DHQ: Digital Humanities Quarterly
2017
Volume 11 Number 2

Exploratory Search Through Visual Analysis of Topic Models

Patrick Jähnichen <jaehnichen_at_informatik_dot_uni-leipzig_dot_de>, Machine Learning Group, Humboldt-Universität zu Berlin
Patrick Oesterling <oesterling_at_informatik_dot_uni-leipzig_dot_de>, Image and Signal Processing Group, Leipzig University, Germany
Gerhard Heyer <heyer_at_informatik_dot_uni-leipzig_dot_de>, Natural Language Processing Group, Leipzig University, Germany
Tom Liebmann <liebmann_at_informatik_dot_uni-leipzig_dot_de>, Image and Signal Processing Group, Leipzig University, Germany
Gerik Scheuermann <scheuermann_at_informatik_dot_uni-leipzig_dot_de>, Image and Signal Processing Group, Leipzig University, Germany
Christoph Kuras <ckuras_at_informatik_dot_uni-leipzig_dot_de>, Natural Language Processing Group, Leipzig University, Germany

Abstract

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 to explore the corpus following an intuitive methodology: First, determine a topic of interest, second, suggest documents that contain the topic with "sufficient" proportion, 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.

Introduction

When dealing with large collections of digitized historical documents, very often only little is known about the quantity, coverage and relations of its content. In order to get an overview, an interactive way to explore the data is needed that goes beyond simple “lookup” approaches. 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 key-word 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 on the data, she then identifies topics of interest and finds the documents related to them. Additionally, these documents may 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 in which specific terms appear, i.e. dissipate semantic ambiguities that may appear. Especially when working with historical texts, this might help to reveal new aspects of known concepts. 1

Topic modeling [Blei 2009] has become one of the main tools to analyze text corpora in such a manner. We gain insight into a corpus' contents by identifying semantic classes of words, coined topic, opening up for exploratory search and analysis based on them. An excellent overview of how topic modeling can help humanists in their research is given in [Blei 2012]. 2

Topic modeling research, however, often focuses on the development of probabilistic models, i.e. incorporating a richer meta-data structure, increasing the speed of inference or using nonparametric models to circumvent model selection 3

problems. Comparatively little effort has been made to develop methods to use the outcome of these models in applications *visually*, although recently this task received growing attention (see section 3).

In this paper, we present a prototypical visual analysis tool to find and display the relations suggested by topic modeling. We derive distinct exploration tasks from the elements of a topic model, and present visual implementations for these tasks to provide the user with interactive means to browse through relations between documents, topics and words. In this way, the user uncovers expected or unexpected facts that eventually lead to interesting documents. More precisely, we represent topics by tag clouds of different size and, by considering pair-wise topic-similarities, we layout these clouds in the plane to provide the user with a topic-centered view on the data. 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.

The rest of the paper is organized as follows: in the next section we briefly discuss the underlying method, topic modeling. We then review related work in section 3 from both the language processing point view and from the direction of presenting topic models (and their alternatives) visually. In section 4, we define elementary exploration tasks applicable to the outcome of topic models, followed by descriptions of their visual implementation in our analysis tool in section 5. We report results from fitting topic models to two different data sets in section 6. We use Stasi records collected from the former East-German Ministry of State Security and the ECCO-TCP[1] data set, a set of classical literature texts, and conclude in section 7.

Topic Models

Topic models are a family of algorithms that decompose the content of large collections of documents into a set of topics and then represent each document as a mixture over these topics (based on the document's content). The outcome is thus a list of words for each topic (showing the probability of a term appearing in this topical context) and the proportion of topics for each of the documents. The key ingredients for finding this structure are word co-occurrences, words in a topic tend to co-occur across documents and hence are interpreted to share a common semantic concept (following the assumptions of distributional semantics (e.g., [de Saussure 2001]). On a more technical level, a topic model is a Bayesian hierarchical probabilistic (graphical) 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. The most general topic model is Latent Dirichlet Allocation (LDA) introduced by [Blei 2003] which is one probabilistic extension of the well-known Latent Semantic Analysis (LSA) technique (see [Landauer 2008]. As in LSA, it makes use of the bag-of-words assumption, in which the order of terms in text is neglected and their document-specific frequencies remain the sole basis of analysis. However, instead of factorizing the emerging document-term matrix algebraically, LDA defines a generative process that describes how documents are constructed. Here, latent variables control document generation: (1) the topics (as sets of word proportions), (2) the documents' topic proportions and (3) the probability of a certain term in a specific documents to belong to a specific topic. Using a Bayesian technique, prior distributions are placed on these factors. They interact as follows:

1. for all topics $k = 1,...,K$, draw topics $\beta_k \sim \mathrm{Dir}_V(\eta)$
2. for all documents $d = 1,...,D$
    - draw document d's topic proportion $\theta_d \sim \mathrm{Dir}_K(\alpha)$
    - for all words $n = 1,...,N_d$ in the document
        - draw the topic assignment $z_{dn} \sim \mathrm{Mult}(\theta_d)$
        - draw the word $w_{dn} \sim \mathrm{Mult}(\beta_{z_{dn}})$

where K is the number of topics, N_d are the document lengths, Dir(·) and Mult(·) respectively denote the Dirichlet and multinomial distribution (see [Kotz 2000]; [Johnson 1997] and η and α are so called hyperparameters (i.e. model parameters) to the Dirichlet distribution. A topic β_k 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 an individual topic sum to 1. The set of words

with highest probability is assumed to be different across different topics and to describe the individual topics thematically. Moreover, the assumption is that only a limited fraction of terms exhibit high probability in each topic. We can ensure this by appropriately setting the topic hyperparameter η. The document topic proportions θ_d are again probability distributions, defined over the topic simplex, i.e. every topic gets some probability in a document. Each document has its own topic proportions (hence the subscript d), the probabilities of topics for a single document also summing to 1. Again, we assume that only a small number of topics is active in each document and set the hyperparameter α accordingly. The words w_{dn} that we see in a document are now generated by first finding a topic z_{dn} through the document's topic proportions θ_d and then finding a word from the chosen topic $\beta_{z_{dn}}$. Both choices are random draws from their respective multinomial distributions. During inference, we seek to reverse this generative process in order to get approximations for the governing latent factors that best give rise to the observed words, i.e. we want to find the setting of the latent factors for which the observed words are highly likely. We end up with a suitable approximation for these factors that describes the generation of words assuming our generative model would be true. Note that we have skipped the technical details of how this approximation is achieved, the interested reader is referred to [Blei 2003] or [Heinrich 2005] for a more thorough technical description.

