White Paper Report

Report ID: 109092

Application Number: HD-51705-13

Project Director: Lauren Klein (lauren.klein@lmc.gatech.edu)

Institution: Georgia Tech Research Corporation

Reporting Period: 5/1/2013-5/31/2015

Report Due: 8/31/2015

Date Submitted: 10/13/2015

FINAL WHITE PAPER

Grant number:
HD-51705-13

Project title:
TOME: Interactive Topic Model and Metadata Visualization

Project director:
Lauren Klein

Co-Project Director:
Jacob Eisenstein

Grantee Institution:
Georgia Institute of Technology

Report Submitted:
October 12, 2015

TOME: Interactive Topic Model and MEtadata Visualization

Lauren F. Klein and Jacob Eisenstein
Georgia Institute of Technology

Abstract

This white paper summarizes our research to date on TOME: Interactive TOpic Model and MEtadata Visualization, a tool designed to support the exploratory thematic analysis of digitized archival collections, funded through an NEH Office of Digital Humanities Startup Grant (2013-15). Thus far, the primary outcome of TOME is a set of visualizations intended to facilitate the interpretation of the topic model and its incorporation into extant humanities research practices. In contrast to other topic model browsers, which present the model on its own terms, TOME is informed by the process of conducting early-stage humanities research. In this way, our research also demonstrates the conceptual conversions--in terms of both design and process—that interdisciplinary collaboration necessarily entails. In making these conversions explicit, and exploring the implications of their successes and failures, we take up the call, as voiced by Johanna Drucker (2011), to resist the "intellectual Trojan horse" of visualization. We seek to model a new mode of interdisciplinary inquiry, one that brings the methodological emphasis of the digital humanities to bear on the practices of humanities research and computer science alike.

Introduction and Overview

William Lloyd Garrison, the nineteenth-century abolitionist, was widely renowned for his writing. The lines that began the first issue of his anti-slavery newspaper, *The Liberator*, in 1831, "I will not equivocate -- I will not excuse -- I will not retreat a single inch -- AND I WILL BE HEARD," quickly became the rallying cry of the abolitionist movement in the United States. But one decade later, with several additional newspapers under his editorial control, Garrison decided to hire a woman, Lydia Maria Child, to edit the *National Anti-Slavery Standard*, the official newspaper of the American Anti-Slavery Society. Child was most famous as a novelist, but she also wrote stories for children and had published a bestselling cookbook. Garrison hoped that Child could bolster female support the cause, "impart[ing] useful hints to the government as well as to the family circle" (Karcher 1994). But did she? And if so, how effective—or how widely adopted—was this change in topic or tone?

Questions like these, about the evolution of issues and ideas, prompted our work on the TOME project, which we describe in this white paper. TOME, short for Interactive TOpic Modeling and MEtadata Visualization, is a tool designed to support the exploratory thematic analysis of digitized archival collections. Muralidharan and Hearst (2012) have proposed that we think of the exploratory phase of humanities research as a process of *sensemaking*, one in which scholars aggregate concrete pieces of evidence into more abstract theoretical models. Their work on the WordSeer project draws upon the concept of Exploratory Data Analysis (EDA), which has played a role in quantitative research since the 1970s (Tukey 1977). In comparison to formal hypothesis testing, EDA is intended to help the researcher develop a general sense of the properties of the dataset before embarking on more specific inquiries. Typically, EDA combines visualizations such as scatterplots and histograms with lightweight quantitative analysis, serving to check basic assumptions, reveal errors in the data-processing pipeline, identify relationships between variables, and suggest preliminary models (Gelman 2004). Ideally, these models motivate additional research questions, questions which generate new evidence, and this evidence in turn leads the scholar to refine or revise the initial model of the data (Russell et al. 1993).

By incorporating ideas about EDA into the humanities research process, tools such as WordSeer and TOME contribute to a small but growing body of work designed to meet the core challenge of humanistic research: humanistic research questions, such as those about gender and influence posed above, are most often formulated in high-level thematic terms. By contrast, most computationally-assisted research tools facilitate low-level modes of inquiry. For instance, the search engine—arguably the most ubiquitous form of computational support for research—operates at the level of query strings, offering at best the flexibility of wildcards and boolean connectors. Corpus-based methods such as concordances and word-frequency plots are similarly confined to the literal; they may provide counts of local textual contexts or visualizations of temporal trends, but they require that scholar have a specific set of words in mind before pursuing further investigation (Michel et al., 2011). These techniques are of significant value at the midpoint of the research process, when the scholar has already identified a specific line of inquiry. But in the earlier, more exploratory phases of research—that is to say, when faced with a new or otherwise unfamiliar archive—the humanities scholar must first ask a more basic question: "What is there?"

Topic modeling is a computational technique developed to answer this question. By automatically analyzing the co-occurrence counts of words or phrases across a collection of documents, the topic model generates a set of topics, often described as themes, consisting of the words or phrases that tend to appear together (Blei 2012). When applied to newspaper datasets, topic models most helpfully point to thematic patterns, and for this reason—as well as for issues of legibility—we employ the terminology of "themes" over "topics" in describing our approach. But it is important to underscore that the topics generated by the model might also identify specific historical events, notable stylistic features, or systematic transcription errors, to name only a handful of non-thematic topical modes. In her study of ekphrastic poetry, for instance, Lisa Rhody has argued for the importance of attending to both the thematic and non-thematic topics generated by the model (2012). In their application and analysis of a topic model of the journal of literary criticism, *PMLA*, Andrew Goldstone and Ted Underwood have proposed that we replace the term "theme" with "discourse," since topics can also reflect "rhetorical frames, cognitive schema, or specialized idioms" (2014). They rightly caution, however, that it remains "the task of the interpreter to decide, through further investigation, whether a topic's meaning is overt, covert, or simply illusory." To be sure, the model's computational nature does not guarantee that its topics will offer up insight, but topic modeling nonetheless provides a new entry-point into the *process* of sensemaking. The recent surge of interest in the application of topic modeling to the humanities (e.g. Newman and Block 2006; Jockers and Mimno 2013; Goldstone and Underwood 2014) suggests the potential of topic models, placed into the sensemaking phase of the humanities research process, to advance active research questions in humanities fields.

