Reading Thomas Jefferson with TopicViz: Towards a Thematic Method for Exploring Large Cultural Archives

Lauren Klein & Jacob Eisenstein
Georgia Institute of Technology

Scholarly and Research Communication
VOLUME 4 / ISSUE 3 / 2013

Abstract

In spite of what Ed Folsom has called the "epic transformation of archives," referring to the shift from print to digital archival form, methods for exploring these digitized collections remain underdeveloped. One method prompted by digitization is the application of automated text mining techniques such as "topic modelling"—a computational method for identifying recurring themes across an archive of documents. We review the nascent literature on topic modelling of literary archives and present a case study, applying a topic model to the Papers of Thomas Jefferson. The lessons from this work suggest that the way forward is to provide scholars with more holistic support for the visualization and exploration of topic model outputs, while integrating topic models with more traditional workflows oriented around assembling and refining sets of relevant documents. We describe our ongoing effort to develop a novel software system that implements these ideas.

Keywords

Archives; Reading; Research; Topic modelling; Search; Visualization; Thomas Jefferson

Lauren Klein is Assistant Professor in the School of Literature, Media, and Communication at the Georgia Institute of Technology, 686 Cherry Street, Atlanta, GA, 30332. Email: lauren.klein@lmc.gatech.edu .

Jacob Eisenstein is Assistant Professor in the School of Interactive Computing at the Georgia Institute of Technology, 85 Fifth Street NW, Atlanta, GA, 30308. Email: jacobe@gatech.edu .

Introduction

In spite of what Ed Folsom has called the "epic transformation of archives," (2007, p. 1571) referring to the shift from print to digital archival form, methods for exploring these digitized archival collections remain underdeveloped. At present, in order to

CCSP Press
Scholarly and Research Communication
Volume 4, Issue 3, Article ID 0301121, 12 pages
Journal URL: www.src-online.ca
Received March 13, 2013, Accepted March 17, 2013, Published December 16, 2013

Klein, Lauren, & Eisenstein, Jacob. (2013). Reading Thomas Jefferson with TopicViz: Towards a Thematic Method for Exploring Large Cultural Archives. *Scholarly and Research Communication, 4*(3): 0301121, 12 pp.

identify relevant documents, scholars must scroll or page through scanned images, or, in the best cases, they can perform keyword searches across sets of documents that have been converted to plain-text form. The strength of such searches is that they have a well-understood workflow, allowing even the most non-technical of researchers to quickly locate documents of interest. However, search-based interfaces require the search terms to be known in advance, making exploratory research difficult. Moreover, as archival collections are being digitized at an increasing rate, these interfaces will no longer suffice; scholars will require new techniques to sift through and make sense of this expanding amount of material.

Several scholars have proposed that "topic modelling", a computational technique for identifying recurring themes across a set of documents, might be applied to such archives. Successful applications of topic modelling to other document collections have been demonstrated in compilations of scientific research articles (Griffiths & Steyvers, 2004), traces of social media (Ramage, Dumais, & Liebling 2010), and more recently, in historical archives (discussed in depth below). While the ability of topic modelling to extract coherent and meaningful themes has been repeatedly demonstrated (Chang, Boyd-Graber, Gerrish, Wang, & Blei, 2009), we are still at an early stage as far as understanding how these themes can be employed to enhance humanities scholarship. How should humanists think about the topics that these models identify? What can these topics reveal, and what do they obscure? Finally, how can we improve upon the minimal interactive capabilities that topic models currently provide, in order to provide comprehensive automated support for the exploration of large cultural archives?

This article addresses this set of questions from three angles. First, we synthesize the outcomes of several pioneering applications of topic models to humanities scholarship, summarizing the lessons learned and the drawbacks encountered. Next, we present a case study of our own application of topic models to the *Papers of Thomas Jefferson Digital Edition*, an archive of approximately 30,000 documents. In our case study, we describe how an interactive mode of topic visualization might be employed to assist scholars in assembling and refining document sets based on their particular research interests. Finally, we describe our ongoing work on a new interactive system for topic model exploration in humanities scholarship, based on a prototype user interface called TopicViz (Eisenstein, Chau, Kittur, & Xing, 2012). TopicViz integrates theme-based exploration with a traditional keyword-based interface, thus helping to reveal documents and search terms that might not have been considered at the outset. Once developed, our system will allow scholars to pursue new pathways of connection among archival documents, facilitating a rhizomatic mode of research and discovery unconstrained by the original (print) form of the archive, or by any prior conception of what the archive might contain.

