Topic Modeling on Historical Newspapers

Tze-I Yang
Dept. of Comp. Sci. & Eng.
University of North Texas
tze-iyang@my.unt.edu

Andrew J. Torget
Dept. of History
University of North Texas
andrew.torget@unt.edu

Rada Mihalcea
Dept. of Comp. Sci. & Eng.
University of North Texas
rada@cs.unt.edu

Abstract

In this paper, we explore the task of automatic text processing applied to collections of historical newspapers, with the aim of assisting historical research. In particular, in this first stage of our project, we experiment with the use of topical models as a means to identify potential issues of interest for historians.

1 Newspapers in Historical Research

Surviving newspapers are among the richest sources of information available to scholars studying peoples and cultures of the past 250 years, particularly for research on the history of the United States. Throughout the nineteenth and twentieth centuries, newspapers served as the central venues for nearly all substantive discussions and debates in American society. By the mid-nineteenth century, nearly every community (no matter how small) boasted at least one newspaper. Within these pages, Americans argued with one another over politics, advertised and conducted economic business, and published articles and commentary on virtually all aspects of society and daily life. Only here can scholars find editorials from the 1870s on the latest political controversies, advertisements for the latest fashions, articles on the latest sporting events, and languid poetry from a local artist, all within one source. Newspapers, in short, document more completely the full range of the human experience than nearly any other source available to modern scholars, providing windows into the past available nowhere else.

Despite their remarkable value, newspapers have long remained among the most underutilized historical resources. The reason for this paradox is quite simple: the sheer volume and breadth of information available in historical newspapers has, ironically, made it extremely difficult for historians to go through them page-by-page for a given research project. A historian, for example, might need to wade through tens of thousands of newspaper pages in order to answer a single research question (with no guarantee of stumbling onto the necessary information).

Recently, both the research potential and problem of scale associated with historical newspapers has expanded greatly due to the rapid digitization of these sources. The National Endowment for the Humanities (NEH) and the Library of Congress (LOC), for example, are sponsoring a nationwide historical digitization project, *Chronicling America*, geared toward digitizing all surviving historical newspapers in the United States, from 1836 to the present. This project recently digitized its one millionth page (and they project to have more than 20 million pages within a few years), opening a vast wealth of historical newspapers in digital form.

While projects such as *Chronicling America* have indeed increased access to these important sources, they have also increased the problem of scale that have long prevent scholars from using these sources in meaningful ways. Indeed, without tools and methods capable of handling such large datasets – and thus sifting out meaningful patterns embedded within them – scholars find themselves confined to performing only basic word searches across enormous collections. These simple searches can, indeed, find stray information scattered in unlikely

Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pages 96–104,
Portland, OR, USA, 24 June 2011. ©2011 Association for Computational Linguistics

places. Such rudimentary search tools, however, become increasingly less useful to researchers as datasets continue to grow in size. If a search for a particular term yields 4,000,000 results, even those search results produce a dataset far too large for any single scholar to analyze in a meaningful way using traditional methods. The age of abundance, it turns out, can simply overwhelm historical scholars, as the sheer volume of available digitized historical newspapers is beginning to do.

In this paper, we explore the use of topic modeling, in an attempt to identify the most important and potentially interesting topics over a given period of time. Thus, instead of asking a historian to look through thousands of newspapers to identify what may be interesting topics, we take a reverse approach, where we first automatically cluster the data into topics, and then provide these automatically identified topics to the historian so she can narrow her scope to focus on the individual patterns in the dataset that are most applicable to her research. Of more utility would be where the modeling would reveal unexpected topics that point towards unusual patterns previously unknown, thus help shaping a scholar's subsequent research.

The topic modeling can be done for any periods of time, which can consist of individual years or can cover several years at a time. In this way, we can see the changes in the discussions and topics of interest over the years. Moreover, pre-filters can also be applied to the data prior to the topic modeling. For instance, since research being done in the History department at our institution is concerned with the "U. S. cotton economy," we can use the same approach to identify the interesting topics mentioned in the news articles that talk about the issue of "cotton."