Our project, TOME, was conceived to address what we see as the two primary impediments to more widespread adoption of the topic models in the humanities. First, the great advantage of topic models is that they are based on nothing more than word statistics; that is to say, they neither require nor assume any model of grammar, discourse, semantics, or argumentation. While it is this agnosticism that makes topic models universally applicable, this strength becomes a weakness when the time comes to marshal a topic model into an argument. There is no inherent reason to believe that words grouped together on the basis of co-occurrence statistics should really mean or prove anything, aside from the winkingly suggestive similarities that these word groups so often display. For topic models to be truly integrated into humanities scholarship, additional interpretive functionality is required: the user must be able to probe the semantic associations that model proposes, and seek out additional perspectives on the model itself.

Second, for topic models to play a significant role in humanities research, they must be integrated into existing scholarly workflows. At present, most topic model browsers (surveyed below) ask the user to interact with the model on its own terms: listing the topics, pivoting from topics to keywords, graphs, and metadata, etc. While these affordances are helpful from the perspective of understanding the topic model, they remain disconnected from the humanities scholar's goal of making sense of the underlying documents, as well as from any more sustained investigation that might follow. Such browsers are similarly disconnected from more familiar modes of query-based search. Given the conceptual demands of topic-driven exploration, we believe it to be important that the scholar not also be required to assimilate a new research process into his or her workflow. We propose that incorporating topic models into a more familiar mode of research will facilitate wider adoption of the technique in the humanities. This research can thus be seen as steps towards integrating topic models with humanities scholarship. We present the designs for two prototype interfaces, each of which addresses the two issues sketched out above—facilitating interpretation of the topic model, and incorporating it into extant research practices—through the medium of *interactive visualization*. In general, interactive visualization is understood as a pathway to insight, enhancing or otherwise amplifying cognition (Ferster 2012; Card et al. 1999). We see an additional generative confluence in the thematic abstraction that results from the topic model and the visual abstraction that is entailed in the design and implementation of visualizations. We explore this confluence in our discussion of our prototypes, which provide a number of visual and interactive affordances for better understanding the links between topics, words, documents, and metadata. We then place the TOME interface in the context of related work, identifying its innovative features and sketching its future development and use.

Finally, we note that we have pursued this research in an interdisciplinary context, combining our backgrounds in interaction design, data visualization, computational text analytics, and nineteenth-century American literature. By bringing together technical expertise with domain-specific knowledge, we have sought to design an interface informed by current computational research that nevertheless remains focused on the scholarship about slavery and its abolition that first motivated the project. Thus, while this paper is on one level about a set of technological and visual design decisions, it is also on another level about the conceptual conversions—in terms of both design and process—that interdisciplinary collaboration necessarily entails. In making these conversions explicit, and exploring the implications of their successes as well as their failures, we take up the call, as voiced by Johanna Drucker (2011), to resist the "intellectual Trojan horse" of visualizations that "conceal their epistemological biases under a guise of familiarity." For visualizations are powerful because they reflect existing assumptions about knowledge production as much as they facilitate new modes of inquiry. By identifying our own disciplinary assumptions, and explaining how our design decisions at times enforce and at times overturn them, we seek to model of new mode of interdisciplinary inquiry, one that brings the methodological emphasis of the digital humanities to bear on the practices of humanities research and computer science alike.

Corpus and Model

Our test corpus was a set of abolitionist newspapers from the nineteenth-century United States in which antislavery advocates mounted moral, social, and political arguments in favor of a general emancipation. These newspapers are significant not only because they function as the primary record of slavery's abolition, but also because they are among the earliest examples in the US of women and men (as well as African Americans and whites, Northerners and Southerners, US citizens and those from abroad) writing together, in the same pages. These newspapers present a particularly compelling dataset for topical analysis, as scholars have long believed that similar ideas were framed differently by (and for) these diverse audiences (e.g. Davis 1997; Clytus 2007; Dudden 2011). In designing our prototype interfaces, we focused on one newspaper, *The Anti-Slavery Bugle*, published in New Lisbon, Ohio, between 1845 and 1861. (We assembled a larger test corpus, described more fully in "Appendix 1" below, although we did not employ it in the project's first iteration which this white paper describes). The *Bugle* is noteworthy because it was the source of substantial reprinting (Golden 2014), and because it underwent several distinct shifts in editorial control. The *Bugle* therefore raises questions about the degree to which the themes expressed in its pages reflect the particular perspective of the editor at the time.

We first employed the plain-text (OCR) version of the *Bugle* provided by the U.S. Library of Congress, available through the *Chronicling America* project website. We then applied standard Latent Dirichlet Allocation (LDA) topic analysis (MALLET; McCallum 2002) with 100 topics and standard parametrization; these settings worked well enough "out of the box" that no further tuning was necessary. Another important direction for supporting the use of topic models would be to facilitate exploration of the topic model parameters, as well as of the dataset and vocabulary. In this research, however, we focus entirely on visualizing the output of the topic model, which we view as a challenge in its own right. The possibility for integrating these complementary research directions is discussed in more detail in the conclusion.
Latent Dirichlet Allocation revealed a number of topics that might intrigue a scholar in the initial phases of research, including:

- T40 ("Civil Rights"): states state law constitution tho government power united laws congress rights people con ohio tion act union question property

- T55 (“Mexican-American War”): war mexico texas mexican army peace president territory troops united government military treaty annexation mexicans country polk taylor republic
- T56 (“Native Americans”): indians indian tribes tribe chiefs frontier dian treaties tiger hawk antelope annuity fiscal lllack hyenas tigers dians avalanche savages
- T59 (“Women’s Rights”): woman women rights husband wife sex sho marriage property married mrs female legal sphere equality estate social duties sexes

These topics each suggest deeper insights, discussed more fully in the sections below. They also reveal the errors inherent in automated digitization processes, such as OCR. (For a discussion of the uses of topic modeling for identifying transcription errors, see Rhody 2012.) Here, we note only that these word lists (and associated topic distributions) point to the limits of what can be learned from the topic model alone. For these topics to play a role in substantive humanities research, scholars must be able to make sense of these topics, and our goal is to go beyond LDA by providing interactive visualizations that support this sensemaking activity.