What are topic models?

Topic modelling is a text mining technique that applies probabilistic inference to identify latent themes, or "topics," in a set of documents. Topics are probability distributions of the words in a document collection's vocabulary; however, due to the difficulty of visualizing a probability distribution over thousands of words, topics are usually summarized by a list of the words with the highest probability relative to other topics.

Klein, Lauren, & Eisenstein, Jacob. (2013). Reading Thomas Jefferson with TopicViz: Towards a Thematic Method for Exploring Large Cultural Archives. *Scholarly and Research Communication, 4*(3): 0301121, 12 pp.

The only inputs into a topic model are the number of topics desired and the text of the documents—no manual annotations are needed. The model's output consists of two parts: a set of topics—that is, clusters of words that appear in similar documents—and the topical composition of each document. For example, our research on the *Papers of Thomas Jefferson* revealed the following topics (selected from 50 topics in total):

- bones, mineral, animal, philosophical, buffon, needle, curiosities, teeth;
- creditors, wayles, assets, bills, balance, drawer, debtors, specie, moiety, debit;
- despotic, monarchy, delegated, leaders, abuses, decisions, monarchical, and bias.

Each topic clearly indicates a different theme in the Jefferson archive: science, debt, and politics. Topics such as these, often summarized by short phrases, constitute the first portion of any topic model's output.

The second portion of the model's output consists of a characterization of each document in terms of the topics it contains. On the basis of the topics described above, one document might be characterized as 80% politics, 15% science, and 5% debt; another might be characterized as 25% politics, 0% science, and 75% debt. It is important to underscore that these themes are not defined in advance, but rather are extracted from the digital archive. In general, the user of the model specifies the number of topics, and then each document is described with a vector of percentages—one for each topic.

While there are many types of topic models, the most simple is the original Latent Dirichlet Allocation (LDA) of Blei, Ng, and Jordan (2003). This model is notable for its strong simplifying assumptions, which we will briefly describe. First, it employs a "bag-of-words" model, in which each document is treated as a bag of words, so that word order (and thus linguistic structure) is irrelevant. Next, the model assumes that each word was chosen randomly by author, but that this random choice has been shaped by the topics that characterize the document. For instance, if a document in the Jefferson archive is characterized as 80% politics, then the assumption is that 80% of the words in the document are drawn (randomly) from the politics topic. This means that words like "despotic" will have a high probability of occurring, and words like "mineral" (not included in the politics topic) will be rare. These assumptions leave no place for rhetorical structure or even basic grammar, but they, nonetheless, constitute the basis for the simplest possible model that can extract topics from text.[1]

With these modelling assumptions in hand, one can write software that works backwards from a document collection in order to infer which topics were discussed in any particular document. For example, if "despotic" and "mammal" never appear in the same document, then they should not have a high probability of occurrence in the same topic. Such probabilities are precisely what topic modelling tools such as MALLET provide (McCallum, 2002). Blei's (2012) tutorial describes this process in more technical detail.

Survey of topic modelling in humanities research

In recent years, scholars from across the humanities have begun to investigate the applicability of topic modelling to humanities scholarship. Among the earliest

examples is Cameron Blevins' work on the diary of Martha Ballard, an eighteenth century New England midwife (2011). Blevins employed MALLET to generate a list of thirty topics contained within the nearly 10,000 diary entries. Blevins then graphed the strength of relevant topics—such as "cold weather" or "housework"—over time. For the most part, the graphs confirmed what a reader might reasonably assume: discussions of cold weather increased during winter months; discussions of housework increased as she aged and faced increasing financial and legal difficulties. Blevins identifies one topic, which he labels "emotion," as worthy of further examination: this topic demonstrates a general increase over time, with anomalous increases in the years between 1803 and 1804. Blevins attributes this increase to the fact that, during that time, Ballard's husband was imprisoned for debt and her son was indicted for fraud. Blevins concludes that in this case, as with others, "MALLET did a better job of grouping words than a human reader," and that it suggests a "new and valuable way of interpreting source materials" (Blevins, 2011).

