Exploratory Thematic Analysis for Historical Newspaper Archives

Eisenstein, Jacob
Georgia Institute of Technology, United States of America

Sun, Iris
Georgia Institute of Technology, United States of America

Klein, Lauren F.
lauren.klein@lmc.gatech.edu
Georgia Institute of Technology, United States of America

Introduction

On July 19th, 1848, 300 concerned United States citizens gathered in Seneca Falls, New York, for the women's rights convention that would culminate in the signing of the Declaration of Rights and Sentiments, the first major document (in the US) to call for women's right to vote. In *The North Star*, Frederick Douglass, the former slave turned abolitionist, extolled the event as a "grand movement for attaining the civil, social, political, and religious rights of women" (1848). In the *Oneida Whig*, the same event was ridiculed as the "most shocking and unnatural event ever recorded in the history of womanity" (1848). As demonstrated by these contradictory accounts, published opinions varied greatly -- about the women's rights movement in the nineteenth-century United States, and about current events generally conceived. Large-scale digitization projects have increasingly enabled humanities scholars to search newspapers, such as those just cited, for significant words and phrases. But exploring more open-ended questions such as, "How did the discourse surrounding women's rights in the United States change in the wake of the 1848 Seneca Falls Convention?" or "Did the women's rights movement borrow language from the nation's contemporaneous anti-slavery campaign?" remains a challenge. Synthesizing current research on exploratory data analysis with techniques from the fields of computational linguistics and data visualization, we propose a new set of methods to assist humanities scholars in computationally-assisted exploratory research.

Background and Overview

Exploratory data analysis (EDA) has played a fundamental role in quantitative research since at least the 1970s (Tukey 1977). In comparison to formal hypothesis testing, exploratory data analysis is more open-ended, and is meant to help the researcher develop a general sense of the properties of the dataset before embarking on more specific inquiries (Russell, Stefik, Pirolli, and Card, 1993). EDA typically 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. More recently, Andrew Gelman (2004) has argued that EDA should be interwoven with formal statistical modeling, facilitating an iterative design process driven by experimenter insight.

The questions about women's rights, posed above, suggest the potential of EDA for humanities research-- a possibility also noted by Muralidharan and Hearst (2012). That team employs automatic syntactic analysis to identify and visualize recurring grammatical patterns, which, when combined with document metadata, reveals insights at the sentence level. By contrast, we combine metadata with techniques such as topic modeling in order to reveal insights at the document level. Inspired by the increasing use of topic models to make literary and cultural arguments (Underwood 2012, Rhody 2012, and Jockers 2013), we ask how the *exploratory thematic analysis* of documents might be incorporated into the initial phase of humanities research.

Our approach encompasses both traditional topic models and innovative visualizations, as well as alternative computational techniques targeted at the questions that topic modeling raises but leaves unanswered. By designing new visualizations and text-mining algorithms within the context of a specific, humanities-driven research effort, we hope to prototype a new mode of multi-disciplinary scholarship that will facilitate the iterative research methodology advocated by Gelman (2004). Specifically, we aim to facilitate the thematic exploration of document archives as a precursor to more informed keyword searching, more sustained close reading, and more systematic evidence gathering.

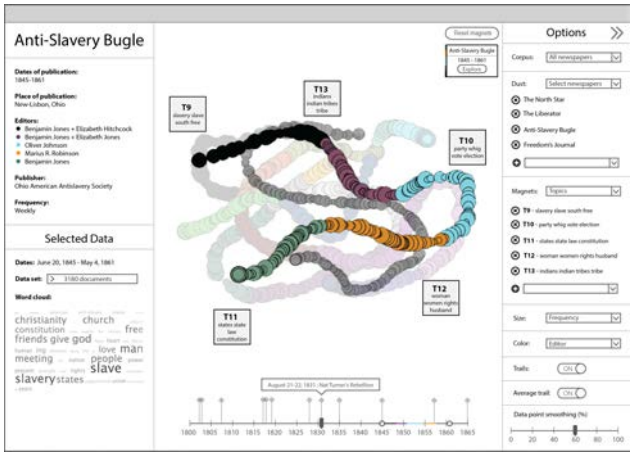
User Scenario and Interface Prototypes

Our focus is on a set of abolitionist newspapers from the nineteenth-century United States, in which antislavery advocates mounted moral, social, and political arguments in favor of emancipation. These newspapers present a particularly compelling dataset for thematic analysis, as similar ideas were purportedly framed differently by (and for) women and men (Dudden 2011). Here, we focus on one newspaper, *The Anti-Slavery Bugle,* published in New Lisbon, Ohio, between 1845 and 1861. Significantly, it was the source of much reprinting (Golden 2013), and underwent several shifts in editorial control.

Standard LDA topic analysis (MALLET; McCallum 2002) with 100 topics and standard parametrization reveals a number of topics that might intrigue a scholar in the initial phases of research, including:

- T40: states state law constitution tho government power united laws congress **rights**people con ohio tion act union question property
- T56: indians indian tribes tribe chiefs frontier dian treaties tiger hawk antelope annuity fiscal lllack hyenas tigers dians avalanche savages
- T59: woman women **rights**husband wife sex sho marriage property married mrs female legal sphere equality estate social duties sexes

Topic 59 (T59) suggests that the *Bugle* may offer insight into the relationship between the antislavery movement and the nascent drive for women's rights. The accompanying metadata reveals that the newspaper was co-edited by a woman between 1845 and 1849; however, this topic peaked in the late 1850s (a time when the women's rights movement was ascendant). At this point, we reach the limit of what can be learned from a topic model alone. We cannot easily answer, for instance, how the treatment of this topic in the *Bugle* might have differed from that of other newspapers; whether the editors of the *Bugle* were early advocates for the women's rights movement; or whether the peak in the late 1850s followed a national trend.