Making Sense of Topic Models: Two Prototype Interfaces

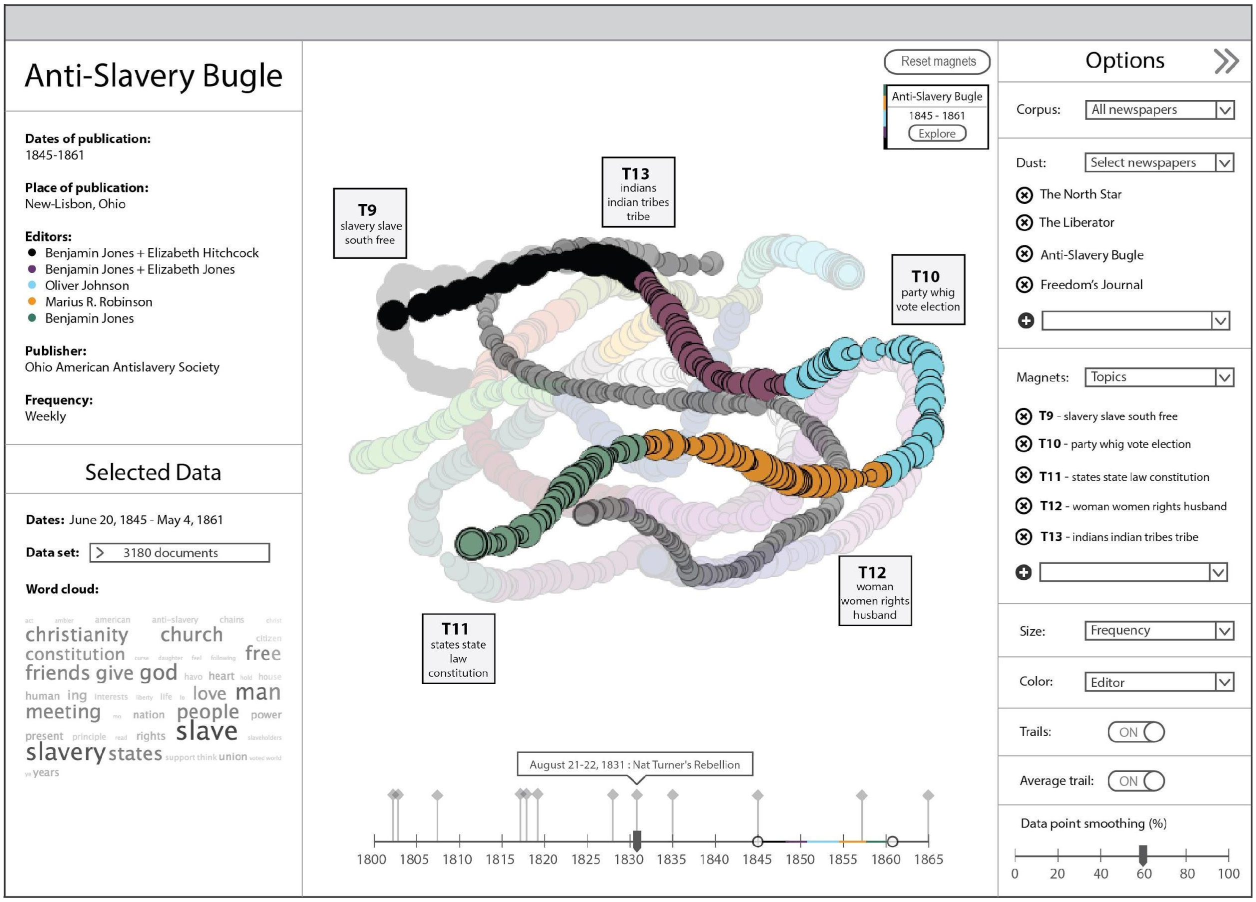
The first step towards prototyping the TOME interface was to consider how to best illuminate the features of the topic model for humanities scholars, and subsequently, how to best align the interface with more familiar methods of conducting humanities research. To address the former, we focused on the importance of visualizing the archive’s *thematic* landscape-- that is, the topics that best reflect the contents of the archive at a given point in time. Since archival materials are almost always associated with temporal metadata—in this case, publication date—we wanted our interface to also include the ability to visualize thematic *change over time*. For some topics, such as the Mexican-American War (T55, above), history tells us that there should be a distinct start and endpoint. But for topics less attached to specific historical events, such as T40, above, which seems to describe civil rights, we expect to see its prominence wax and wane. Indeed, it is this type of topic-- broader and more thematic, and of the sort that Goldstone and Underwood might hold up as a “discourse”—that holds the most promise for humanities research. To return to the example of Lydia Maria Child, one might therefore employ a *thematic* visualization of *change over time* in order to ask: Did Child employ the language of the home to advocate for equal rights, as Garrison hoped she would? Or did she merely adopt the more direct line of argument that other (male) editors employed?

These research questions lead to the second aspect of our interface’s underlying conceit: its alignment with existing humanities research practices. There is a difference, we came to realize, between a scholar who is new to an archive, and therefore primarily interested in understanding its overall thematic landscape; and a scholar who already has a general sense of the archive’s contents, and has a specific set of research questions in mind. This is the distinction between the process of exploration theorized by Tukey (and adapted for the humanities by Hearst and Muralidharan), and a later-stage process we might call investigation. The prototype interfaces discussed in this section were both designed with this two-phase research process in mind. Each rests on a distinct conception of the visual space, and foregrounds a specific mode of conducting research. We view the first as *comparative*, and the second as *multimodal*. We present both designs because we believe that their success as well as their failures help render visible—to humanistic and computational fields alike—the complex processes of conducting archival research.