Robert Nelson (2012), in his work on topic modelling the Richmond, Virginia, newspaper, the *Daily Dispatch*, between 1860–65, identifies both the "classificatory power as well as the limitations of topic modeling for individual documents within a larger corpus." Like Blevins, he also observes that the "real potential of topic modeling" is that it "allows us to step back from individual documents and look at larger patterns among all documents, to practice not close but *distant reading*, to borrow Franco Moretti's memorable phrase" (Nelson, 2012, author emphasis). Nelson produced a website called *Mining the Dispatch,* which provides an interface that allows users to explore the words and documents associated with each topic. While the interface is an important step towards interactivity, a particularly significant contribution of the project is the analysis that Nelson himself provides of the topics (and associated graphs). Nelson's analysis of the waning of advertisements that solicit slaves for hire versus those that advertise fugitive slaves, for instance, or his interpretation of the multiple topics that describe types of soldiers—deserters, casualties, prisoners, etc.—suggest how topic modelling, made accessible to the non-technical user, can serve as a substantive starting point for future historical research.

Taking a more literary approach, Ted Underwood and Jordan Sellers have applied LDA to a broad set of eighteenth and nineteenth century texts (Underwood & Sellers, 2012). Their most significant conclusions for the humanities relate to the differentiation of literary and non-literary diction in the transition to the nineteenth century. While traditionally, scholars tend to credit William Wordsworth with inciting a revolution in poetic language, pushing it closer to everyday speech, Underwood and Sellers's analysis shows that all literary forms move further from non-literary language as the century progresses. In conversation with Lisa Rhody, whose work is described below, Underwood also addresses the question of how the topics produced by topic modelling should be understood. He suggests that they can most generally be considered "trends," but that their meaning shifts in relation to the size of the dataset. For smaller datasets, such as Rhody's (2012), they might be considered "semantic topics," while for larger datasets, they are more accurately theorized as "discourses." We address these conceptualizations of topics, and offer additional distinctions, later in this article.

Klein, Lauren, & Eisenstein, Jacob. (2013). Reading Thomas Jefferson with TopicViz: Towards a Thematic Method for Exploring Large Cultural Archives. *Scholarly and Research Communication, 4*(3): 0301121, 12 pp.

Turning to Rhody's work, we see how LDA can be used to explore a set of 4,500 works of twentieth century poetry. Rhody's work departs from Underwood and Sellers', and those above, in her emphasis on the exploration and analysis of figurative language. As with Blevins, Nelson, and Underwood and Sellers, graphical visualizations of topics and their association with metadata play a key role. While her overall project, an investigation into the use of ekphrasis—or, the textual description of a visual object—is still ongoing, her theorization of topic modelling when applied to literary texts is important to the field. Rhody (2012) identifies four types of topics that appear when modelling poetry: Optical character recognition (OCR) errors and foreign language topics; "large chunk" topics, which result from an unusually large document that skews the statistical analysis; "semantically evident" topics, in which the thematic connection is clear; and, "semantically opaque" topics, in which the words have no clear thematic relationship to each other. Like Nelson (2012), she underscores that the topics revealed by LDA models applied to "highly figurative texts must be the starting point for an interpretive process". She urges scholars to explain why, for instance, "a topic with keywords like 'night, light, moon, stars, day' isn't just about time of day." She argues that, "more likely, it's about the use of time of day as images, metaphors, and other figurative proxies for another conversation and none of that is evident without a combination of close and 'networked' reading."

Finally, Matthew Jockers' *Macroanalysis: Digital Methods and Literary History* (2013) promises provides the first book-length investigation into topic modeling (among other machine learning techniques) applied to literary texts. Jockers's website provides an interface in which users can select from a set of 500 themes extracted from a corpus of nineteenth century novels and view the theme's associated words in a word cloud. Additional charts show the distribution of the topic in terms of the gender of the author, the country of publication, and the year of publication. This linking of topics and metadata is a core element of each of these analyses, although each researcher had to cobble together multiple tools to produce visualizations of this linking. Thus, more holistic software support for visualizing the relationship between topics and metadata could facilitate exploration and lead to new insights.

