MAPPING TEXTS: COMBINING TEXT-MINING AND GEO-VISUALIZATION TO UNLOCK THE RESEARCH POTENTIAL OF HISTORICAL NEWSPAPERS

A White Paper for the National Endowment for the Humanities

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

Rada Mihalcea
University of North Texas
rada.mihalcea@unt.edu

Jon Christensen
Stanford University
jonchristensen@stanford.edu

Geoff McGhee
Stanford University
gmcghee@stanford.edu

In September 2010, the University of North Texas (in partnership with Stanford University) was awarded a National Endowment for the Humanities Level II Digital Humanities Start-Up Grant (Award #HD-51188-10) to develop a series of experimental models for combining the possibilities of text-mining with geospatial mapping in order to unlock the research potential of large-scale collections of historical newspapers. Using a sample of approximately 230,000 pages of historical newspapers from the *Chronicling America* digital newspaper database, we developed two interactive visualizations of the language content of these massive collections of historical documents as they spread across both time and space: one measuring the quantity and quality of the digitized content, and a second measuring several of the most widely used large-scale language pattern metrics common in natural language processing work. This white paper documents those experiments and their outcomes, as well as our recommendations for future work.

Project Website:
http://mappingtexts.org

TABLE OF CONTENTS

PROJECT OVERVIEW

"Mapping Texts" is a collaborative project between the University of North Texas and Stanford University whose goal has been to develop a series of experimental new models for combining the possibilities of text-mining and geospatial analysis in order to enable researchers to develop better quantitative and qualitative methods for finding and analyzing meaningful language patterns embedded within massive collections of historical newspapers.

The broader purpose behind this effort has been to help scholars develop new tools for coping effectively with the growing challenge of doing research in the age of abundance, as the rapid pace of mass digitization of historical sources continues to pick up speed. Historical records of all kinds are becoming increasingly available in electronic forms, and there may be no set of records becoming available in larger quantities than digitized historical newspapers. The *Chronicling America* project (a joint endeavor of the National Endowment for the Humanities and the Library of Congress), for example, recently digitized its one millionth historical newspaper page, and projects that more than 20 million pages will be available within a few years. Numerous other digitization programs, both in the public and for-profit sectors, are also digitizing historical newspapers at a rapid pace, making hundreds of millions of words from the historical record readily available in electronic archives that are reaching staggering proportions.

What can scholars do with such an immense wealth of information? Without tools and methods capable of handling such large datasets—and thus sifting out meaningful patterns embedded within them—scholars typically find themselves confined to performing only basic word searches across enormous collections. While such basic searches can, indeed, find stray information scattered in unlikely places, they become increasingly less useful as datasets continue

to grow in size. If, for example, 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 researchers, as the sheer volume of available digitized historical newspapers is beginning to do.

Efforts among humanities scholars to develop more effective methods for sifting through such large collections of historical records have tended to concentrate in two areas: (1) sifting for language patterns through natural language processing (usually in the form of text-mining), or (2) visualizing patterns embedded in the records (through geospatial mapping and other techniques).

Both methods have a great deal to offer humanities scholars. Text-mining techniques can take numerous forms, but at base they attempt to find—and often quantify—meaningful language patterns spread across large bodies of text. If a historian, for example, wanted to understand how Northerners and Southerners discussed Abraham Lincoln during the American Civil War, he or she could mine digitized historical newspapers to discover how discussions of Lincoln evolved over time in those newspapers (looking for every instance of the word "Lincoln" and the constellation of words that surrounded it). Visualization work, on the other hand, focuses on understanding the patterns in large datasets by visualizing those relationships in various contexts. Often this takes the form of mapping information—such as census and voting returns—across a landscape, as scholars seek to understand the meaning of spatial relationships embedded within their sources. A researcher who wanted to understand what U. S. census data can tell us about how populations over the last two centuries have shifted across North America, for example, might map that information as the most effective means of analyzing it.

The goal of our project, then, has been to experiment with developing new methods for discovering and analyzing language patterns embedded in massive databases by attempting to combine the two most promising and widely used methods for finding meaning in such massive collections of electronic records: text-mining and geospatial visualization. And to that end, we have also focused on exploring the records that are being digitized and made available to scholars in the greatest quantities: historical newspapers.