Visualizing the results of this model is one solution to unveil knowledge hidden in the data. However, the outcome (i.e. the topics and documents' topic proportions) 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. 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 key elements containing interesting insights about the data. 8

We emphasize that LDA is just one model that subsumes document collection content into topics. There exists a numerous amount of different topic models that can be used alternatively. Besides LDA, others take additional meta-data into account. These include e.g. the Author-Topic model [Rosen-Zvi 2005], inferring probability distributions over topics for each author instead of each document, the Hierarchical Dirichlet process topic model [Teh 2009, 887], a nonparametric model that uses Dirichlet Process priors instead of Dirichlet distributions and language models based on probabilistic nonnegative matrix factorization like the LDA interpretation of [Gopalan 2013], among others. Our visualization approach takes two matrices as input: a document-topic (in LDA the matrix formed by the individual θ_ds) and a topic-term matrix (formed by the β_ks). Every model that produces this output (or whose output can be transformed to these structures) is amenable to our visualization. This includes all of the above models (whereas in the Author-Topic model, authors would replace documents conceptually), in fact we could also visualize the outcome of the LSA model which follows completely different approach as topic modeling. 9

We also note that we do not go into the analysis of the models themselves but rather restrict our discussion to the outcome that they produce. Assessing the quality of the models' results is a research field on its own (e.g.,[Boyd-Graber 2009]; [Mimno 2011]) and we do not make any assumptions on its quality. [Gelman 2013] argue that for Bayesian methods in general, outcome must undergo thorough inspection and interpretation by domain experts. 10

Related Work

Traditional linguistic approaches such as the vector space model [Salton 1988] translate documents into high-dimensional feature vectors (typically in combination with, e.g., the tf-idf [Sparck Jones 1972] term weighting). Visualizing such high-dimensional structures is a research field on its own and several different methods have been proposed in the literature. Themeriver [Havre 2000] is a chart-based flow illustration that focuses on the change of themes over time based on word frequencies. The visualization has a strong focus on the themes and no support of 11

relations between topics, documents, and words. We currently ignore time information and consider the data set as static although there are plans to enhance our tools to be able to follow topic evolution through time (cf. section 7).

12

Closely related to our approach is that of [Chaney 2012], an attempt to directly visualize the output of topic models. The authors describe the main functionalities of such a system that are quite naturally similar to the definition of our exploration tasks in section 4. However, their approach includes generating a set of static websites that can be browsed to explore the data set, providing a lightweight, largely text-based application to solve the tasks. On the other hand, this method lacks user interaction and also is rather document-centered. 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. Numerous solutions extending this concept exist, e.g. [Snyder 2013] or [Hinneburg 2012], enriching or refining the resulting presentation with different kinds of metadata. [Cao 2010] propose a visualization technique for entities extracted from texts which they call FaceAtlas; a graph-based network visualization augmented with density maps to visually analyze text corpora with documents having relations based on different facets. This approach is similar to ours in that semantic similarity of entities determines spatial distance in the visualization. However, the method of how we arrive at our data model considerably differs. They use Named Entity Recognition (NER) to extract named entities from texts and visualize their relations whereas we extract latent structures (the topics) from the text that define distributions over the vocabulary. Relations between them are implicitly defined by measuring similarity of those distributions with suitable metrics (see e.g., [AlSumait 2009]; [Niekler 2012]). TopicNets [Gretarsson 2012] is a graph-based, interactive analysis tool that incorporates topic models into the mechanics of graph visualization and facilitates the collapsing of nodes based on semantic association, topic-based deformation of node sets, or real-time topic modeling on graph subsets at various levels. Topic Islands~[Miller 1998] uses stereoscopic depictions of topics using wavelets to describe thematic characteristics. ThemeScapes [Wise 1995] uses a terrain-like landscape metaphor to illustrate topics as hills with documents on top. Less complex linguistic approaches translate documents into high-dimensional feature vectors using the vector space model [Salton 1988] in combination with, e.g., the tf-idf [Sparck Jones 1972] term weighting. In this space, words serve as dimensions and documents are finally represented as a point cloud; with (sub-)clusters of documents for each (sub-)topic. Finding and visualizing this high-dimensional structure is a research field on its own. Established approaches include projective techniques, like the Text Map Explorer [Paulovich 2006] Multidimensional Scaling (MDS) [Kruskal 2009], e.g. Sammon's mapping [Sammon 1969]; neuro-computational algorithms like Kohonen's Self Organizing Maps (SOM) [Kohonen 2001], scatterplot matrices [Elmqvist 2008] or topological analysis based on density functions [Oesterling 2010]. However, compared to the modeling approach used in this paper, the insights obtained from the vector space model is rather limited. Our work differs in that we focus on the visual representation of topic model elements to provide more thorough and quicker visual access to the data. We use a more flexible depiction of probability distributions as tag clouds and illustrate topics in an overview image to permit the identification of related topic groups or outliers. Furthermore, we extend the analysis to word-based tasks like finding polysemous and homonymic relations, we use smooth level-of-detail instead of using a fixed number of keywords, and we allow the user to quantify the relative impact of related topics or documents. We also note that there exist a wide variety of different topic models. For example models that impose a network structure on document during model design provide the possibility to interpret the links between documents on a semantic level ([Chang 2010]; [Mei 2008]). However, we want to keep our tool applicable to the widest range of data possible and thus neglect models that make use of other meta data then word frequencies. Further, we do not aim at visualizing links between documents but rather links between topics (which are given by distributional similarity, see below).

Exploration Tasks

13

Using the aforementioned outcome of topic models, we aim to provide the user with exploratory means to analyze a corpus by creating a largely topic-centered view on the data and letting latent topics act as the user's main interface to the documents. In this section, we recall the elements of a topic model and classify them into exploration tasks to relate topics to words and documents, and vice versa. The analysis process then consists of concatenating elementary exploration tasks. That is, motivated by intermediate insights about expected or unexpected relationships, the user interactively browses through the data via linked tasks.