Prototype I: Comparative Research in Thematic Space

Our first prototype is premised on a thematic use of space, translating the topic model into two dimensions in order to facilitate comparative mode of research. In this design, we built upon the concept of a *dust-and-magnets* visualization (Yi et al., 2005), initially developed in order to visualize the topics in scientific journal abstracts. In our adaptation of this model, shown in Figure 1 (below), each newspaper is represented as a trail of dust, with each circle corresponding to a single issue of that newspaper. The position of each point is determined by the issue's topical composition—but only with respect to the specific topics displayed on the screen (in this case, T9, T10, T11, T12, and T13). The topics, here represented as squares, exert a "magnetic" force on each point, pulling each point closer to the topics that are more prominent in that issue. For example, if a given newspaper issue contained words from T9 and T13 in equal measure, with no indication of T10, T11, or T12, then the corresponding circle would be positioned in between the squares representing T9 and T13. (This scenario is closest to the darkest portion of the highlighted dust trail, at the top left of the image below). The area of each circle represents the extent to which it is composed from the topics on the screen; a small circle indicates that the issue is largely composed of other topics, which the user could then add to the view.

We have highlighted the dust trail of the *Anti-Slavery Bugle* as it might relate to five topics, such as civil or women's rights. (The figure below is a mockup; it does not employ actual data). We also used color to convey additional metadata; here, each color corresponds to a different editor. By comparing newspaper dust trails to each other, as well as by looking at individual newspaper trails, the thematic differences between (or within) publications become illuminated. So, for instance, one could see how the *Bugle*'s coverage of civil rights compared to the *National Anti-Slavery Standard*'s; or one could see how Child's treatment of the topic, in the *Standard*, compared to the (male) editors who preceded her.



Anti-Slavery Bugle

Dates of publication: 1845-1861

Place of publication: New-Lisbon, Ohio

Editors:
- ● Benjamin Jones + Elizabeth Hitchcock
- ● Benjamin Jones + Elizabeth Jones
- ● Oliver Johnson
- ● Marius R. Robinson
- ● Benjamin Jones

Publisher: Ohio American Antislavery Society

Frequency: Weekly

Selected Data

Dates: June 20, 1845 - May 4, 1861

Data set: 3180 documents

Word cloud:

Options

Corpus: All newspapers

Dust: Select newspapers
- The North Star
- The Liberator
- Anti-Slavery Bugle
- Freedom's Journal

Magnets: Topics
- **T9** - slavery slave south free
- **T10** - party whig vote election
- **T11** - states state law constitution
- **T12** - woman women rights husband
- **T13** - indians indian tribes tribe

Size: Frequency

Color: Editor

Trails: ON

Average trail: ON

Data point smoothing (%): 60

Figure 1: Prototype I: Comparative Research in Thematic Space

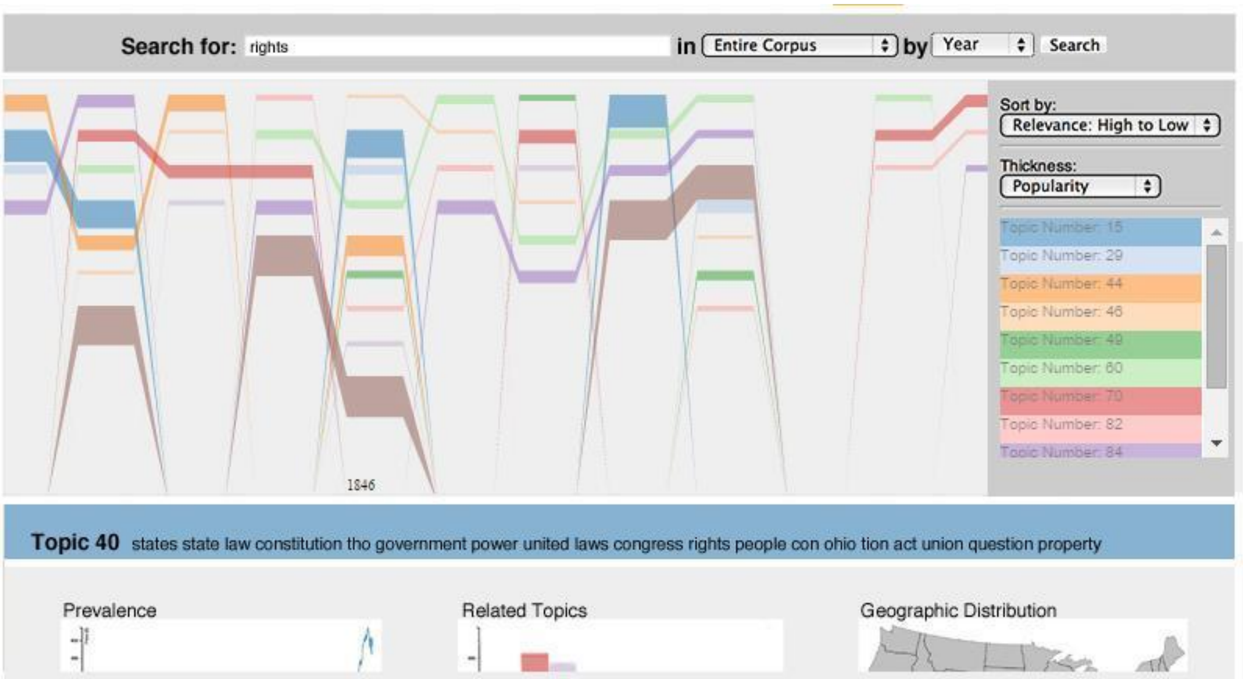
Any visualization entails an abstraction, and for this reason, information is necessarily lost. In this prototype, what is lost is the multidimensional nature of the topic model itself. In other words, documents are almost always composed of more than two topics. For the topical composition of these newspapers to be represented most accurately, the topics should be arranged in an n-dimensional space, where "n" corresponds to the total number of topics. As with the initial dust-and-magnets concept, our prototype cannot express this multidimensional information. We can, however, expose the additional dimensions through interaction: as the user adjusts the position of each topic-magnet, the points representing each newspaper issue move interactively, enabling the exploration an infinite number of spatializations. Facilitating multiple views might therefore allow the engaged researcher to develop an understanding of the overall topic distribution of the archive.

