D. M. and Lafferty, J. D.** (2009). Topic Models. In Srivastava, A. N. and Sahami, M. (eds), *Text Mining: Classification, Clustering, and Applications*. CRC Press, p. 71.

**Chaney, A. and Blei, D. M.** (2012). Visualizing Topic Models. *Sixth International AAAI Conference on Weblogs and Social Media*.

**Jones, K. S.** (1972). A Statistical Interpretation of Term Specificity and Its Application in Retrieval. *Journal of Documentation,* **28**(1): 11–21.

**Kohonen, T.** (2001). *Self-Organizing Maps*. Springer Science & Business Media.

**Kruskal, J. B. and Wish, M.** (1978). *Multidimensional Scaling*. SAGE Publications.

**Marchionini, G.** (2006). Exploratory Search. *Communications of the ACM,* **49**(4): 41–46.

**Paulovich, F. V. and Minghim, R.** (2006). Text Map Explorer: A Tool to Create and Explore Document Maps. *Tenth International Conference on Information Visualisation (IV’06)*, pp. 245–51.

**Salton, G. and Buckley, C.** (1988). Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing & Management,* **24**(5): 513–23.

**Sammon, J. W.** (1969). A Nonlinear Mapping for Data Structure Analysis. *IEEE Transactions on Computers*.