Definition of Exploration Tasks

The probability distributions resulting from a topic model relate the whole vocabulary to latent topics, and the latter to all documents. That is, the outcome of this model is very complex in that all words occur in every topic, and all topics appear in every document – both with certain "significance" (in fact probability). Because such complex data is hard to handle as a whole, we split the analysis process into distinct exploration tasks to reveal possible relations between single conceptual entities, e.g. documents, and one or more topics, between topics related to a single documents, or between topics related to single words. Based on a simple input-output scheme, every task requires certain information produced by the topic model or provided by the input data and it discloses potentially existent relationships between them. 14

Exploration Task 1 - Examining a Topic

Examining a single topic is difficult because it is a probability distribution over potentially thousands of words in the vocabulary. Technically, this task involves the following information: the topic's overall significance in the corpus, a meaningful sorting of the words for appropriate topic description, and actual word significances to provide pivotal keywords and their relative importance. The quantity of a topic's overall significance can easily be computed as a relative measure using the model outcome of topic models: topic-significance$_k$ = $(\sum_d \theta_{dk} \cdot N_d)/(\sum_d N_d)$, where θ_{dk} is the d-th document's topic proportion of topic k, satisfying $\theta_{dk} \leq 0$, k=1,...,K and $\sum_k \theta_{dk} = 1$. N_d is the length of document $d. Relevance determination of a topic's words involves finding a suitable sorting because both tasks are based on the word probabilities provided by the topic model. Since we can assume that the largest part of the vocabulary does not carry topic-specific information, it is reasonable to sort the words by decreasing probability and increasingly disregard their relevance for that topic. Another approach involves determining the words' relevances using a tf-idf [Sparck Jones 1972] flavored procedure; each topic is interpreted as a document and word probabilities are treated as scaled document frequencies. Using a basic tf-idf scheme, we could easily identify words that are highly descriptive exclusively to their respective topics. In the visualization, this information is used to help the user to quickly identify the key words and their relative importance for a topic. 15

Exploration Task 2 - Overview over the Topics

The second exploration task is to summarize the set of latent topics found by the topic model. This includes the following information: the number of topics, their overall significance for (or impact on) the corpus, and similarities between topics defined by some measure. While the overall significance is equal to that in Exploration Task 1, for topic similarity, different metrics are possible. One score that is motivated by similar topic distributions was described by [Chaney 2012]. Another approach is to understand topics as vectors in a high-dimensional space where words of the vocabulary serve as dimensions. Then topic similarities are described by proximities in the vector distribution on the surface of a unit d-simplex. Possible metrics are the Jensen-Shannon divergence [Manning 1999] or distance-based measures like, e.g., simple Euclidean distance. Using any appropriate metric, topic (dis)similarity shall be described by spatial proximity in the overview. 16

Exploration Task 3 - Finding Different Polysemous and Homonymic Semantics of Terms

One advantage of topic models is the automatic disambiguation of semantic meanings of words into topics. A word with different meanings[2] automatically appears in different topics that correspond to its different semantic contexts. Note that each word has some probability in every topic. Instead of introducing a significance threshold, i.e. a hard threshold in probability, we resort to the idea that if the significance in another topic falls below a certain level that is not visually presentable anymore, the relation is of no importance. This corresponds to defining an implicit threshold in probability through visual limitations. To highlight the occurrence of polysemous and homonymic terms including their relative importance in other topics, this task consists of the following requirements: selecting a word of interest, the quantification of relevant topics that share this word with “sufficient” probability, and the relevance of the selected word in these topics to evaluate semantic diversity. In the visualization, the user should be able to quickly deduce potential semantical ambiguity by selecting a word in any topic and seeing the impact of this word in other topics provided by the 17

topic model.

Exploration Task 4 - Identifying Documents Covering a Topic

Having identified one or more interesting topics $\mathscr{K} = \{k : k \in \{1,...,K\}\}$, the user may want to look at documents that cover these topics. This task is at the core of exploratory analysis. The information required for this are the topics of interest k_i and a list of documents sorted in decreasing order by the combined impact of topics $\mathscr{K}$ on the documents. Given $\mathscr{K}$, we can easily read off the probability of these topics in all of the documents $\theta_{\cdot k}$, $k \in \mathscr{K}$. After sorting the documents by their proportions of the impact product of all topics $\mathscr{K}$ (we use $\Pi_{k \in \mathscr{K} k}\, p(\theta_{d\hat{k}} | \hat{k} = k)$ to approximate the combined impact of $\mathscr{K}$ on each document d), we obtain a list of documents that exhibit the topics of interest with decreasing significance. 18

Exploration Task 5 - Finding Related Topics of a Document

Once an interesting document has been identified, the user may want to inspect other topics related to it, or, in a transitive way, documents related to these other topics. Again, this task aims at giving the user a tool for exploring related documents (and thus the corpus) through picking interesting topics. The following information are involved in this task: a document of interest, the proportions of other topics in this document, and document-related documents. While the related topics of a document simply result from the topic model, the latter information could be obtained from considering the similarity between topic distributions between two documents (an example metric is given in [Chaney 2012]). Because we focus on a topic-centered navigation through the data, we drop document-document relationships and primarily focus on a document's topic distribution. 19

Visualization Approach

In this section, we explain our analysis tool and provide visual implementations for each exploration task defined in section 4. Furthermore, we describe interactive means to navigate through the data by letting the user concatenate individual tasks, stimulated by the feedback of previous insights and following an intuitive analysis methodology: overview first, details on demand. 20

Visual Implementation of Exploration Tasks

The visualization of a topic model should provide quick visual access to the key features and relations in the data. This task is difficult because the broad and complex probability distributions produced by a topic model contain large amounts of irrelevant information. That is, only a minor part of the vocabulary is meaningful to describe a topic, and only some topics have considerable impact on a certain document. Hence, visualizing the topic model as a whole rapidly creates cluttered visualizations. We pursue a visualization approach that illustrates the crucial information of individual exploration tasks, but that also allows the user to refine the level-of-detail in an intuitive and interactive way. 21

Visual Implementation of Exploration Task 1 - Examining a Topic

To visualize a single topic we make use of *tag clouds* [Steele 2010], a popular visual metaphor for weighted word-lists. Although tag cloud implementations can be highly sophisticated, we keep it simple and only focus on the information required for the exploration task. Taking the sorted words, we create labels with size and opacity proportional to the words' probabilities and arrange them in a spiral layout around the most significant word. That is, for each word, we start a spiral from a center point until sufficient space is found to place this word. As a consequence, small and increasingly transparent irrelevant words are positioned at the cloud's border or in the gaps between relevant words. The user can smoothly change the level-of-detail by zooming in and out to make small words appear and to adjust a word's readability (size and opacity) proportional to the zoom-factor. The minimum level-of-detail shows at least the top keyword per topic at full opacity. Furthermore, the tag cloud's extent is scaled by the topic's overall significance in the corpus and each cloud is assigned a distinguishable color to ease further analysis. Example topic clouds are shown in Figure 2. 22