This design also lends itself to the goal of allowing the scholar to "drill down" to a subset of topics or documents. For example, the scholar might want to compare the conversation around civil rights with the one that framed the discussion of women's rights. (As scholars of the nineteenth-century United States know well, these conversations took place contemporaneously, but advocates often disagreed about whether ending slavery and enfranchising women could be achieved through the same actions.) To do so, she would remove the topic-magnets for all extraneous topics, and the dust trails would adjust.

This design nevertheless carries some substantial disadvantages, as we came to see after exploring additional usage scenarios. For one, the topic distributions computed for each newspaper are not guaranteed to vary with any consistency—that is, while some topics appear and then disappear forever, others repeatedly surge and diminish. In such cases, the resultant trails are not spatially coherent unless smoothing is applied post hoc. This has the effect of diminishing the accuracy of the representation, and raises the question of how much smoothing is enough. A second disadvantage is that while the visualization facilitates the comparison of the *overall* thematic trajectories of two newspapers, it is not easy to align these trajectories-- for instance, to determine the thematic composition of each newspaper at the same point in time. We considered interactive solutions to this problem, such as adding a clickable timeline that would highlight the relevant point on each dust trail. (This is sketched in our prototype). However, the various interactive solutions moved us further from a visualization that was immediately intuitive.

Prototype II: Multimodal Research in Temporal Space

Motivated by the strengths and weaknesses of our first prototype, we designed and implemented a second prototype interface with complementary affordances. Rather than merging exploratory and investigatory modes into a single comparative view, this interface facilitates a *multimodal* research process. Here, *time* provides the structure for the interface, anchoring each research mode—exploration and investigation—in a single frame. Figure 2 (below) represents the topics in "timeline" form. (The timeline-based visualization also includes smooth zooming and panning, using D3's built-in zoom functionality). The user begins by entering a query term, as in a traditional keyword search. Instead of a list of websites or documents, however, what is displayed is a visualization of the topics that contain that keyword, ranked from top to bottom in terms of *relevance*—the frequency with which the query appears in each topic.



Search for: rights in Entire Corpus by Year

Sort by: Relevance: High to Low

Thickness: Popularity

Topic Number: 15
Topic Number: 29
Topic Number: 44
Topic Number: 46
Topic Number: 49
Topic Number: 60
Topic Number: 70
Topic Number: 82
Topic Number: 84

Topic 40 states state law constitution tho government power united laws congress rights people con ohio tion act union question property

Figure 2: Prototype II: Multimodal Research in Temporal Space

Figure 2 displays the result for a search for the word “rights,” with each relevant topic presented as a block with a unique color; the height of each block corresponds to its overall popularity. Given this query on “rights,” for example, T59, in blue at the top right, may be most relevant—with “rights” as the most statistically significant keyword—but is also relatively rare in the entire corpus. Topic 40, on the other hand, which deals with civil rights, has “rights” as a much less meaningful keyword, yet is extremely common in this anti-slavery newspaper archive. Each of these topics holds significance for the scholar, but in different ways.

This interface aims for a balance between figure and ground, ordering the topics by relevance to the query, while sizing them by their overall importance. We envision a notion of “query” that is broad and inclusive: essentially, any method for selecting subsets of text can function as a query in this framework, since we can compute relevance and overall popularity. At the same time, the user can elect to focus on a narrower subset of topics if desired—for example, focusing on only those topics that deal with issues of women’s rights. By testing and refining increasingly complex queries on increasingly refined topic sets, the user can move seamlessly from an initial exploratory phase to more specific investigatory questions.

Of course, as with the first prototype, visualization necessitates tradeoffs. In this case, we have sacrificed Prototype I’s ability to explore fine-grained combinations of topics in favor of a clearer understanding of the overall shifts in relevance and popularity over time. The appropriateness of this design decision is motivated by our sense of what questions will be most salient in the analysis of a newspaper corpus, but only real user experiences can provide the definitive answer on which design is most useful. We return to this issue, and our plans for *in situ* usage analysis, in the sections on next steps and conclusions.

Interpreting Topics

Thus far, we have described how different conceptions of space can illuminate the topic model and its relationship to the archive. These visualizations focus on the relationships *between* topics, and on the change in the thematic landscape over time. But full support for the thematic analysis of an archive also requires a "micro" view on individual topics. This view can help to answer questions about how the words *within* each topic relate to each other, as well as about how the topic relates to other relevant metadata—such as author, editor, or place of publication—not just time (Brown et al., 2012).

Recall that topics are defined by sets of words, with the implicit assumption that each word has a single meaning across all usage contexts. However, humanities scholarship commonly entails a sensitivity to shifting meanings and uses. A scholar may wonder, for instance, how women's "rights" (as indicated by the keyword in T59) were described in relation to the legal "rights" featured in T40. She may ask if the rhetoric of one borrowed from the other, or if the use of the word "rights" changed when it was employed to describe women's rights in the wake of the struggle to end slavery. Again, the scholar seeks to know more than what can be inferred by the topic model alone. We propose to link LDA's high-level thematic analysis with visualizations that drill down to the level of individual examples. Building on the traditional keyword-in-context (KWIC) models (Rockwell 2003, Rockwell et al. 2010), we are developing a computational algorithm for selecting contexts that are both strongly associated with each topic of interest (for example, the contexts for "rights" in T40 and T59), while simultaneously revealing the full range of thematic possibilities within each topic (see "Next Steps," below).

While the range of connotations of individual words in a topic presents one kind of interpretive challenge, the topics themselves can at times present another: when a topic includes words associated with seemingly divergent themes. In T56, the scholar might observe a (seemingly) obvious connection, for the nineteenth-century, between words that describe Native Americans and those that describe nature. However, unlike the words "antelope" or "hawk," the words "tiger" and "hyena," also included in the topic, do not describe animals indigenous to North America. Does an explanation lie in a figurative vocabulary for describing native peoples? Or is this collection of words merely an accident of statistical analysis, a result of being built on a randomized algorithm? As noted by Chuang et al (2012), such questions lead inevitably to considerations of *trust*—why should the user trust that the algorithm was right to group hawks with hyenas?