2 Topic Modeling

Topic models have been used by Newman and Block (2006) and Nelson (2010)[1] on newspaper corpora to discover topics and trends over time. The former used the probabilistic latent semantic analysis (pLSA) model, and the latter used the latent Dirichlet allocation (LDA) model, a method introduced by Blei et al. (2003). LDA has also been used by Griffiths and Steyvers (2004) to find research topic trends by looking at abstracts of scientific papers. Hall et al. (2008) have similarly applied LDA to discover trends in the computational linguistics field. Both pLSA and LDA models are probabilistic models that look at each document as a mixture of multinomials or topics. The models decompose the document collection into groups of words representing the main topics. See for instance Table 1, which shows two topics extracted from our collection.

Topic
worth price black white goods yard silk made ladies wool lot inch week sale prices pair suits fine quality
state states bill united people men general law government party made president today washington war committee country public york

Table 1: Example of two topic groups

Boyd-Graber et al. (2009) compared several topic models, including LDA, correlated topic model (CTM), and probabilistic latent semantic indexing (pLSI), and found that LDA generally worked comparably well or better than the other two at predicting topics that match topics picked by the human annotators. We therefore chose to use a parallel threaded SparseLDA implementation to conduct the topic modeling, namely UMass Amherst's MAchine Learning for LanguagE Toolkit (MALLET)[2] (McCallum, 2002). MALLET's topic modeling toolkit has been used by Walker et al. (2010) to test the effects of noisy optical character recognition (OCR) data on LDA. It has been used by Nelson (2010) to mine topics from the Civil War era newspaper *Dispatch*, and it has also been used by Blevins (2010) to examine general topics and to identify emotional moments from Martha Ballards Diary.[3]

3 Dataset

Our sample data comes from a collection of digitized historical newspapers, consisting of newspapers published in Texas from 1829 to 2008. Issues are segmented by pages with continuous text containing articles and advertisements. Table 2 provides more information about the dataset.

[1] `http://americanpast.richmond.edu/dispatch/`

[2] `http://mallet.cs.umass.edu/`

[3] `http://historying.org/2010/04/01/`

Property	
Number of titles	114
Number of years	180
Number of issues	32,745
Number of pages	232,567
Number of tokens	816,190,453

Table 2: Properties of the newspaper collection

3.1 Sample Years and Categories

From the wide range available, we sampled several historically significant dates in order to evaluate topic modeling. These dates were chosen for their unique characteristics (detailed below), which made it possible for a professional historian to examine and evaluate the relevancy of the results.

These are the subcategories we chose as samples:

- **Newspapers from 1865-1901:** During this period, Texans rebuilt their society in the aftermath of the American Civil War. With the abolition of slavery in 1865, Texans (both black and white) looked to rebuild their post-war economy by investing heavily in cotton production throughout the state. Cotton was considered a safe investment, and so Texans produced enough during this period to make Texas the largest cotton producer in the United States by 1901. Yet overproduction during that same period impoverished Texas farmers by driving down the market price for cotton, and thus a large percentage went bankrupt and lost their lands (over 50 percent by 1900). As a result, angry cotton farmers in Texas during the 1890s joined a new political party, the Populists, whose goal was to use the national government to improve the economic conditions of farmers. This effort failed by 1896, although it represented one of the largest third-party political revolts in American history.

  This period, then, was dominated by the rise of cotton as the foundation of the Texas economy, the financial failures of Texas farmers, and their unsuccessful political protests of the 1890s as cotton bankrupted people across the state. These are the issues we would expect to emerge as important topics from newspapers in this category. This dataset consists of 52,555 pages over 5,902 issues.

- **Newspapers from 1892:** This was the year of the formation of the Populist Party, which a large portion of Texas farmers joined for the U. S. presidential election of 1892. The Populists sought to have the U. S. federal government become actively involved in regulating the economy in places like Texas (something never done before) in order to prevent cotton farmers from going further into debt. In the 1892 election, the Populists did surprisingly well (garnering about 10 percent of the vote nationally) and won a full 23 percent of the vote in Texas. This dataset consists of 1,303 pages over 223 issues.

- **Newspapers from 1893:** A major economic depression hit the United States in 1893, devastating the economy in every state, including Texas. This exacerbated the problem of cotton within the states economy, and heightened the efforts of the Populists within Texas to push for major political reforms to address these problems. What we see in 1893, then, is a great deal of stress that should exacerbate trends within Texas society of that year (and thus the content of the newspapers). This dataset consists of 3,490 pages over 494 issues.