OUR DATA SET: THE NEWSPAPERS

For this project, we experimented on a collection of about 232,500 pages of historical newspapers digitized by the University of North Texas (UNT) Library as part of the National Digital Newspaper Program (NDNP)'s *Chronicling America* project. The UNT Library has spent the last several years collecting and digitizing surviving newspapers from across Texas, covering the late 1820s through the early 2000s. These newspapers were available to us—and anyone interested in them—through the *Chronicling America* site (http://chroniclingamerica.loc.gov/) and UNT's *Portal to Texas History* site (http://texashistory.unt.edu/). Working in partnership with the UNT library, we determined to use their collection of Texas newspapers as our experimental dataset for several reasons:

1. **With nearly a quarter million pages, we could experiment with scale.** Much of the premise of this project is built around the problem of scale, and so we wanted to work with a large enough dataset that scale would be a significant factor (while also keeping it within a manageable range).

2. **The newspapers were all digitized according to the standards set by the NDNP's *Chronicling America* project, providing a uniform sample**. The standards set by the NDNP (http://www.loc.gov/ndnp/guidelines/archive/NDNP_201113TechNotes.pdf) meant that whatever techniques we developed could be uniformly applied across the entire collection, and that our project work would also be applicable to the much larger collected corpus of digitized newspapers on the *Chronicling America* site.
3. **The Texas orientation of all the newspapers gave us a consistent geography for our visualization experiments**. Because we would be attempting to create visualizations of the language patterns embedded in the newspapers as they spread out across time and space, we needed to have a manageable geographic range. Texas, fortunately, proved to be large enough to provide a great deal of geographic diversity for our experiments, while also being constricted enough to remain manageable.

PROJECT GOALS

The focus of our work, then, was to build a series of interactive models that would experiment with methods for combining text-mining with visualizations, using text-mining to discover meaningful language patterns in large-scale text collections and then employ visualizations in order to make sense of them. By "model" we mean to convey all of the data, processing, text-mining, and visualization tools assembled and put to work in these particular processes of exploration, research, and sense-making. We also mean to convey a "model" that can be used by others for the particular datasets that we employed as well as other, similar datasets of texts that have significant temporal and spatial attributes.

Our original concept had been to build these new models around a series of particular research questions, since the long-term goal of our work is to help scholars sift these collections in order to better answer important questions within various fields of research. At a very early stage in our work, however, we realized that we needed to build better surveying tools to simply understand what research questions *could* be answered with the available digital datasets available to us. For example, we had originally planned to compare the differences in language patterns emanating from rural and urban communities (hoping to see if the concerns of those two differed in any significant way, and if that had changed over time). We soon realized, however, that before we could begin to answer such a question we would first need to assess how much of the dataset represented rural or urban spaces, and whether there was enough quantity and quality of the data from both regions to undertake a meaningful comparison.

We therefore shifted the focus of our models to take such matters into account. Because almost all research questions would first require a quantitative survey of the available data, we determined that the first model we built should plot the quantity and quality of the newspaper content. Such a tool would, we hoped, provide users with a deep and transparent window into the amount of information available in digitized historical newspapers—in terms of sheer quantity of data, geographic locations of that information, how much was concentrated in various time spans, and the like—in order to enable users of digitized historical newspapers to make more informed choices about what sort of research questions could, indeed, be answered by the available sources. We were, however, unwilling to abandon our original focus on developing *qualitative* assessments of the language embedded in digitized newspapers. Indeed, we remained committed to also developing a qualitative model of the newspaper collection that would reveal

large-scale language patterns, which could then complement and work in tandem with the quantitative model. Between these two models—the quantitative and qualitative—we hoped to fulfill the project's central mission.