Case study: Topic models of the *Papers of Thomas Jefferson*

We conducted our case study of topic modeling on the *Papers of Thomas Jefferson Digital Edition*, which was made available to us through an agreement with the Rotunda imprint of the University of Virginia Press. The digitization process is ongoing; our current dataset includes all of the Jefferson papers included in Volumes 1–36 of the print version of the *Papers of Thomas Jefferson*, Main Series, as well as Volumes 1–4 of the Retirement Series, published by Princeton University Press. Each series is ordered chronologically. As such, our dataset contains documents composed between January 14, 1760 and March 3, 1802, and between March 4, 1809 and April 30, 1812. Volumes 37–39 of the Main Series, and 5–9 of the Retirement Series, are awaiting digitization and encoding by Rotunda. The remaining volumes, covering the years between 1803 and 1809, and 1816 and 1826, are awaiting publication.

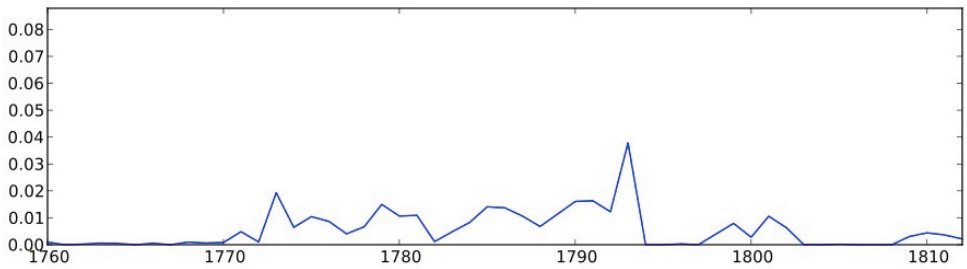
In parsing the digitized files, which we received in XML form, we treated any tag as a potential document if it contained an <FGEA:author> child, resulting in a total of 15,111 letters. From this set, we randomly sampled 10,000 letters to use as our corpus.

Within each letter in the corpus, we considered only text marked as English, and broke the text into tokens using NLTK, the Natural Language Processing Toolkit (Bird, Klein, & Loper, 2009). We converted each token to lowercase, and considered only tokens that contained all alphabetic characters; we did not perform any additional processing. We limited the vocabulary to the 8,000 most frequent words in the corpus. We applied a custom implementation of LDA, as described above, closely based on the original Blei, Ng, and Jordan (2003) model. We used 50 topics, and ran the model for 100 iterations. The choices of 10,000 documents, 8,000 words, 50 topics, and 100 iterations were based on the goal of obtaining the most precise and accurate model under the limits of what could be computed overnight on a standard laptop.

Many of the 50 topics fell into the following broad categories: epistolary convention and style (T3, T7, T8, T12, T20, T23, T25, T44, T45, T46, and T48), politics (T0, T1, T2, T9, T15, T18, T24, T31, and T34), war (T10, T11, T15, T16, and T24), commerce (T5, T6, T21, T41, and T42), finance (T4, T28, T41, and T47), medicine (T13 and T27), law (T29 and T30). Others were more specific: education (T27), farming (T5), family (T32), printing (T38), ideology (T31), Africa (19), Compte du Buffon (T22), Native Americans (T33), fishing (T42), the Caribbean (T43), French Language (T49). The remainder were inconclusive (T14, T26, T35, T36, T37, and T39).[2] A sustained analysis of any of these topics might yield intriguing results. We focus on Topic 43, which we have selected for its relevance to the Caribbean, so as to suggest how scholars might use the topics generated by topic models to inform and direct future research.

The twenty words most closely associated with Topic 43 are: crescent, Muley, Landais, Suliman, Puchelberg, gunner, brigantine, lordships, deck, Martinique, Jamaica, pilot, Hispanola, privateer, undersigned, crew, vessel, brig, captured, and sloop. The chart below indicates the prevalence of this topic in the *Papers of Thomas Jefferson* over time. Specifically, the y-axis of the figure is the estimated proportion of words generated from this topic in the documents that were written in each year on the x-axis.

Chart 1: Prevelance of Topic 43 over time



This topic appears most often in letters composed by Thomas Jefferson, George Hammond (a British diplomat), Thomas Barclay (an American diplomat), James Maury (an American Anglican minister), Joshua Johnson (a merchant and U.S. consul), David Humphreys (Jefferson's secretary), Samuel Smith (an American Presbyterian minister), George Washington, and John Paul Jones (a naval officer). The most frequent recipients of these letters are: Thomas Jefferson, George Washington, George Hammond, the Board of Trade, Edmond Charles Genet (the French ambassador to the

Klein, Lauren, & Eisenstein, Jacob. (2013). Reading Thomas Jefferson with TopicViz: Towards a Thematic Method for Exploring Large Cultural Archives. *Scholarly and Research Communication, 4*(3): 0301121, 12 pp.