- **Newspapers from 1929-1930:** These years represented the beginning and initial onset in the United States of the Great Depression. The United States economy began collapsing in October 1929, when the stock market crashed and began a series of economic failures that soon brought down nearly the entire U. S. economy. Texas, with its already shaky economic dependence on cotton, was as devastated as any other state. As such, this period was marked by discussions about how to save both the cotton economy of Texas and about possible government intervention into the economy to prevent catastrophe. This dataset consists of 6,590 pages over 973 issues.

Throughout this era, scholars have long recognized that cotton and the economy were the dominating issues. Related to that was the rise and fall

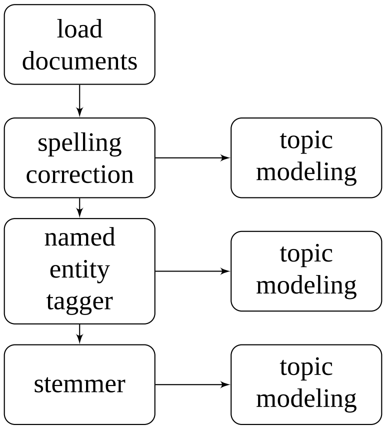

Figure 1: Work flow

of the Populist Party during the 1890s, as farmers sought to use the political system as a means of dealing with their economic problems. As such, we would expect to see these concerns as major (perhaps dominating) topics in the newspapers from the time.

3.1.1 "Cotton" data

Within the date ranges listed above, we also select all mentions of the topic "cotton" – as pertaining to possible discussion relevant to the "U. S. cotton economy." Cotton was the dominating economic force in Texas throughout this period, and historians have long recognized that issues related to the crop wielded tremendous influence on the political, social, and economic development of the state during this era. Problems related to cotton, for example, bankrupted half of all Texas farmers between 1865 and 1900, and those financial challenges pushed farmers to create a major new political party during the 1890s.

3.2 Data Processing

Before applying topic modeling on our data, some pre-processing steps were applied. Some challenges in processing the dataset come from errors introduced by the OCR processing, missing punctuations, and unclear separation between different articles on the same page. Multi-stage pre-processing of the dataset was performed to reduce these errors, as illustrated in Figure 1.

The first phase to reduce errors starts with spelling correction, which replaces words using the Aspell dictionary and de-hyphenates words split across lines. Suggested replacements are used if they are within the length normalized edit distance of the originals. An extra dictionary list of location names is used with Aspell.

Next, the spelling corrected dataset is run through the Stanford Named Entity Recognizer (NER).[4] Stanford NER system first detects sentences in the data then labels four classes of named entities: PERSON, ORGANIZATION, LOCATION, and MISCELLANEOUS (Finkel et al., 2005). The model used in conjunction with the tagger is provided by the software and was trained on the CoNLL 2003 training data using distributional similarity features. The output is then massaged so that entities with multiple words would stay together in the topic modeling phase.

Property	**# of Unique**	**# of Total**
LOC entities	1,508,432	8,620,856
ORG entities	6,497,111	14,263,391
PER entities	2,846,906	12,260,535
MISC entities	1,182,845	3,594,916
Named entities	12,035,294	38,739,698

Table 3: Properties of the newspaper collection after named entity recognition

Lastly, the words that are not tagged as named entities pass through an English stemmer while the named entities stay unchanged. We are using the Snowball stemmer.[5]

At the end of each of the pre-processing stage, we extract subsets from the data corresponding to the sample years mentioned earlier (1865-1901, 1892, 1893, and 1929-1930), which are then used for further processing in the topic modeling phase.

We made cursory comparisons of the outputs of the topic modeling at each of the three stages (spelling correction, NER, stemming). Table 4 shows sample topic groups generated at the three stages. We found that skipping the named entity tagging and stemming phases still gives comparable results. While the named entity tags may give us additional information ("dallas" and "texas" are locations), tagging the entire corpus takes up a large slice of processing time. Stemming after tagging

[4] http://nlp.stanford.edu/software/

[5] http://snowball.tartarus.org

Topic: spell
worth fort city texas county gazette tex special state company dallas time made yesterday night business line railroad louis
Topic: spell + NER
city county texas_location company yesterday night time today worth made state morning fort special business court tex dallas_location meeting
Topic: spell + NER + stemmer
state counti citi texas_location year ani time made worth fort peopl good line special tex land busi work compani

Table 4: Comparison of the three topic output stages: Each entry contains the top terms for a single topic

may collapse multiple versions of a word together, but we found that the stemmed words are very hard to understand such as the case of "business" becoming "busi". In future work, we may explore using a less aggressive stemmer that only collapses plurals, but so far the first stage output seems to give fairly good terms already. Thus, the rest of the paper will discuss using the results of topic modeling at the spelling correction stage.