And so we planned, developed, and deployed the following two experimental models for combining text-mining and visualizations:

(1) ASSESSING DIGITIZATION QUALITY: This interactive visualization plots a *quantitative* survey of our newspaper corpus. Users of this interface can plot the quantity of information by geography and time periods, using both to survey the amount of information available for any given time and place. This is available both at the macro-level (that is, Texas as a region) and the micro-level (by diving into the quantity and quality of individual newspaper titles), and can be tailored to any date range covered by the corpus. The central purpose of this model is to enable researchers to expose and parse the amount of information available in a database of digitized historical newspapers so they can make more informed choices about what research questions they can answer from a given set of data. (The creation of this interface, and how it works, is described in greater detail in the section below.)

(2) ASSESSING LANGUAGE PATTERNS: This interactive visualization offers a *qualitative* survey of our newspaper corpus. Users of this interface can plot and browse three major language patterns in the newspaper corpus by geography and time periods. This can be done at both the regional level (Texas) and for specific locations (individual cities), as well as for any given date range covered by the corpus. For this model, we made available three of the most widely used methods for assessing large-scale language patterns: overall word counts, named entity counts, and topic models of particular date ranges. (The details of each of those categories, as

well as the creation and operation of this model, is also described in greater detail below). The overarching purpose of this visualization is to provide users with the ability to survey the collected language patterns that emanate from the newspaper collection for any particular location or time period for the available data.

PROJECT TEAMS

Because the project required deep expertise in multiple fields, we built two project teams that each tackled a distinct side of the project. A team based at the University of North Texas focused on the language assessment, quantification, and overall text-mining side of the project. A team at Stanford University worked on designing and constructing the dynamic visualizations of those language patterns. The two teams worked in tandem—as parallel processes—to continually tailor, adjust, and refine the work on both sides of the project as we sought to fit these two sides together.

The University of North Texas team was headed by Andrew J. Torget, a digital historian specializing in the American Southwest, and Rada Mihalcea, a nationally-recognized computer science expert in natural language processing. Tze-I "Elisa" Yang (a graduate student in UNT's computer science department) took the lead in data manipulation and processing of the text-mining efforts, while Mark Phillips (Assistant Dean for Digital Libraries at UNT) provided technical assistance in accessing the digital newspapers.

The Stanford team was headed by Jon Christensen (Executive Director for the Bill Lane Center for the American West) and Geoff McGhee (Creative Director for Media and Communications at the Lane Center). Yinfeng Qin, Rio Akasaka and Jason Ningxuan Wang

(graduate students in Stanford's computer science department), and Cameron Blevins (graduate student in history) assisted in the development of the quantitative visualization model, as well as website design for the project. Maria Picone and her team at Wi-Design (http://wi-design.com/) worked with the project to develop the qualitative visualization model.

These collaborations built on a partnership forged between Andrew Torget and Jon Christensen during an international workshop, "Visualizing the Past: Tools and Techniques for Understanding Historical Processes," held in February 2009 at the University of Richmond. (For more information about this workshop, and the white paper it produced, see http://dsl.richmond.edu/workshop/.) That workshop, sponsored by an earlier grant from the National Endowment for the Humanities, provided the springboard for this project.

BUILDING A QUANTITATIVE MODEL: ASSESSING NEWSPAPER QUALITY

Following our initial assessments of the newspaper corpus, we determined to build our first model to examine the quality and quantity of information available in our data set.

THE NEED FOR DATA TRANSPARENCY

Part of the problem with current tools available for searching collections of historical newspapers—typically limited to simple word searches—is that they provide the user with little or no sense of how much information is available for any given time period and/or geographic location. If, for example, a scholar was interested in how Abraham Lincoln was represented in Georgia newspapers during the Civil War, it would be highly useful to be able to determine how much information a given database contained from Georgia newspapers during the 1861-1865 era. Without such information, it would be remarkably difficult for a researcher to evaluate whether a given collection of digitized historical newspapers would likely hold a great deal of potentially useful information or would likely be a waste of time. Indeed, without such tools for data transparency, it would be difficult for a researcher to know whether a search that produced a small number of search results would indicate few discussions of Lincoln from that era or simply that few relevant resources were available within the dataset.