U.S.), John Brown Cutting (a war veteran and friend of Jefferson, Edmund Randolph (an attorney and politician), and the American Commissioners.

Several things about this topic are worth noting. In terms of the words that comprise the topic, several are proper names: Muley (the first name of several North African leaders), Landais (a notorious U.S. naval captain), Suliman (a Moroccan leader at the center of the First Barbary War), and Puchelberg (a German merchant), indicating individuals at the centre of a range of events that took place on or across an ocean. Others are geographic locations: Martinique, Jamaica, and Hispanola, indicating the regions in which certain events took place. Of the remaining words, several indicate roles assigned on a ship: gunner, brigantine, pilot, privateer, and crew. Others describe the ship itself: deck, vessel, brig, and sloop. Of the remaining words, two can be attributed to epistolary style: lordships and undersigned. The final words, crescent and captured, also appear related to the topic, but cannot be generalized without reference to the original documents.

In terms of the distribution of this topic throughout the *Jefferson Papers*, the range spans the period between 1771 and 1803, with a resurgence around 1810. Spikes occur in 1773, 1779–81, and 1792–93, with the largest spike occurring in 1793. Major political events during those years include: the Boston Tea Party (1773), the Spanish entrance to the Revolutionary War and its subsequent siege of Gibraltar (1779), the Great Hurricane of 1780, which killed 20,000–30,000 people in the Caribbean, the escalation of war between the British and the Dutch (1780–81), the end of the Revolutionary War (1781), the murder of enslaved passengers on the slave ship Zong (1781), and the commencement of the French Revolutionary Wars (1792–1793). Additional analysis of the letters is required to determine the specific events under discussion.

In terms of the authors and recipients of the letters, it is interesting to note that while certain correspondence, such as the letters exchanged between Jefferson and George Hammond, was reciprocal, most was unidirectional. A visualization of this correspondence network, as proposed in the section that follows, would help to substantiate this data, or any claims that might be made.

While certain aspects of the topic suggest some significance – such as the prevalence of ship-related words, the inclusion of "crescent," or the connotations of violent conflict in many of the words ("captured," "gunner," "brigantine") – the immediate response of this scholar is to ask more questions: How can I read the original letters? Which specific events were being discussed, and by whom? What other topics might contain ship-related vocabulary, and do those topics carry the same overtones of conflict? Are there other topics that contain the same (or nearby) geographic regions? Each of these questions requires a return to the larger set of topics, and an ability to manipulate the topics as well as to access the original letters. With current topic modelling tools, such as MALLET, these operations are difficult—if they can be accomplished at all. We therefore propose how a tool to visualize and interact with topic models might allow for scholars to navigate the thematic landscape of their archives, quickly building and refining document sets based on the topics relevant to their research.

Interactive visualization of topic models

Our study of the *Papers of Thomas Jefferson Digital Edition* echoes what previous investigations into the uses of topic modelling for humanities research have shown: that while topic models can offer a certain degree of insight by connecting the latent themes in the text with associated metadata, the output of the topic model is useful only to the extent to which it can support further humanistic inquiry. The unanswered question—indeed, the question that motivates our project as a whole—is how to integrate topic models and related automated text-mining tools with existing modes of scholarship, while leaving open the possibility for new syntheses of computational and humanistic modes of inquiry. We propose that the purview of topic modelling software need not stop at providing a list of topics, but rather, should work toward providing holistic support for exploratory analysis. The conceptual challenge is to provide such support without requiring the user to follow a single search path. We now consider several recent efforts to provide more interactive control of topic models, before describing our own approach.

A handful of websites now provide interactive interfaces for exploring the output of topic models, although each site is designed around a specific archive. *Mapping Texts* displays lists of topics associated with a series of historical time periods (Yang, Torget, & Mihalcea, 2011). However, there is little integration of the topics with the document metadata, or with the documents themselves, thereby limiting the utility of the topics to facilitate additional research. *The Open Encyclopedia of Classical Sites* (Mimno, 2012) displays the topics associated with a set of locations referenced in data from Google Books. It also provides visualizations of the metadata associated with each of the topics: spatial information is shown using the Google Maps application programming interface API, and temporal strength is shown with a graph. But this site, like *Mapping Texts*, displays a limited list of topics, each represented by a series of words. It is not clear how this interface would scale to models involving more topics, since current models result in hundreds or even thousands of topics (see Mimno, Hoffman, & Blei, 2012). It seems that no static visualization could possibly support this level of detail. We therefore advocate an interactive approach that allows scholars, through an iterative process, to refine the topics and texts most relevant to their research.