Visual Implementation of Exploration Task 2 - Overview over the Topics

The topic overview visualization is the main view on the data and the starting point of any other exploration task. To visualize the required information for this task, we layout tag clouds in the plane to present them on the screen. Their number reflects the number of latent topics found by the topic model and their pair-wise distances in the layout approximate their similarities; understood either as the difference between probability distributions or distances in the high-dimensional *topic space* on the surface of the unit d-simplex. There are plenty of algorithms to create a layout of the clouds that reflects pair-wise (dis)similarities as distances in the plane. For the sake of simplicity, we use either a force-directed approach or Sammon's mapping. Although the probability distributions of the topics are assumed to be sufficiently diverse to minimize cloud occlusions, the user can also scale all pair-wise cloud distances to disperse accumulations. In the overview, the user can quickly distinguish cloud sizes and identify related topics as nearby clouds or cloud accumulation. 23

Visual Implementation of Exploration Task 3 - Finding Different Polysemous and Homonymic Terms

To identify polysemous terms and homonyms, the selected word of interest (selection mechanisms will be explained in section 5.2) is highlighted in every topic cloud which provides quick access to this word's significances in other topics via their labels' sizes. Moreover, clouds corresponding to topics in which the selected word is considered insignificant are decolored, i.e. bleached out to facilitate focusing only on those topics with relevant word probability. Because a word's size in another topic could be marginal relative to all other label sizes, or because the current zoom-level is not high enough to identify the word of interest in every cloud, we also provide a chart in a head-up display (HUD) to denote the proportions of this word's probability in each topic. Every part of the chart is colored according to the topic it represents. By inspecting the highlighted labels' size or their corresponding parts in the chart, the user can quickly judge the diversity and partial quantities of a word's different meanings. The topic chart is also a starting point for topic-related exploration tasks (cf. section 5.2). 24

Visual Implementation of Exploration Task 4 - Documents Covering a Topic

Given one or more selected topics of interest, the sorted list of documents covering these topics is presented in the head-up display. Each of the scrollable list's entries shows the document's name. The list is also a starting point for document-related exploration tasks (cf. section 5.2). 25

Visual Implementation of Exploration Task 5 - Finding Related Topics of a Document

Each of the document list's entries additionally shows a small chart of the document's overall topic distribution. From these charts, the user can directly read off and compare the impact of the selected topics on every document; also in contrast to all other topics. Selecting a document activates a magnified version of the topic chart to the right of the list in the HUD and serves as a source for topic-related exploration tasks, like examining its words or updating the document list. 26

Figure 1. Overview over our analysis tool. The visual context of the topic space is always preserved in the background, moving the camera if necessary to center selected topics currently not visible on the screen. Widgets in the head-up display (HUD) summarize relations between user-selected elements. In this example scenario, the user first selects the word “newcastle” in the top green cloud. By doing so, this word gets highlighted (red) in any other cloud and the probabilities of this word in other topics are additionally shown in a pie chart (bottom left). Clouds with insignificant parts in this chart are decolored. Selecting the other topics in which “newcastle” appears (black border) we can create the document list (sorted by the impact of these topics) in the middle of the HUD. Selecting a single document (red line) creates a chart (bottom right) with this document's topic probabilities, revealing other significant topics (colored) in the topic overview that appear in the document.

User Interaction Mechanisms

The user browses interactively through the data by concatenating exploration tasks. The visual implementations of these tasks can be linked in order to launch subsequent tasks based on intermediate insights about the data (cf. Figure 1 for the different visualization components). 27

Selecting a Topic

Topics can be selected in two ways: by right-clicking one or more clouds in the overview visualization, or by selecting the corresponding part of a topic chart (see the next to selection mechanisms). A selection of one or more topics triggers the following actions: an accentuation of the corresponding clouds with an additional border to highlight selected topics, and an update of the document list in the HUD to present those documents that share the selected topics, sorted by decreasing combined impact. In addition, if a topic is currently not visible in the cloud overview, a camera movement centers and magnifies it on the screen so that representative words can be read off to get a quick understanding of the topic. Note that this camera movement is actually also part of exploration task 1 because being able to read words associated to a topic is an inherent part of examining it. 28

Selecting a Word

As part of exploration task 3, words are selected by left-clicking them in any of the topic clouds. A word selection triggers three actions: the word's accentuation in every other cloud, the decoloring of those clouds that feature only insignificant probabilities for this word, and the creation of a pie chart showing an aggregation of the word's significance in different topics in the HUD, creating starting point for exploration task~4 or exploration task 1. 29

Selecting a Document

Documents are selected by clicking on them in the HUD's document list. A document selection triggers the creation of the topic distribution chart for this document, which is placed to the right of the document and is used to trigger the topic-based exploration tasks 1 and 4. Further, topics not relevant for this document are decolored in the cloud overview to quickly identify those that are. 30

Cyclic Analysis Process and Visual Context of Topics

Using these interaction mechanisms, the analysis process is carried out by combining exploration tasks in a transitive or cyclic way. That is, the selection of topics, words, or documents highlights other visual entities and updates widgets which triggers the next exploration task. For example, clicking on a word in one of the clouds (task 1 and 2) creates a topic chart (task 3) in which click-events create the document list for certain selected topics (task 4). Clicking on a document creates a chart for related topics (task 5) whose selection centers a topic cloud (task 1) and updates the document list (task 4)-and so on. 31

Note that camera movements related to topic selections constantly preserve the visual context in the topic space. That is, for selected topics, the user can always read off keywords to evaluate their meaning and importance and related topics and their overall significance can be identified by examining nearby clouds. By bleaching out topics that do not significantly contribute to one of the topic charts, the user can quickly identify the spatial relation between selected (colored) topics in the overview. This can help to reveal interesting words appearing in deselected topics. Note that we understand our framework as a tool to navigate through the data based on relations between topics and both words and documents. Once interesting documents are identified, their content is presented to the user in a linked view or in the head-up display. 32

Experiments

Data Sets

We report use cases of fitting topic models to two different data sets. The first 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*”) [Münkel 2014] that reveal the Stasi's[3] specific view of the GDR, containing references to real and perceived oppositional conduct as well as to economic and supply problems. The second data set is the ECCO-TCP[4], a set of classical literature and non-fiction texts. We preprocessed the raw text and performed a set of standard preprocessing steps (including stop word removal, minimum frequency pruning and lower casing). In the following we walk through the exploration tasks as described in section 4 and provide screen shots of the respective visual implementations, including how the spatial proximity of topics helps to identify semantic similarities between them. We fitted non-parametric topic models [Teh 2009, 887] on each of the two data sets using the Gibbs sampling approach described there. Our visualization prototype is based on the OpenWalnut visualization engine[5] and is implemented as a plugin for this framework. Our current workflow is pretty rudimentary: given a collection of documents, we learn a topic model. We then take the resulting document-topic and topic-term probability matrices and feed them into our prototype. OpenWalnut runs on a wide variety of platforms, we have used a ready-to-use Linux distribution[6] providing installation packages for OpenWalnut and added our plugin. Experiments could smoothly be carried out using virtualBox[7] on an Apple MacBook Air laptop. 33

Examining a topic

Figures 2 and 3 show two topics extracted from the data sources. Words' sizes are determined by their probability in the topic's distribution over the vocabulary. The topics can easily be identified to circle around the literary genre of drama and communist-party propaganda concerning the youth respectively. As the user zooms in, more words become visible that were hidden because of lesser relevance to the topic, uncovering them reveals a semantic refinement of the topic. This shows that not only the most significant terms define a topic, they merely allude to the topic's semantic meaning 34

that is subsequently defined by the other words with considerable probability mass in it.