OCR QUALITY

In a digital environment, assessing the quantity of information available also necessitates assessing the quality of the digitization process. The heart of that process for historical newspapers is when scanned images of individual pages are run through a process known as

optical character recognition (OCR). OCR is, at base, a process by which a computer program scans these images and attempts to identify alpha-numeric symbols (letters and numbers) so they can be translated into electronic text. So, for example, in doing an OCR scan of an image of the word "the," an effective OCR program should be able to recognize the individual "t" "h" and "e" letters, and then save those as "the" in text form. Various versions of this process have been around since the late 1920s, although the technology has improved drastically in recent years. Today most OCR systems achieve a high-level of recognition accuracy when used on printed texts and calibrated correctly for specific fonts.

Images of historical newspapers, however, present particular challenges for OCR technology for a variety of reasons. The most prolific challenge is simply the quality of the images of individual newspaper pages: most of the OCR done on historical newspapers relies upon microfilmed versions of those newspapers for images to be scanned, and the quality of those microfilm images can vary enormously. Microfilm imaging done decades ago, for example, often did not film in grayscale (that is, the images were taken in essentially black-and-white, which meant that areas with shadows during the imaging process often became fully blacked out in the final image) and so OCR performed on poorly imaged newspapers can sometimes achieve poor results because of the limitations of those images. Another related challenge is that older newspapers, particularly those from the nineteenth century, typically employed very small fonts in very narrow columns. The tiny size of individual letters, by itself, can make it difficult for the OCR software to properly interpret them, and microfilm imaging done without ideal resolution can further compound the problem. Additionally, the small width of many columns in historical newspapers also means that a significant percentage of words can also be lost during the OCR

process because the widespread use of hyphenation and word breaks (such as “pre-diction” for “prediction”) which newspaper editors have long used to fit their texts into narrow columns.

OCR on clean images of historical newspapers can achieve high levels of accuracy, but poorly imaged pages can produce low levels of OCR recognition and accuracy. These limitations, therefore, often introduce mistakes into scanned texts (such as replacing “l” with “1” as in “1imitations” for “limitations”). That can matter enormously for a researcher attempting to determine how often a certain term was used in a particular location or time period. If poor imaging—and therefore OCR results—meant that “Lincoln” was often rendered as “Linco1n” in a data set, that *should* affect how a scholar researching newspaper patterns surrounding Abraham Lincoln would go about his or her work.

As a result, we needed to develop methods for allowing researchers to parse not just the quantity of the OCR data, but also some measure of its quality as well. We therefore set about experimenting with developing a transparent model for exposing the quantity and quality of information in our newspapers database.

SCRUBBING THE OCR

Because the newspaper corpus was so large, we had to develop programmatic methods of formatting and assessing the data. Our first task was to scrub the corpus and try to correct simple recurring errors introduced by the OCR process:

- Common misspellings introduced by OCR could be detected and corrected, for example, by systematically comparing the words in our corpus to English-language dictionaries. For this task, we used the GNU Aspell dictionary (which is freely available and fully compatible with

UTF-8 documents), and then ran a series of processes over the corpus that checked every word in our newspaper corpus against the dictionary. Within Aspell we also used an additional dictionary of place names gathered from Gazetters. This way, Aspell could also recognize place names such as "Denton" or "Cuahtemoc," and also suggest them as alternatives when there was a slight misspelling. Whenever a word was detected that did not match an entry in the dictionary, we checked if a simple replacement for letters that are commonly mis-rendered by OCR (such as "Linco1n" for "Lincoln") would then match the dictionary. We made these replacements only with the most commonly identified errors (such as "1" for "l" and "@" for "a"), and we experimented with this numerous times in order to refine our scripts based on hand-checking the results, before running the final process over the corpus.