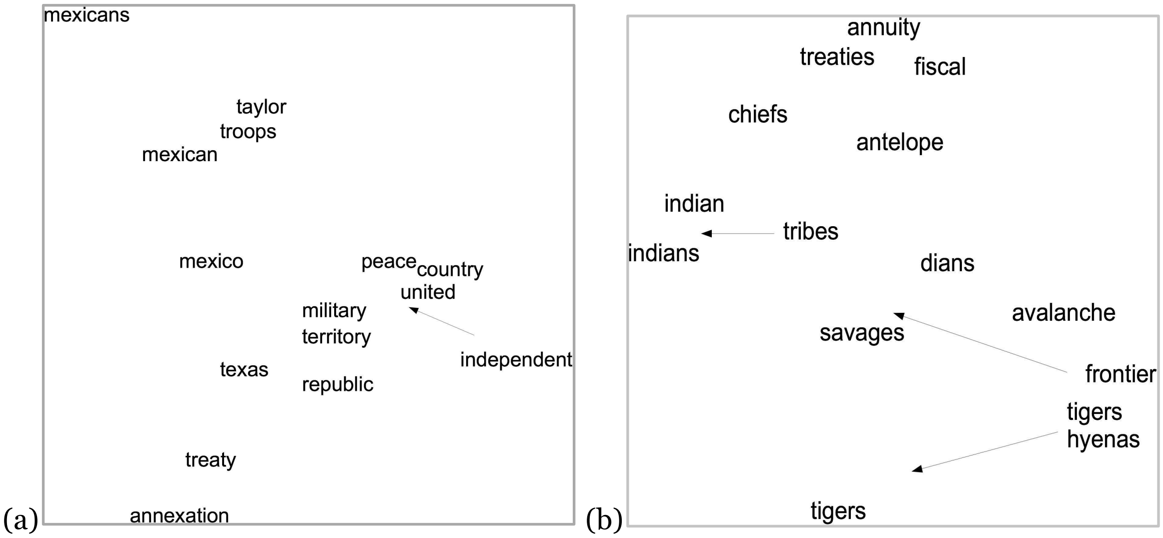


Figure 3:Multidimensional scaling of topics 55 (left) and 56 (right).
Arrows are used to show the position of words that would otherwise be occluded.

The above figures employ a spatial visualization using multidimensional scaling (Cox and Cox, 2010) to position the keywords for each topic according to their contextual similarity. Multidimensional scaling is an algorithm for spatially locating instances (in this case, words), such that similar instances are near each other, and distinct instances are further apart. Here, this means that the projection of words into space will be governed by whether they co-occur in the same sets of documents. In Figure 3b, the terms "indian", "indians", and "tribes" are located apart from "hyena", "tiger", and "tigers", which are themselves closely associated. This suggests a close connection—a high rate of co-occurrence—for the pairs of words *within* each group, and a relatively weak connection—a low rate of co-occurrence—for the pairs of words *between* the two groups. For comparison, Figure 3a displays the spatial visualization for a topic relating to the Mexican-American War, in which terms related to the conduct of the war ("Taylor", "troops") are spatially distinguished from those related to its outcome ("treaty", "annexation"). These views provide a visual alternative to the well-worn topic keyword list, allowing the researcher to spatially explore the complex network of lexical semantic relationships wrapped up within each topic.

Related Work

The work presented in this paper represents the culmination of the first phase of a larger project aimed at supporting the interactive exploration and visualization of archival collections. In this way, we place our work in the lineage of digital humanities tools designed for text analysis. The oldest and most widely used of these tools is Voyant (2003; 2012), which was conceived to enhance reading through lightweight text analytics such as word frequency lists, frequency distribution plots, and KWIC displays. More recently, WordSeer (2013) has built upon the basic functionality modeled by Voyant, incorporating automated syntactic analysis to identify part-of-speech sequences and grammatical dependency relations. Like topic modeling, automated syntactic analysis can also be seen as a method that improves upon more basic word-frequency models by adding context: syntactic analysis accounts for context at the level of individual sentences, while topic analysis accounts for context across the entire document. We thus see WordSeer as a peer project that complements our own.

As previously mentioned, the digital humanities have seen a proliferation of topic modeling analysis, including several topic model browsers. Andrew Goldstone developed a browser to explore various facets of the topics generated from a corpus of journal articles from *PMLA* (2013). The InPhO Topic Explorer (Murdock et al. 2014) employs a comparative perspective, facilitating the exploration of individual documents within a corpus in terms of their topic distribution. However, the structure of each of these browsers is determined by the topic model, while our browser attempts to emulate the process of conducting humanities research. While we do not view TOME as a general-purpose topic model browser, there have been many efforts in recent years, outside the humanities, to provide interactive visualizations for topic models. In this regard, our project draws on prior research on spatial visualizations (Alsakran et al., 2012; Eisenstein et al., 2012) and their limitations (Chuang et al., 2012), as well as work on timeline-based interfaces to topic models (Cui et al., 2011; Wei et al., 2010). A key difference from this work is our focus on the specific sensemaking needs of humanities scholars. Our approach is thus aligned with Chuang et al. (2012), but they focus on the *domain* of scholarship while we focus on the *task* of scholarship.

We note, finally, an array of digital humanities projects that make use of historical newspapers and related archives. The Trove project (2014) encompasses an array of tools designed to navigate the archives of the National Library of Australia, with a special emphasis on historical newspapers. The Mapping Texts (2011) project employs interactive visualizations, including some of topic models, to assess newspaper quality and language patterns. 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. Mining the Dispatch (2013) employs topic modeling in order to produce analyses of historical newspapers. Finally, Viral Texts (2013) identifies instances of textual reuse within a large set of historical newspapers. Taken together, these projects attest to the richness of this archival material.

Conclusions

Our work on TOME is premised on the assumption that what is needed to advance scholarship in the digital humanities is not merely a new algorithm that confirms prior humanities research, nor a new humanistic domain for the application of existing algorithms. Rather, we seek to address the higher-level question of what can be gained from bringing together the humanities with computation through interdisciplinary collaboration. Pursuing this question has led us to the intersections of automated thematic analysis, sensemaking, and interactive visualization. Some of the challenges that we have encountered are characteristic of any research on text visualization: text is an extraordinarily high-density carrier of information, and the problem of how to transform it into an accurate and efficient spatial representation has occupied visualization researchers for decades (Wise, 1999). Other challenges are more unique to the humanities, in that the humanities scholar is exceptionally attuned to subtle shifts in meaning and context, shifts which are difficult to convey without recourse to the text itself. Our research has nevertheless led to two high-level designs for visualization of thematic shifts in newspaper corpora, each of which provide a lens on the evolution of topics over time. In addition, we have identified a set of peripheral visual affordances that, once implemented, may prove critical for humanities scholarship.