4 Historical Topics and Trends

We are interested in automatically discovering general topics that appear in a large newspaper corpus. MALLET is run on each period of interest to find the top one general topic groups. We use 1000 iterations with stopword removal. An extra stopword list was essential to remove stopwords with errors introduced by the OCR process. Additionally, we run MALLET on the 1865-1901 dataset to find the top ten topic groups using 250 iterations.

In addition, we also find the topics more strongly associated with "cotton." The "cotton" examples are found by extracting each line that contains an instance of "cotton" along with a window of five lines on either side. MALLET is then run on these "cotton" examples to find the top general topic groups over 1000 iterations with stopword removal.

5 Evaluation and Discussion

The topic modeling output was evaluated by a historian (the second author of this paper), who specializes in the U.S.-Mexican borderlands in Texas and is an expert in the historical chronology, events, and language patterns of our newspaper collection. The evaluator looked at the output, and determined for each topic if it was relevant to the period of time under consideration.

The opinion from our expert is that the topic modeling yielded highly useful results. Throughout the general topics identified for our samples, there is a consistent theme that a historian would expect from these newspapers: a heavy emphasis on the economics of cotton. For example, we often see words like "good," "middling," and "ordinary," which were terms for evaluating the quality of a cotton crop before it went to market. Other common terms, such as "crop," "bale," "market," and "closed" (which suggests something like "the price *closed* at X") evoke other topics of discussion of aspects of the buying and selling of cotton crops.

Throughout the topics, market-oriented language is the overwhelming and dominate theme throughout, which is exactly what our expert expected as a historian of this region and era. You can see, for example, that much of the cotton economy was geared toward supplies the industrial mills in England. The word "Liverpool," the name of the main English port to where Texas cotton was shipped, appears quite frequently throughout the samples. As such, these results suggest a high degree of accuracy in identifying dominate and important themes in the corpus.

Within the subsets of these topics, we find more fine-grained patterns that support this trend, which lend more credence to the results.

Table 5 summarizes the results for each of the three analyzes, with accuracy calculated as follows: $Accuracy(\text{topics}) = \frac{\text{\# of relevant topics}}{\text{total \# of topics}}$ $Accuracy(\text{terms}) = \frac{\text{\# of relevant terms in all topics}}{\text{total \# of terms in all topics}}$. Tables 6, 7 and 8 show the actual analyzes.

5.1 Interesting Finding

Our historian expert found the topic containing "houston april general hero san" for the 1865-1901 general results particularly interesting and hypothesized that they may be referring to the Battle of San Jacinto. The Battle of San Jacinto was the final fight in the Texas Revolution of 1836, as Texas sought to free themselves from Mexican rule. On April 21, 1836, General Sam Houston led about 900