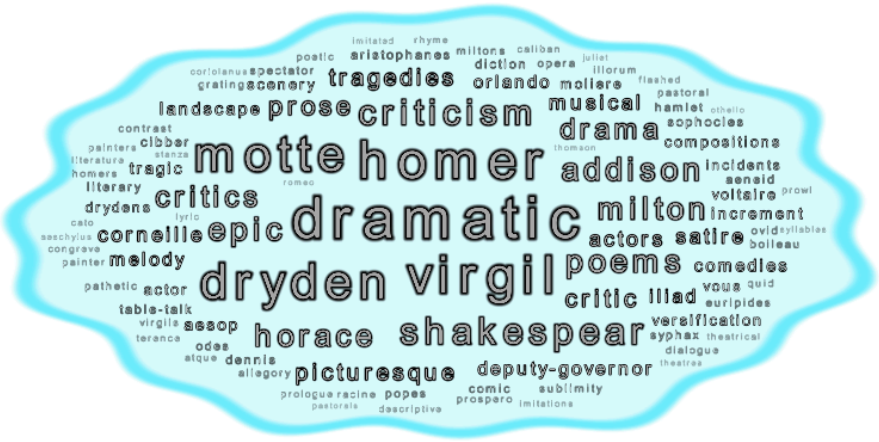

Figure 2. A topic covering several classical drama authors and their work from the ECCO-TCP data set.

Figure 3. A topic covering communist party propaganda in connection with the youth from the Stasi files data set.

Overview of the Topics

Figures 4 and 5 show an overview over the topics found. As described above, the size of the tag clouds represents the overall topic-significance in the whole data set and spatial proximity indicates closer semantic relatedness of topics. One example (cf. Figure 4) is the cluster of topics that cover different aspects of religion and its role in colonization and history. To determine the differences between them the user can follow the methodology for examining a topic. The objective of this view is to provide an initial starting point for further analysis and to draw the user's attention to interesting parts of the data. Figure 5 shows a cluster of topics that are concerned with taking appropriate measure towards different problems in economy and society in GDR. The user gains a first insight into the corpus and is motivated to continue her exploration of the data by concatenating further tasks.

Figure 4. Example topics from the ECCO-TCP data set.

Figure 5. Examples from the Stasi files data set.

Finding different semantics

To disambiguate the semantics of a given term, the user selects the word of interest leading to other topics that exhibit the word with sufficient significance being highlighted. In Figure 6 the selected word in the visualization of the ECCO-TCP corpus is “greeks”. Other topics that include this term with sufficient probability are the topics about medicinal findings about organs, Jesus and religion, and drama. Clearly, “greeks” appears in different semantical contexts here: it relates to the Greek political system and the ancient Greek society in that it had a modern understanding of medicine (for their time) and stands for famous Greek drama authors (most prominently Homer). Figure 7 shows the different semantics of the word "untersuchungen" (investigations) from the Stasi corpus. The term is used in connection with the ever suspicious Stasi (topics dealing with addresses, name, Berlin, “*wohnhaft*” (lives at) etc. but also with investigations of accidents in factories (“*veb*” meaning Peoples-owned enterprise) and the state railroad (“*reichsbahn*”).

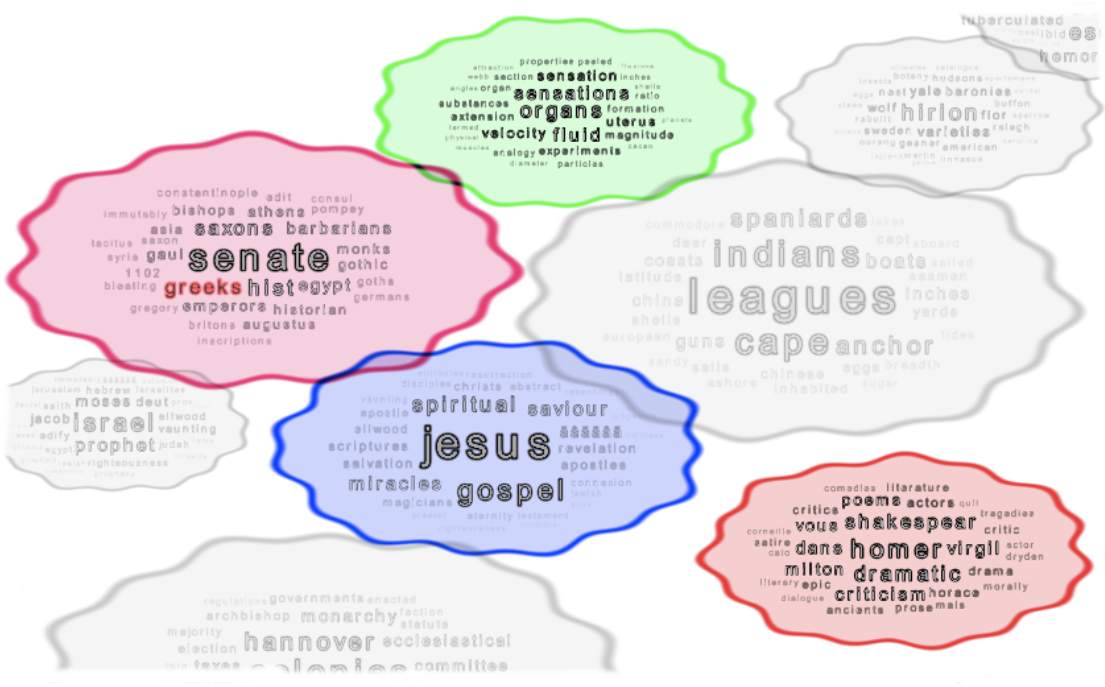

Figure 6. Different usages of the term “greeks” in the ECCO-TCP data set. Usages in connection with religion, literature, science and politics are apparent.