TOME is designed to address the problems of understanding and exploring topic models. A related literature tackles the issue of how to incorporate user input in building the topic model itself, which involves setting numerical parameters such as the number of topics, fine-tuning lists of excluded terms (stopwords), removing "junk" topics, and so on. Chuang et al (2013) employ the method of small multiples to visualize the sensitivity of the topic modeling algorithm to changes in the number of topics and other parameters, allowing the user to rapidly explore a large, multi-dimensional space of possible topic models. More ambitiously, Hu et al (2014) enable the user to provide feedback about which topics are meaningful, and which conflate multiple underlying concepts; their interactive topic modeling algorithm then incorporates this feedback to produce better topics. Our visualizations are designed to support deep exploration of an individual topic model, but they could be integrated into a larger framework to support iterative re-training of the model. For our corpus, we were able to obtain high quality topics on the first try, with minimal fine-tuning of the model parameters, stopword lists, and preprocessing.

Next Steps

Our plans for continuing the project involve both short-term and long-term goals. In the short term, our goals are twofold: to document the current code so that it can be more easily employed and adapted for use in other contexts; and to ensure that the TOME section of the DH Lab website reflects all presentations and publications related to the project.

In the long term, we plan to pursue the outstanding research questions (and related tool development) through a combination of internal funding mechanisms and structures and external grant funding. In particular, we plan to focus on further refinement of the visualization framework and its accompanying affordances. We are particularly interested in more tightly linking the KWIC view with the topic model by showing contexts that vary as much as possible across topics.

We also plan to hold annual Digital Humanities "Hack Days," in addition to the hack day held last year (described in "Appendix II"). These hack days, run by the co-PI (Eisenstein), allow graduate and undergraduate students in computational linguistics to experiment with datasets and research questions in new domains. We will continue to employ the abolitionist newspaper dataset, advancing some of the algorithmic work required to pursue the open questions identified above.

Our most pressing goal, however, is to gather *in situ* studies of TOME as a component of real humanities research. The experiences of scholars invested in the abolitionist newspapers that constitute our principal dataset will highlight the strengths and weaknesses of the design decisions that we have already made, and point towards more generalizable applications of text mining and visualization for the humanities. Our hope is that, in facilitating the exploration and investigation of substantive research questions from the nineteenth century, we will push forward the development of twenty-first century tools and techniques.

Appendix 1: Project Personnel and Activities

Aside from the PI and co-PI, the team included three research assistants: one master's student in Georgia Tech's Digital Media program, and two undergraduates (one Computational Media major, and the other computer science). Our specific activities involved:

- Researching and prototyping potential interface designs and libraries for their implementation, focusing specifically on the challenges of incorporating text analytics in humanities scholarship.
- Meeting weekly with a team of graduate and undergraduate research assistants (one graduate student and one undergraduate in the first year; one undergraduate in the second).
- Implementing topic modeling algorithms and their associated visualizations. This involved coding in Python and JavaScript (primarily the D3.js visualization library).
- Compiling two datasets: the first from the *Chronicling America* website (described above), and the second from Accessible Archives (http://www.accessible-archives.com/). The latter dataset is manually keyed, and more completely covers the newspapers of interest than the *Chronicling America* data. Due to the time required to negotiate a license with Accessible Archives, however, we did not obtain it in time to incorporate it in our published research, although it was the topic of a DH "Hack Day" in Spring 2015 and, as we have renewed our licensing agreement with them, it will remain a focus of our research.
- Documenting progress and disseminating results in the form of one journal article, three conference/workshop presentations, and five invited talks.

Each of these publications are available (or summarized in blog form) on the Georgia Tech DH Lab website: http://dhlab.lmc.gatech.edu/tome/

In addition, code associated with the TOME project is publicly available on GitHub: https://github.com/GeorgiaTechDHLab/TOME

Appendix II: Educational Impact and Outreach

TOME has played a substantial educational role in both humanities and computer science courses at Georgia Tech, as students enrolled in courses taught by the PI and Co-PI have been exposed the progress and results of this research. The PI (Klein) has shared the TOME project with students enrolled in LMC 3314 ("Studies in Communication and Culture") as an example of humanities-focused visualization (Spring 2013, enrollment 29; Spring 2016, enrollment TBD), as well as in LMC 8801/CS 8001 ("Special Topics in Information Design and Technology"), an interdisciplinary graduate seminar focused on information visualization research (Spring 2015, enrollment approx. 10; Fall 2015, enrollment approx. 15). Results and progress from the research were also presented as examples of the role of topic models and the importance of visualization in CS 4650 and CS 7650 ("Natural Language Understanding"), a yearly course offered by the Co-PI (Eisenstein), with current enrollment of 86 students. They were also presented in CS 8803-CSS ("Computational Social Science"), a graduate seminar taught by Eisenstein in the spring semester of 2014 and 2015.

An additional internal audience for this project were the participants in the 2015 Digital Humanities "Hack Day", in which approximately fifteen literary scholars and computer scientists, ranging from undergraduates to professors, came together to apply text analytics and visualization techniques to the dataset of antislavery newspapers gathered as part of this project. The participants formed small teams, and their projects included new topic model formulations for historical texts, named entity recognition and other information extraction methods, exploration of new visualization techniques for historical text data.

Both Klein and Eisenstein have been invited to speak on the project at universities including: University of Saint Andrews, Scotland (Eisenstein, Computer Human Interaction visiting speaker series 2013); Bucknell University (Klein, Program in Comparative Humanities, 2015); University of Oklahoma (Klein, Presidential Dream Course Speaker Series, 2015); Emory University (Klein, Institute for Quantitative Theory and Methods, 2015). These audiences included humanities scholars and computer science researchers.