A new project, *Paper Machines*, suggests the wide-ranging impact of a tool that can visualize the output of topic modeling software. *Paper Machines* provides a graphical interface for a range of text analysis tools (such as MALLET, geoparser, and DBpedia Annotation), as well as related visualizations, including word clouds and phrase nets. It accepts as its input the contents of a Zotero library (Zotero is a robust citation management tool), and generates a set of visualizations based off the contents of that library. In addition to the metadata associated with the citations themselves, *Paper Machines* is also able to analyze the full text of any website "snapshots" in the library, as well as any the text of any PDF that can be extracted with OCR. Currently *Paper Machines* works best with large Zotero collections, allowing scholars to visualize the contents of a single collection, or to compare the contents of multiple collections in terms of topics, locations, dates, etc. *Paper Machines'* synthesis of topic modelling techniques with relevant metadata – and its incorporation of advanced visualization techniques – suggests how a more general tool might allow scholars to track the

Klein, Lauren, & Eisenstein, Jacob. (2013). Reading Thomas Jefferson with TopicViz: Towards a Thematic Method for Exploring Large Cultural Archives. *Scholarly and Research Communication, 4*(3): 0301121, 12 pp.

themes that recur across any document collection, and, with more robust interactivity, might employ visualization as a method of archival search.

While these efforts are promising, we see significant unsolved problems in relating topic models to humanities research. One of the primary strengths of topic models is that the topics are not defined in advance, but rather arise directly from the text itself. It is this capability that allows topic models to lead researchers in new directions unforeseen by previous assumptions about the document collection. But this also means that researchers face the challenge of determining the significance (or lack thereof) of dozens or even hundreds of topics, and then of identifying the topics that are relevant to their specific research questions. Another issue is that many existing visualizations of and interfaces for topic models are centred on the topic model itself. This approach calls to mind the critique often attributed to computer science pioneer Edsger Dijkstra (1986) that "computer science is no more about computers than astronomy is about telescopes" (p. 8). That is, the affordances of digital humanities software should be designed around the needs of humanities scholarship, rather than the predefined characteristics of any particular technical approach. It is therefore important to note that existing workflows built around keyword search are not only well entrenched but are also often quite successful. An interactive visualization that integrates topic models into this already familiar workflow may be more immediately useful than an approach that begins with the topic model and then asks how it might best be presented to the user.

Developing support for topic modelling on humanities archives

The authors of this article are currently developing a software system that integrates interactive topic model exploration with keyword search. The first prototype of this system, TopicViz, was designed for collections of scientific research articles. The next version of this software – TOME, for Interactive Topic Model and Metadata Visualization – will support the exploration of historical archives. We now describe the basic capabilities of the prototype system, and consider how these capabilities might be brought to bear on the types of archives which humanities scholars rely upon for their research.