Figure 7. Usages of the term “*untersuchungen*” (Engl. investigations) from the Stasi files data set.

Identifying documents covering a topic

Akin to example 7, assume that the user is interested in documents that include the topic described by terms like “indian”, “cape”, “spaniards”, “china” or “anchor”. Selecting the topic creates a list of documents that cover it (in this case travel narratives and reports from the English colonies) in the middle of the HUD as shown in Figure 8. The user can

scroll through this list and finish one goal of exploratory analysis. She has identified a topic of interest, examined it to identify its semantic nature and found a list of documents that cover it. Figure 9 solves this task on the Stasi data set. The selected topic is about problems in medical care in former GDR (indicated by the terms "*medizinischen*" (medical), "*versorgung*" (care) and "*probleme*" (problems). The resulting list includes mainly statistical reports but also documents about the unions' evaluation to the problem and even about the planned departure of a physician. This allows for a content driven access to the data. It is also possible to combine different topics so that we can display documents that cover a combination of topics (see Figure 1 for one such example).

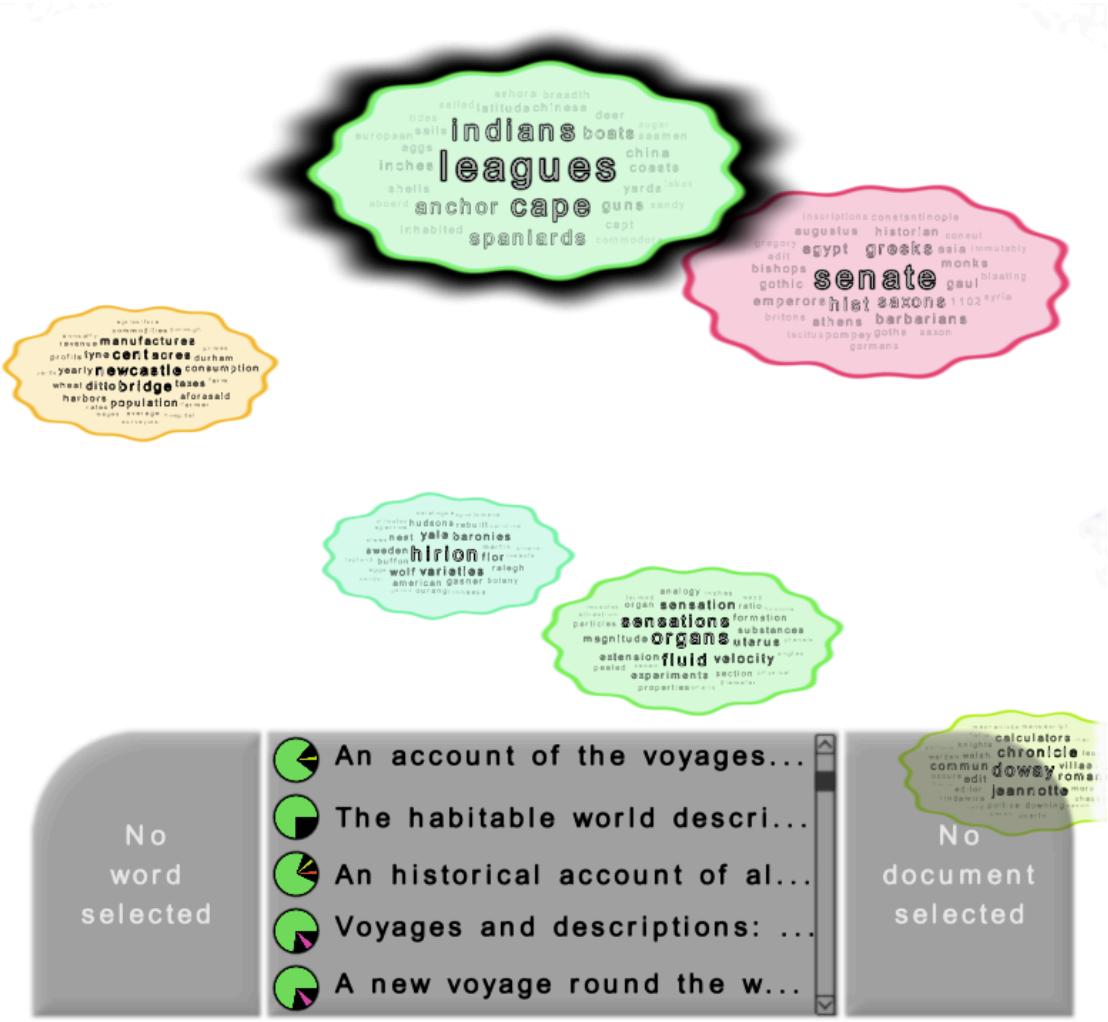

Figure 8. Documents from the ECCO-TCP data set that are related to the topic about "indians", "cape", "anchor", "spaniards" etc., the documents are mainly travel narratives and reports from English colonies.