Works Cited

Alsakran, J., Chen, Y., Luo, D., Zhao, Y., Yang, J., Dou, W., & Liu, S. (2012). Real-time visualization of streaming text with a force-based dynamic system. *IEEE computer graphics and applications*, *32*(1), 34-45.

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

Brown, T., Baldridge, J., Esteva, M., & Xu, W. (2012). The substantial words are in the ground and sea: Computationally linking text and geography. *Texas Studies in Literature & Language*, *54*(3), 324-339.

Card, S. K., Mackinlay, J. D., & Shneiderman, B. (Eds.). (1999). *Readings in information visualization: using vision to think*. Morgan Kaufmann.

Chuang, J., Ramage, D., Manning, C., & Heer, J. (2012). Interpretation and trust: Designing model-driven visualizations for text analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 443-452). ACM.

Chuang, J., Gupta, S., Manning, C., & Heer, J. (2013). Topic Model Diagnostics: Assessing Domain Relevance via Topical Alignment. In *Proceedings of the International Conference on Machine Learning* (pp. 612-620). JMLR: W&CP volume 28.

Clytus, R. (2007). *Envisioning Slavery: American Abolition and the Primacy of the Visual.* Yale UP.

Cox, T. F., & Cox, M. A. (2010). *Multidimensional scaling*. CRC Press.

Cui, W., Liu, S., Tan, L., Shi, C., Song, Y., Gao, Z., *et al.* (2011). Textflow: Towards better understanding of evolving topics in text. *IEEE Transactions on Visualization and Computer Graphics*, *17*(12), 2412-2421.

Cutting, D. R., Karger, D. R., Pedersen, J. O., & Tukey, J. W. (1992). Scatter/gather: A cluster-based approach to browsing large document collections. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 318-329). ACM.

Davis, D.B. (1997). *Antebellum American Culture: An Interpretive Anthology*. Penn State UP.

Drucker, J. (2011). "Humanities Approaches to Graphical Display." *Digital Humanities Quarterly* 5(1).

Drucker J. (2014). *Graphesis: Visual Forms of Knowledge Production*. Harvard UP.

Dudden, F. (2011). *Fighting Chance: The Struggle Over Woman Suffrage and Black Suffrage in America*. Oxford UP.

Ferster, B. (2012). *Interactive Visualization: Insight Through Inquiry*. The MIT Press.

Gelman, A. (2004). Exploratory data analysis for complex models. *Journal of Computational and Graphical Statistics*, *13*(4).

Golden, V. (2014). *Consolidated Listings of Ohio Newspapers*. Manuscript in preparation. American Antiquarian Society.

Goldstone, A. (2014). Topic Model Browser. http://agoldst.github.io/dfr-browser/demo/

Goldstone, A., & Underwood, T. (2012). "What Can Topic Models of PMLA Teach Us About the History of Literary Scholarship?" *Journal of Digital Humanities* 2(1).

Goldstone, A., & Underwood, T. (2014). The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us. *New Literary History*, forthcoming.

Hu, Y., Boyd-Graber, J., Satinoff, B., & Smith, A. (2014). Interactive Topic Modeling. *Machine Learning* 95(3), 423-469.

Jockers, M. (2013). *Macroanalysis: Digital Methods and Literary History*. U Illinois P.

Jockers, M. L., & Mimno, D. (2013). Significant themes in 19th-century literature. *Poetics*, *41*(6), 750-769.

Karcher, C. (1997). *The First Woman in the Republic: A Cultural Biography of Lydia Maria Child*. Duke UP.

Kwok, J. T. and Adams, R. P. (2012). "Priors for Diversity in Generative Latent Variable Models." *Advances in Neural Information Processing Systems*.

McCallum, A. K. (2002). "MALLET: A Machine Learning for Language Toolkit."

Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., *et al.* (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331* (6014), 176-182.

Muralidharan, A. and Hearst, M. A. (2012). Supporting Exploratory Text Analysis in Literature Study, *Literary and Linguistic Computing*, 28(2), 283-295.

Murdock, J. et al. (2014). InPhO Topic Explorer. http://inphodata.cogs.indiana.edu/

Nelson, R. (2013). Mining the Dispatch. http://dsl.richmond.edu/dispatch/

Newman, D. J., & Block, S. (2006). Probabilistic topic decomposition of an eighteenth-century American newspaper. *Journal of the American Society for Information Science and Technology*, *57*(6), 753-767.

Rhody, L. (2012). "Topic Modeling and Figurative Language." *Journal of Digital Humanities* 2 (1).

Rhody, L. (2012). "Some Assembly Required: Understanding and Interpreting Topics in LDA Models of Figurative Language." Lisa @ Work.

Rockwell, G. (2003). What is text analysis, really? *Literary and Linguistic Computing*, 18(2), 209-219.

Rockwell, G., Sinclair, S.G., Ruecker, S., and Organisciak, P. (2010). Ubiquitous text analysis. *Poetess Archive Journal*, 2(1), 1-19.

Russell, D. M., Stefik, M. J., Pirolli, P., & Card, S. K. (1993). The cost structure of sensemaking. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems* (pp. 269-276). ACM.

Sherratt, T. and the National Library of Australia. (2014). Trove. http://trove.nla.gov.au/

Sinclair, S., Rockwell, G., et al. (2012). Voyant Tools. http://voyant-tools.org/

Smith, D., Cordell, R., and Dillon, E.M. (2013). “Infectious Texts: Modeling Text Reuse in Nineteenth-Century Newspapers.” In *Proceedings of the Workshop on Big Humanities*. IEEE.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley.

Underwood, T. (2011). “The Differentiation of Literary and Nonliterary Diction, 1700-1900.” The Stone and the Shell.

Wei, F., Liu, S., Song, Y., Pan, S., Zhou, M. X., Qian, W., et al. (2010). TIARA: a visual exploratory text analytic system. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 153-162). ACM.

Wise, J. A. (1999). The ecological approach to text visualization. *JASIS*, *50*(13), 1224-1233.

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)*, June 2011. 96-104

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.