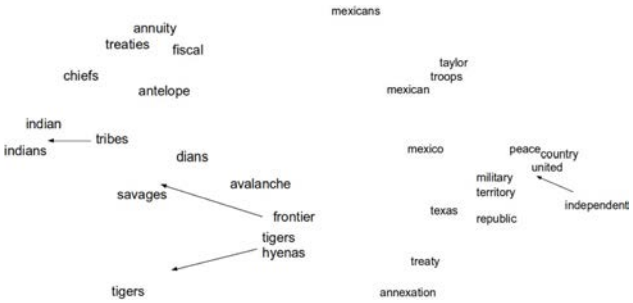


The above screenshot documents a prototype interface for the visualization and analysis of topic models that can begin to answer these questions. We apply a *dust-and-magnet*visualization (Yi et al., 2005), in which user-selected topics exert a magnetic "force" on individual issues of the

newspaper (represented as “dust”). The temporal trajectories of several newspapers are shown as “dust trails” in the visual space, with colors indicating the terms of different editorial teams, and with the *Bugle* highlighted so as to facilitate comparison with contemporaneous newspapers.

Next, we address the topics themselves. In such models, topics are defined by sets of words, with the assumption that each word has a single meaning across all usage contexts. However, much humanities scholarship 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 vs. legal rights. Again, the scholar seeks to know more than what can be inferred by LDA 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, 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.

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?



To address this question, we propose a spatial visualization using multidimensional scaling (Cox and Cox, 2010) to position the keywords for each topic according to their contextual similarity. As shown in the figure above-left, the terms “indian”, “indians”, and “tribes” are located apart from “hyena”, “tiger”, and “tigers”, which are themselves closely associated. The spatial layout suggests a relatively weak connection between these terms. For comparison, we also include the spatial visualization for a topic relating to the Mexican-American War, above-right, in which terms related to the conduct of the war (“Taylor”, “troops”) are spatially distinguished from those related to its outcome (“treaty”, “annexation”).

Conclusion and Next Steps

The goal of our ongoing work on exploratory thematic analysis is to provide a comprehensive set of algorithms and visualizations for understanding newspaper archives. Topic modeling is an important first step, but if we are to move beyond suggestive word lists in order to contribute to humanities scholarship, topic models must be linked to relevant metadata and concrete examples. Moreover, scholars must be provided with new visual modes that illuminate the substructures within the generalized themes that the topic model produces. Such techniques can reveal new insights about the transmission and circulation of ideas among social and political coalitions, and how the framing of these ideas relates to authors’ genders. By linking technical innovation with real humanistic inquiry, we hope to produce algorithms and visualizations that will meet the needs of substantive humanities research.

References

“Bolting Among the Ladies.” The Oneida Whig, August 1, 1848.

Cox, Trevor F., and Michael AA Cox (2010). *Multidimensional scaling*. CRC Press.

Douglass, Frederick (1848). *“The Rights of Women.”*The North Star, July 28.

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

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

Kwok, James T. and Ryan P. Adams. (2012). *“Priors for Diversity in Generative Latent Variable Models.”* Advances in Neural Information Processing Systems.

Rhody, L. (2012) *“Some Assembly Required: Understanding and Interpreting Topics in LDA Models of Figurative Langauge.”* Lisa @ Work.

McCallum, Andrew Kachites (2002). *"MALLET: A Machine Learning for Language Toolkit."*

Muralidharan, Aditi. and Hearst, Marti A. (2012), *Supporting Exploratory Text Analysis in Literature Study*, Literary and Linguistic Computing, 27 (4), Dec 24.

Russell, Daniel M., Mark J. Stefik, Peter Pirolli, and Stuart K. Card. (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.

Tukey, John Wilder (1977). *Exploratory Data Analysis.* Addison-Wesley.

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

Yi, Ji Soo, Rachel Melton, John Stasko, and Julie A. Jacko (2005). *"Dust & magnet: multivariate information visualization using a magnet metaphor."*Information Visualization 4, no. 4: 239-256.

Literary Canon and Digital Bibliographies: The Case of the United States

Ferrer, Carolina
ferrer.carolina@uqam.ca
Université du Québec à Montréal

In this research, I propose an alternative technique to the traditional method of constitution of the literary canon. Instead of basing the determination of the canon on different values, I scrutinize the *Modern Language Association International Bibliography* database in order to determine the most cited authors and literary works. Specifically, I study the literature of the United States of America. Thus, through the process of data mining, I obtain a sample of over 290,000 references that allows us to observe the chronological evolution and the linguistic distribution of the critical bibliography about USA literature. This quantitative technique yields a corpus of more than 100 titles and 100 writers that are cited more than 100 times in the database. Consequently, this bibliography is not the result of subjective selection criteria, but is based on the law of large numbers. Furthermore, this study shows that the quantitative analysis of bibliographic databases is an effective way to bring new light to the field of literary studies.