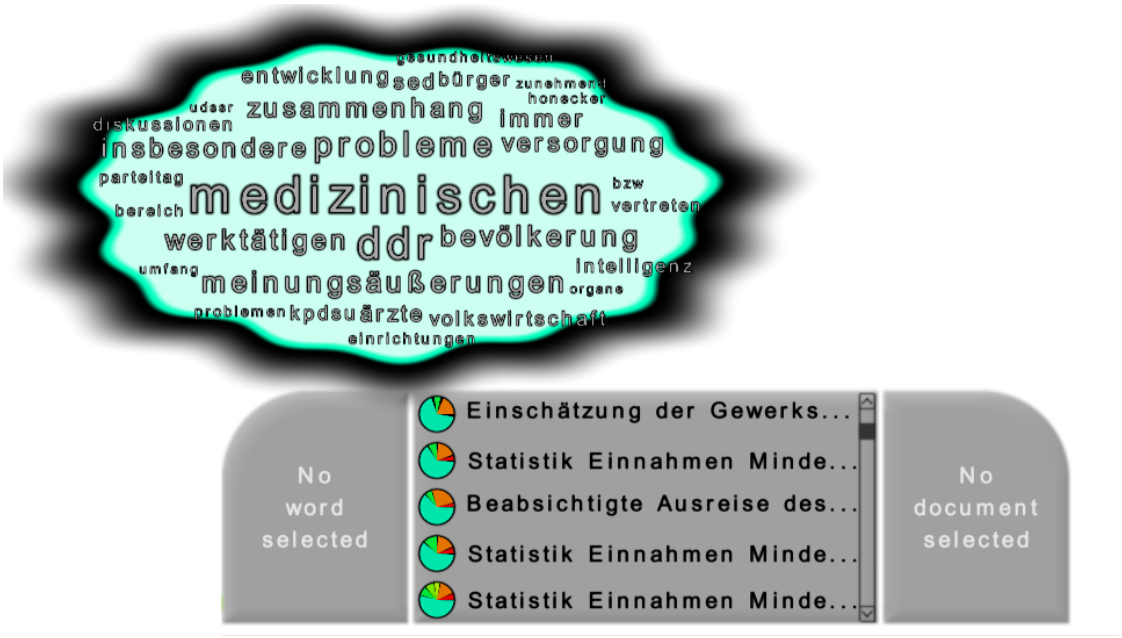

Figure 9. Documents from the Stasi data set related to the topic about problems in medical care in former GDR. The documents are mainly statistical reports, one about the unions' evaluation to the problem and the planned departure of a physician.

Finding related topics of a document

While exploring the data set and reading documents that cover a topic of interest, it is often the case that this topic is not the only one covered by a document. By displaying a chart of the portions of other topics covered by this document as in Figure 10 the user is encouraged to continue her exploration by examining these other topics. In our example the user first selected a topic about anatomy (with terms like “organs”, “fluids”, “sensations”, “uterus” etc.) and then a document covering this topic. We find a connection to another topic about religion (“jesus”, “gospel”, “saviour” etc.). Indeed the selected document's title fully 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”. (In this case the connection is of course not that surprising given the title.)

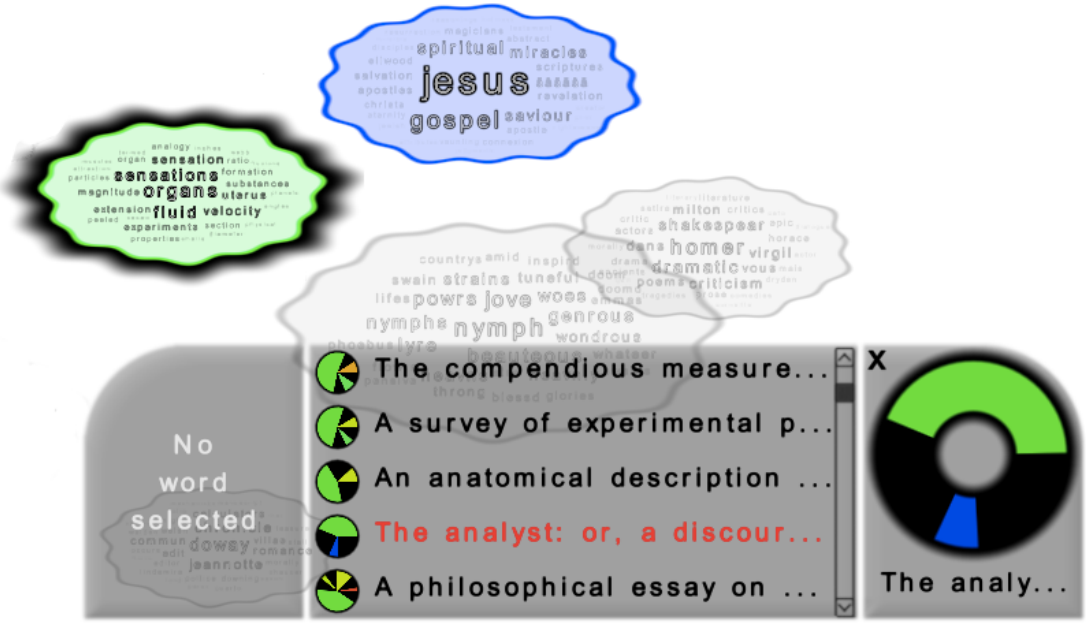

Figure 10. Solving task 5, finding topics related to a document. The selected document has been found by first selecting the anatomy topic (experiments, velocity, uid, organs, magnitude etc.). A second topic contained by this document has typical words like “Jesus”, “savior”, “spiritual”, “salvation” etc. as would be expected considering the document's title.

Discussion and Future Work

In this paper we have described a visual tool using a tag clouds based approach to visualize the outcome of topic models. Showing series of word probabilities as tag clouds easily provides quick visual access to a topic's meaningful keywords including their significance and qualitative difference to other words in the topic. That is, compared to a simple list of sorted words, the user can quickly judge topical distinctness by the ratio between words of high and low probability, and pivotal keywords also stand out visually in the clouds. Furthermore, by zooming in and out to change the level-of-detail, the user can quickly adjust a topic's expressiveness in terms of its keywords; while still minimizing unnecessary information by keeping the remaining words small and translucent. We also advanced [Chaney 2012]'s topic model visualization by introducing additional word-based exploration tasks, providing the user with additional information (absolute values and proportions) to rank related topics and documents, and by the usage of a topic clouds layout to identify related topic groups or outliers easily in a global overview. Camera movements in the cloud space triggered by topic selections also preserve the context of the topic-centered analysis process. 39

We understand our tool as a topic-centered navigator to visually disclose and present structure hidden in the outcome of the topic model. That is, reading documents and other document-related exploration tasks are currently not considered in our tool. We leave the investigation of these tasks and their visual implementations, and also the expansion of the visual analysis to time-dependent data for future work. We also omitted further possible improvements in the preprocessing of data, i.e. before learning a topic model. These may include stemming, lemmatization and restricting the vocabulary by confining to certain parts-of-speech. Also, as of now, we are not able to export findings from our approach and are restricted to identifying document titles. However, plans to incorporate our visualization into a larger NLP-toolbox (the Leipzig Corpus Miner[8] currently under development) would readily solve both these shortcomings. Ultimately, we hope to use the application to restrict an existing corpus to the set of documents the user is interested in, enabling her to perform further analysis steps on the now semantically localized subcorpus. 40

Notes

[1] The Eighteenth Century Collection Online Text Creation Partnership.

[2] A polysemous word has multiple related meanings, e.g. bank as the financial institution and as the building in which the institution resides. A

homonym has different meanings that are unrelated, e.g. bank as the financial institution and the river bank.

[3] Literally, Stasi abbreviates "*STAatsSIcherheit*", i.e. State Security.