Topics	**Explanation**
black* price* worth* white* goods* yard* silk* made* lot* week ladies wool* inch* ladles* sale* prices* pair* suits* fine*	Reflects discussion of the market and sales of goods, with some words that relate to cotton and others that reflect other goods being sold alongside cotton (such as wool).
state* people* states* bill* law* made united* party* men* country* government* county* public* president* money* committee* general* great question*	Political language associated with the political debates that dominated much of newspaper content during this era. The association of the topic "money" is particularly telling, as economic and fiscal policy were particularly important discussion during the era.
clio worth mid city alie fort lino law lour lug thou hut fur court dally county anil tort iron	Noise and words with no clear association with one another.
tin inn mid tint mill* till oil* ills hit hint lull win hut ilia til ion lot lii foi	Mostly noise, with a few words associated with cotton milling and cotton seed.
texas* street* address* good wanted houston* office* work city* sale main* house* apply man county* avenue* room* rooms* land*	These topics appear to reflect geography. The inclusion of Houston may either reflect the city's importance as a cotton market or (more likely) the large number of newspapers from the collection that came from Houston.
worth* city* fort* texas* county* gazette tex* company* dallas* miss special yesterday night time john state made today louis*	These topics appear to reflect geography in north Texas, likely in relation to Fort Worth and Dallas (which appear as topics) and probably as a reflection that a large portion of the corpus of the collection came from the Dallas/Ft. Worth area.
houston* texas* today city* company post* hero* general* night morning york men* john held war* april* left san* meeting	These topics appear to an unlikely subject identified by the modeling. The words `Houston`, `hero`, `general`, `april` and `san` (perhaps part of `San Jacinto`) all fit together for a historian to suggest a sustained discussion in the newspapers of the April 1836 Battle of San Jacinto, when General Sam Houston defeated Santa Anna of Mexico in the Texas Revolution. This is entirely unexpected, but the topics appear to fit together closely. That this would rank so highly within all topics is, too, a surprise. (Most historians, for example, have argued that few Texans spent much time memorializing such events until after 1901. This would be quite a discovery if they were talking about it in such detail before 1901.)
man time great good men years life world long made people make young water woman back found women work	Not sure what the connections are here, although the topics clearly all fit together in discussion of the lives of women and men.
market* cotton* york* good* steady* closed* prices* corn* texas* wheat* fair* stock* choice* year* lower* receipts* ton* crop* higher*	All these topics reflect market-driven language related to the buying and selling cotton and, to a much smaller extent, other crops such as corn.
tube tie alie time thaw art ton ion aid ant ore end hat ire aad lour thee con til	Noise with no clear connections.

Table 6: 10 topic groups found for the 1865-1901 main set. Asterisks denote meaningful topic terms.

Period	**Topics**	**Explanation**
1865-1901	texas* city* worth* houston* good* county* fort* state* man* time* made* street* men* work* york today company great people	These keywords appear to be related to three things: (1) geography (reflected in both specific places like `Houston` and `Fort Worth` and more general places like `county`, `street`, and `city`), (2) discussions of people (`men` and `man`) and (3) time (`time` and `today`).
1892	texas* worth* gazette* city* tex* fort* county* state* good* march* man* special* made* people* time* york men days feb	As with the 1865-1901 set, these keywords also appear to be related to three things: (1) geography, (2) discussions of people and (3) time.
1893	worth* texas* tin* city* tube* clio* time* alie* man* good* fort* work* made street year men county state tex	As with the 1865-1901 set, these keywords also appear to be related to three things: (1) geography, (2) discussions of people and (3) time.
1929-1930	tin* texas* today* county* year* school* good* time* home* city* oil* man* men* made* work* phone night week sunday	As with the 1865-1901 set, these keywords also appear to be related to three things: (1) geography, (2) discussions of people and (3) time. The time discussion here appears to be heightened, and the appearance of economic issues for Texas (`oil`) makes sense in the context of the onset of the Great Depression in 1929-30.

Table 7: Main topics for years of interest for the main set

Period	**Topics**	**Explanation**
1865-1901	cotton* texas* good* crop* bales* county* york* houston* spot middling* year* corn* market* worth* oil* closed* special* ordinary* today	All market-oriented language that reflects all aspects of the cotton market, in particular the evaluation of cotton quality. The geography of New York (`york`) and Houston could reflect their importance in the cotton market or (just as important) sources of news and information (with Houston being a central producer of the newspapers in our corpus).
1892	cotton* bales* spot gazette* special* march middling* ordinary* steady* closed* futures* lots* good* texas* sales* feb low* ton* oil*	Market-oriented language that reflects, in particular, the buying and selling of cotton on the open market. The inclusion of February and March 1892, in the context of these other words associated with the selling of cotton, suggest those were important months in the marketing of the crop for 1892.
1893	cotton* ordinary* texas* worth* belt middling* closed* year bales* good* route* crop* city* cents* spot oil* corn* low* return*	Market-oriented language focused on the buying and selling of cotton.
1929-1930	cotton* texas* county crop* year good* today* york* points* oil* market* farm* made* seed* state* price* tin bales* july*	Market-oriented language concerning cotton. What is interesting here is the inclusion of words like `state`, `market`, and `price`, which did not show up in the previous sets. The market-language here is more broadly associated with the macro-economic situation (with explicit references to the `market` and `price`, which seems to reflect the heightened concern at that time about the future of the cotton market with the onset of the Great Depression and what role the `state` would play in that.