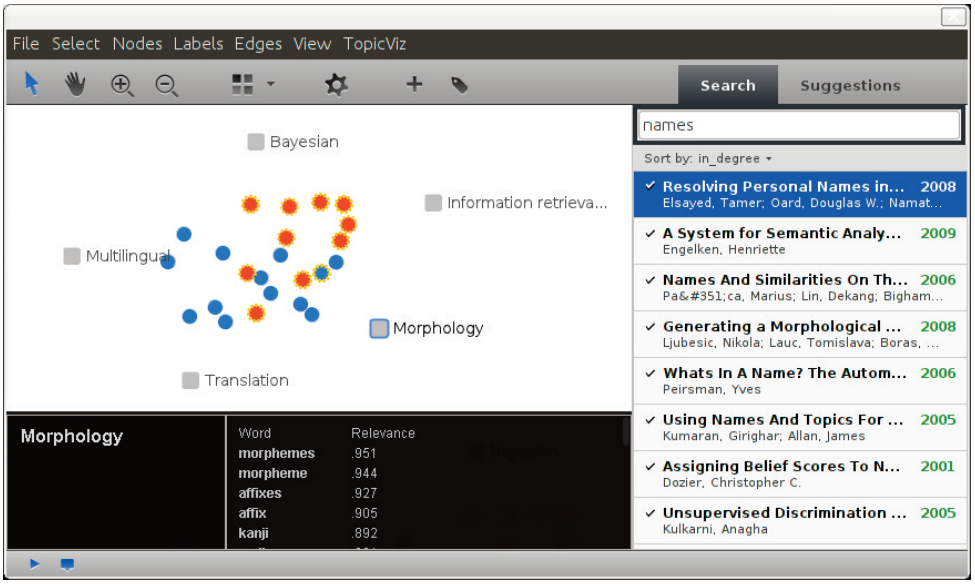
The central idea of TopicViz, which we will echo in TOME, is to bridge the gap between keyword search and topic model exploration. The interaction begins with a keyword query – say, for "emancipation" – but rather than return a list of documents, as in a traditional search interface, TopicViz displays the documents in an interactive spatial layout, organized by the topics that best match the main query results. The user can then navigate the thematic landscape, identifying the topics relevant to her research along with their metadata properties. For example, the documents relating to "emancipation" might focus on three primary themes: political measures, moral arguments, and supporting institutions. TopicViz allows the scholar to rapidly identify the documents that focus on each of these themes, and to refine the working set of documents – for example, by removing all documents whose main emphasis is "moral arguments."

This visual, topic-based method of exploration allows the humanities scholar to quickly build and refine document sets based on their thematic characteristics, rather than by

assembling large and possibly error-prone lists of keywords. It can also help to reveal additional search terms that might not have initially been considered. In the case of institutions developed in the wake of emancipation, for instance, the scholar may not think to search for "prison," and as a result, would not see the documents that describe the institution that arguably carries the most profound implications for the present day. We anticipate an iterative interaction, in which the user identifies an initial set of search terms, and then employs the visualization in order to both explore the most relevant topics, and to discover thematically-relevant documents that missed the initial query.

Given a query and a set of document "hits," TOME will reveal their main topics through a dust-and-magnet visualization (Yi, Melton, Stasko, & Jacko, 2005). In this mode of visualization, each topic acts as a "magnet", exerting force on nodes representing the documents. The stronger the relationship between the topic and the document, the greater the force that the topic magnet exerts. By rearranging the topic magnets, the user can create layouts that foreground specific comparisons or contrasts between topics. By adding or removing topics or documents, the user can drill down to more focused sections of the archive. Thus, even though a topic model may have hundreds of topics, the user need only interact with a small subset that connects with the current query. In the same way that a research task involves iteratively refining the set of relevant documents, the user of TOME can iteratively refine the set of relevant topics.

Figure 1: Screenshot from the TopicViz system



An example of this type of interaction is shown in Figure 1: a screenshot from the TopicViz system for exploring scientific research literature (Eisenstein et al., 2012). In this example, the user has already entered a search query and retrieved a set of documents. She is then presented with a spatial visualization of the relationship between her search terms and the five most relevant topics. In TopicViz, topics are shown by lists of relevant words (see the bottom centre panel), but our research on TOME will investigate visual presentations that use the same form of dust-and-magnet

Klein, Lauren, & Eisenstein, Jacob. (2013). Reading Thomas Jefferson with TopicViz: Towards a Thematic Method for Exploring Large Cultural Archives. *Scholarly and Research Communication, 4*(3): 0301121, 12 pp.

visualization to display the words that are associated with each topic. This will allow the scholar to focus her research on the words most relevant to her research.

Future directions

Analyses of topic models of historical and literary texts have tended to focus on the words that characterize each topic. However, the topic model also computes the topical composition of each document. In the example of "emancipation," given above, the topic model may determine a particular document to be composed of 10% political measures, 30% moral arguments, and 60% supporting institutions. One powerful application of topic models combined with metadata would be to aggregate the topics of all documents composed by a single author, or set of authors, allowing scholars to compare and contrast each author's thematic interests.