- End-of-line hyphenations and dashes could also be programmatically identified and corrected in the OCR'd text. If a word in the original newspaper image had been hyphenated to compensate for a line-break (such as "historical" being broken into "hist-orical"), that would create in the OCR text two nonsensical words "hist-" and "orical" which would not match any text searches for "historical" even though the word did appear in the original text. To correct for this, we ran a script over the corpus that looked for words that end with a hyphen, and was followed by a word that did not match any entries in our dictionary. The two parts ("hist-" and "orical") were then reconnected with the hyphen removed ("historical"), and if that reconnected word now matched an entry in the dictionary, we made the correction.
- We also experimented with the use of language models as a way to correct potential misspellings in the data. Specifically, we considered the construction of unigram and bigram

probabilistic models starting with our existing newspaper dataset. These models can be used to suggest corrections for words occurring very rarely, which are likely to be misspellings. For efficiency reasons, we ended up not applying these models on the dataset we worked with because the methods did not scale up well, but the initial results were promising, which suggests this as a direction for future investigations.

- To give an idea of the coverage and efficiency of this spelling correction phase, we collected statistics on a random sample of 100 documents. From a total of 209,686 words, Aspell identified as correct 145,718 (70%), suggested acceptable replacements for 12,946 (6%), and could not find a correction for 51,022 (24%). (The processing of this set of documents took 9 minutes and 30 seconds.)

The objective of this work was simply to automate some basic clean-up of known OCR errors so that we could get a finer and more accurate sense of how much true "noise" (that is, unrecognizable information) was present in the corpus compared to recognizable words and content.

It is worth noting that we do not believe—and are not claiming—that this scrubbing process was without flaws. There were almost certainly words in the corpus that had variant spellings that did not match the dictionary, and were therefore counted as "noise" when they were not. It is also likely that, on occasion, when the scripts made a correction of "l" for "1" that the resulting word was not what had appeared in the original.

We attempted to guard against these problems by rigorously spot-checking (that is, having human readers verify the scrubbing results) the corrections as we developed our scripts in order to ensure that this scrubbing process was correcting errors rather than introducing them. Those

spot-checks reassured us that, yes, the scripts were overwhelmingly correcting common errors, and whatever errors they introduced were likely quite few in number (especially when compared to the enormous size of the overall corpus). And because of the magnitude of our corpus, there was simply no other way to handle such common errors (since proof-reading by hand would be impossible) unless we simply ignored them. We chose not to ignore them because that seemed to artificially increase the level of noise in the corpus, and we wanted to represent as refined—and thus as accurate—a sense of the quality of the corpus as possible.

FORMATTING THE DATA

Once we had our corpus scrubbed from easily corrected errors introduced by the OCR process, we then ran the full newspaper data set against the dictionary once more to produce a word count of recognized words ("good" content, in the sense that the OCR rendered usable text) to unrecognized words ("bad" content, noise of various sorts introduced by the OCR process that had rendered unusable text). This provided a database of metrics of the quality of the data, which we then organized by newspaper title and year. So for every newspaper title, we had counts of the "good" and "bad" words per year, giving us a finely grained database of the quantity and quality of our newspaper data as it spread out across both time and space.

BUILDING THE VISUALIZATION

As we worked on developing these language metrics at UNT, the Stanford team began developing a dynamic interface that would enable people to visualize and explore those data

points. From the outset, we knew there would be two ways that we would want to index the collection: by time period and by geography.

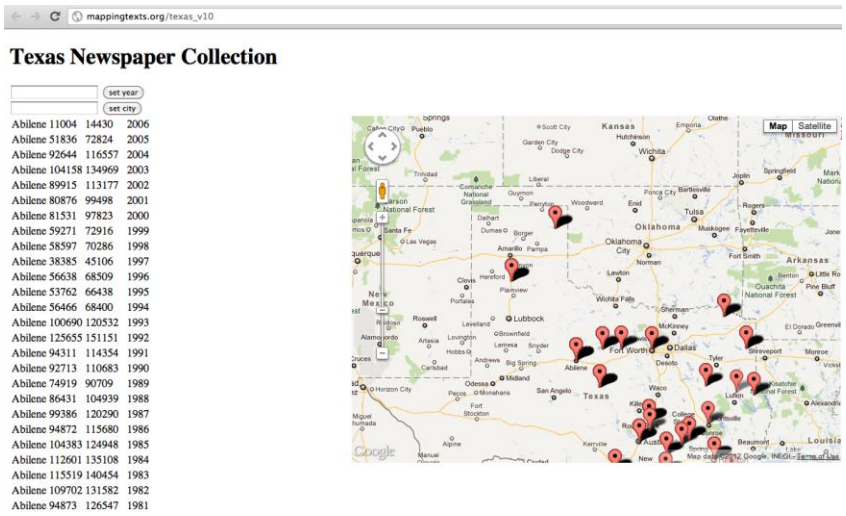
In order to build as reusable and flexible a visualization as possible, the team opted to use "off the shelf" interface widgets to construct the interactive display of scan quality and collection size in order to minimize the amount of development time for creating interface elements, and produce an application that would be as easy as possible to re-deploy with other datasets in the future.