Table 8: Main topics for the `cotton` subset

		Accuracy	
	Topic Groups	Topics	Terms
General	Ten for 1865-1901	60%	45.79% (74.56%)
	One for 1865-1901	100%	73.68%
	One for 1892	100%	78.95%
	One for 1893	100%	63.16%
	One for 1929-1930	100%	78.95%
Cotton	One for 1865-1901	100%	89.47%
	One for 1892	100%	84.21%
	One for 1893	100%	84.21%
	One for 1929-1930	100%	84.21%

Table 5: Accuracy of topic modeling: In parenthesis is the term accuracy calculated using relevant topics only.

Texans against Mexican general Antonio Lopez de Santa Anna. Over the course of an eighteen minute battle, Houston's forces routed Santa Anna's army. The victory at San Jacinto secured the independence of Texas from Mexico and became a day of celebration in Texas during the years that followed.

Most historians have argued that Texas paid little attention to remembering the Battle of San Jacinto until the early twentieth century. These topic modeling results, however, suggest that far more attention was paid to this battle in Texas newspapers than scholars had previously thought.

We extracted all the examples from the corpus for the years 1865-1901 that contain ten or more of the top terms in the topic and also contain the word "jacinto". Out of a total of 220 snippets that contain "jacinto", 125 were directly related to the battle and its memory. 95 were related to other things. The majority of these snippets came from newspapers published in Houston, which is located near San Jacinto, with a surge of interest in the remembrance of the battle around the Aprils of 1897-1899.

6 Conclusions

In this paper, we explored the use of topical models applied on historical newspapers to assist historical research. We have found that we can automatically generate topics that are generally good, however we found that once we generated a set of topics, we cannot decide if it is mundane or interesting without an expert and, for example, would have been oblivious to the significance of the San Jacinto topic. We agree with Block (2006) that "topic simulation is only a tool" and have come to the conclusion that it is essential that an expert in the field contextualize these topics and evaluate them for relevancy.

We also found that although our corpus contains noise from OCR errors, it may not need expensive error correction processing to provide good results when using topic models. We may explore combining the named entity tagged data with a less aggressive stemmer and, additionally, evaluate the usefulness of not discarding the unstemmed words but maintaining their association with their stemmed counterpart.

Acknowledgment

We would like to thank Jon Christensen and Cameron Blevins from Stanford, who have been working with us on the larger project, "Mapping Historical Texts: Combining Text-Mining and Geo-Visualization to Unlock the Research Potential of Historical Newspapers", which subsumes the work we have presented in this paper. This work has been partly supported by the NEH under Digital Humanities Start-Up Grant (HD-51188-10). Any views, findings, conclusions or recommendations expressed in this publication do not necessarily represent those of the National Endowment for the Humanities.

References

[Blei et al.2003] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

[Blevins2010] Cameron Blevins. 2010. Topic Modeling Martha Ballard's Diary.

[Block2006] Sharon Block. 2006. Doing More with Digitization: An Introduction to Topic Modeling of Early American Sources. *Common-Place*, 6(2), January.

[Boyd-Graber et al.2009] Jonathan Boyd-Graber, Jordan Chang, Sean Gerrish, Chong Wang, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems*.

[Finkel et al.2005] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pages 363–370.

[Griffiths and Steyvers2004] Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5228.

[Hall et al.2008] David Hall, Daniel Jurafsky, and Christopher Manning. 2008. Studying the History of Ideas Using Topic Models. In *Proceedings from the EMNLP 2008: Conference on Empirical Methods in Natural Language Processing*, October.

[McCallum2002] Andrew Kachites McCallum. 2002. MALLET: A Machine Learning for Language Toolkit.

[Nelson2010] Robert K. Nelson. 2010. Mining the *Dispatch*.

[Newman and Block2006] David J. Newman and Sharon Block. 2006. Probabilistic topic decomposition of an eighteenth-century American newspaper. *Journal of the American Society for Information Science and Technology*, 57(6):753–767.

[Walker et al.2010] Daniel D. Walker, William B. Lund, and Eric K. Ringger. 2010. Evaluating models of latent document semantics in the presence of OCR errors. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 240–250. Association for Computational Linguistics.