TOME will visualize these comparisons in two distinct ways. For a small number of authors, the relevant document nodes in the visualization can be color coded, as in the image above. For a larger numbers of authors, we will invert the layout, using the authors as magnets (nodes whose position is determined by the user) and the topics as dust (nodes whose position is determined by the forces exerted by the magnets). A topic that is more closely associated with a particular author will be more strongly attracted to that author's magnet (this will be reflected in the spatial position of the topic's node). Visualization of the change in topical composition over time can offer an intuitive sense of how the interests of an author, publication, community, or geographic region evolved. Such a visualization can take the form of a simple line graph; alternatively, within the inverted dust-and-magnet layout described above, we can use magnets to represent time periods, and represent the topics as dust, whose position is determined by the prevalence of the topic during each relevant time period.

A final aspect targeted for further development is the ability to track the spread of topics, or influence, across social networks. Since influence is in its most basic sense a form of causality, we propose to determine it by first identifying which topics originate with a single author, newspaper, or region, and then by determining which other authors—or which other sources—later took up the same topics. Not only might this research yield humanistic insight, it also goes well beyond the capabilities of traditional topic models. As such, it demonstrates the powerful potential of direct collaborative research between scholars from the humanities and computer science.

Acknowledgments

The authors wish to thank David Sewell, Editorial and Technical Manager of the Rotunda imprint of the University of Virginia Press, for granting us access to the XML files of *The Papers of Thomas Jefferson Digital Edition.*

Notes

1. More linguistically plausible models are possible, but they are computationally more complex, and may not yield substantially more coherent topics.

2. Appendices: to see Appendix A, List of Topics and Appendix B, Topics with metadata, please visit http://dhlab.lmc.gatech.edu/projects/tj-topicviz .

References

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. Sebastopol, CA: O'Reilly Media.

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Blei, D. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.

Blevins, C. (2011). Topic modeling historical sources: Analyzing the diary of Martha Ballard. *Proceedings of Digital Humanities*. Stanford, CA: Digital Humanities.

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., & Blei, D. (2009). Reading tea leaves: How humans interpret topic models. *Proceedings of Neural Information Processing Systems (NIPS)*. Vancouver, BC: NIPS.

Dijkstra, E.W. (1986). On a cultural gap. *The Mathematical Intelligencer*, *8*(1), 48-52.

Eisenstein, J., Chau, D. H., Kittur, A., & Xing, E.P. (2012). TopicViz: Semantic navigation of document collections. *Supplemental Proceedings of Conference on Human Factors in Computing Systems (CHI)*. Austin, TX: CHI.

Folsom, E. (2007). Database as genre: The epic transformation of archives. *PMLA*, *122*(5), 1571-1579.

Griffiths, T.L., & Steyvers, M. (2004). Finding scientific topics. *PNAS*, *101*, 5228-5235.

Jockers, M. (2013). *Macroanalysis: Digital methods and literary history*. Champaign-Urbana, IL: University of Illinois Press.

McCallum, A.K. (2002). *MALLET: A machine learning for language toolkit*. URL: http://mallet.cs.umass.edu .

Mimno, D. (2012). *The open encyclopedia of classical sites*. URL: http://www.cs.princeton.edu/~mimno/oecs .

Mimno, D., Hoffman, M., & Blei, D. (2012). Sparse stochastic inference for latent Dirichlet allocation. *International Conference on Machine Learning*. Edinburgh, Scotland, UK.

Nelson, R. (2012). *Mining the dispatch*. URL: http://dsl.richmond.edu/dispatch .

Ramage, D., Dumais, S., & Liebling, D. (2010). Characterizing microblogs with topic models. *Proceedings of International AAAI Conference on Weblogs and Social Media (ICWSM)*. Washington, DC: ICWSM.

Rhody, L. (2012). Chunks, topics, and themes in LDA. *Lisa @ Work*. URL: http://lisa.therhodys.net/2012/04/chunks-topics-and-themes-in-lda .

Underwood, T., & Sellers, J. (2012). The emergence of literary diction. *Journal of Digital Humanities*, *1*(2). URL: http://journalofdigitalhumanities.org/1-2/the-emergence-of-literary-diction-by-ted-underwood-and-jordan-sellers .

Yang, T., Torget, A.J., & Mihalcea, R. (2011). Topic modeling on historical newspapers. *Proceedings of the Association for Computational Linguistics workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (ACL LaTeCH)*, 96-104. Portland, OR: ACL LaTeCH.

Yi, J.S., Melton, R., Stasko, J., & Jacko, J.A. (2005). Dust & magnet: multivariate information visualization using a magnet metaphor. *Information Visualization*, *4*(4), 239-256.