[4] see http://www.textcreationpartnership.org/tcp-ecco/

[5] http://www.openwalnut.org

[6] http://neuro.debian.net/

[7] https://www.virtualbox.org/

[8] http://www.epol-projekt.de/tools-nlp/leipzig-corpus-miner-lcm/

Works Cited

AlSumait 2009 AlSumait, L., Barbará, D., Gentle, J., Domeniconi, C. "Topic significance ranking of LDA generative models", *Machine Learning and Knowledge Discovery in Database*s (2009): 67–82.

Blei 2003 Blei, D. M., Ng, A. Y., Jordan, M. I. "Latent dirichlet allocation", *The Journal of Machine Learning Research*, 3 (2003):993–1022.

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

Blei 2012 Blei, D. M. "Topic Modeling and Digital Humanities", *Journal of Digital Humanities*, 2.1, January 2012.

Boyd-Graber 2009 Boyd-Graber, J., Chang, J., Gerrish, S., Wang, C., Blei, D. M. "Reading tea leaves: How humans interpret topic models", *Advances in Neural Information Processing Systems*, 31 (2009).

Cao 2010 Cao, N., Sun, J., Lin, Y. R., Gotz, D., Liu, S., Qu, H. "FacetAtlas: Multifaceted Visualization for Rich Text Corpora", *IEEE Transactions on Visualization and Computer Graphics*, 16.6 (2010): 1172–1181.

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

Chang 2010 Chang, J., Blei, D. M. "Hierarchical relational models for document networks", *The Annals of Applied Statistics*, 4.1 (2010): 124–150.

Elmqvist 2008 Elmqvist, N., Dragicevic, P., Fekete, J. D. "Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation", *IEEE Transactions on Visualization and Computer Graphics*, 14.6 (2008): 1539–1148.

Gelman 2013 Gelman, A., Shalizi, C. "Philosophy and the practice of Bayesian statistics", *British Journal of Mathematical and Statistical Psychology*, 66.1 (2013): 8–38.

Gopalan 2013 Gopalan, P., Hofman, J. M., Blei, D. M. :Scalable Recommendation with Poisson Factorization", *arXiv.org* (2013).

Gretarsson 2012 Gretarsson, B., O'donovan, J., Bostandjiev, S., Höllerer, T., Asuncion, A., Newman, D., Smyth, P. "Topicnets: Visual analysis of large text corpora with topic modeling", *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3.2 (2012): 23.

Havre 2000 Havre, S., Hetzler, B., Nowell, L. *ThemeRiver: Visualizing theme changes over time* (2000): 115–123.

Heinrich 2005 Heinrich, G. *Parameter estimation for text analysis* (2005). http://www.arbylon.net/publications/textest

Hinneburg 2012 Hinneburg, A., Preiss, R., Schröder, R. *TopicExplorer: Exploring document collections with topic models* (2012): 838–841.

Johnson 1997 Johnson, N. L., Kotz, S., Balakrishnan, N. *Discrete Multivariate Distributions*. John Wiley & Sons (1997).

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

Kotz 2000 Kotz, S., Balakrishnan, N., Lloyd Johnson, N. *Continuous Multivariate Distributions: Models and Applications*, 1. John Wiley & Sons, 2nd edition (2000).

Kruskal 2009 Kruskal, J. B., Wish, M. *Multidimensional Scaling. Quantitative Applications in the Social Sciences*. SAGE

Publications (1978).

Landauer 2008 Landauer, T., Dumais, S. "Latent semantic analysis", *Scholarpedia*, 3.11 (2008): 4356.

Manning 1999 Manning, C. D., Schütze, H. *Foundations of statistical natural language processing*. MIT Press (1999).

Marchionini 2006 Marchionini, G. "Exploratory Search", *Communications of the ACM*, 49.4 (2006): 41–46.

Mei 2008 Mei, Q., Cai, D., Zhang, D., Zhai, C. X. "Topic modeling with network regularization." In *17th international conference on World Wide Web* (2008): 101–110.

Miller 1998 Miller, N. E., Wong, P. C., Brewster, M., Foote, H. *TOPIC ISLANDS TM-a wavelet-based text visualization system* (1998): 189–196.

Mimno 2011 Mimno, D., Blei, D. M. "Bayesian checking for topic models." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (2011): 227–237.

Münkel 2014 Münkel, D. *Die DDR im Blick der Stasi. Für den Bundesbeauftragten für die Unterlagen des Staatssicherheitsdienstes der ehemaligen Deutschen Demokratischen Republik* (BStU) (2014).

Niekler 2012 Niekler, A., Jähnichen, P. *Matching Results of Latent Dirichlet Allocation for Text* (2012): 317–322.

Oesterling 2010 Oesterling, P., Scheuermann, G., Teresniak, S., Heyer, G., Koch, S., Ertl, T., Weber, G. H. "Two-stage framework for a topology-based projection and visualization of classified document collections", *Transactions of the IRE Professional Group on Audio*, (2010): 91–98.

Paulovich 2006 Paulovich, F. V., Minghim, R. "Text Map Explorer: a Tool to Create and Explore Document Maps", *Tenth International Conference on Information Visualisation (IV'06)*, (2006): 245–251.

Rosen-Zvi 2005 Rosen-Zvi, M., Griffiths, T. L., Steyvers, M. "Learning Author Topic Models from Text Corpora", *The Journal of Machine Learning Research* (2005).

Salton 1988 Salton, G., Buckley, C. "Term-Weighting Approaches in Automatic Text Retrieval", *Information processing & management*, 24.5 (1988): 513–523.

Sammon 1969 Sammon, J. W. "A nonlinear mapping for data structure analysis", *IEEE Transactions on Computers* (1969).

Snyder 2013 Snyder, J., Knowles, R., Dredze, M., Gormley, M. R., Wolfe, T. "Topic Models and Metadata for Visualizing Text Corpora." In *HLT- NAACL* (2013): 5–9.

Sparck Jones 1972 Sparck Jones, K. "A statistical interpretation of term specificity and its application in retrieval", *Journal of documentation*, 28.1 (1972): 11–21.

Steele 2010 Steele, J., Iliinsky, N. *Beautiful Visualization. Looking at Data through the Eyes of Experts*. O'Reilly Media, Inc. (2010).

Teh 2009 Teh, Y. W., Jordan M. I. "Hierarchical Bayesian nonparametric models with applications", *Bayesian Nonparametrics* (2009): 158.

Wise 1995 Wise, J. A., Thomas, J. J., Pennock, K., Lantrip, D., Pottier, M., Schur, A., Crow, V. *Visualizing the non-visual: spatial analysis and interaction with information from text documents* (1995): 51-58.

de Saussure 2001 de Saussure, F. *Grundfragen der allgemeinen Sprachwissenschaft*. de Gruyter, 3rd edition (2001).