Freely available or open source widgets used for the visualization included the following:

- Google Maps for plotting spatial data on a scrollable, zoomable basemap.
- The Google Finance time series widget for dynamically querying different time ranges.
- A scrollable timeline of Texas history, built using MIT's "Simile" collection of open-source (http://www.simile-widgets.org/timeline/) interface widgets.
- The Protovis (http://vis.stanford.edu/protovis) charting library developed at Stanford was used for plotting ratios of recognized to unrecognized words over time for individual newspapers.

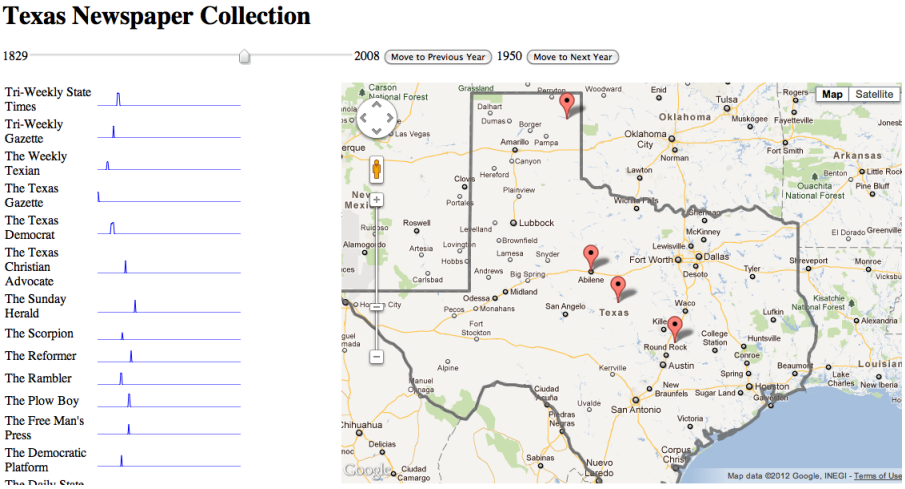
Looking at the visualization interface as a work in progress, one can clearly see the steps and decisions that go into refining a visual tool for exploring data. The team began by simply plotting the data on a basemap without any symbology, which immediately revealed the heavy representation in the collection of newspapers from the eastern portions of Texas. This naturally tracks with the concentration of Texas cities in both the contemporary and historical periods, but might understandably give pause to a historian interested in West Texas. The first iteration also provided a primitive results display interaction, as moving the mouse over a city would provide a

tabular display of recognized words out of total words per year. Also, the interface included form fields that would allow a user to set a single year as a time query:

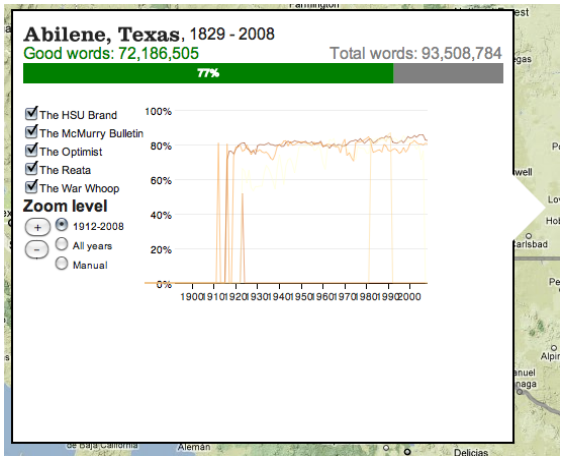


Over time, refinements included the following:

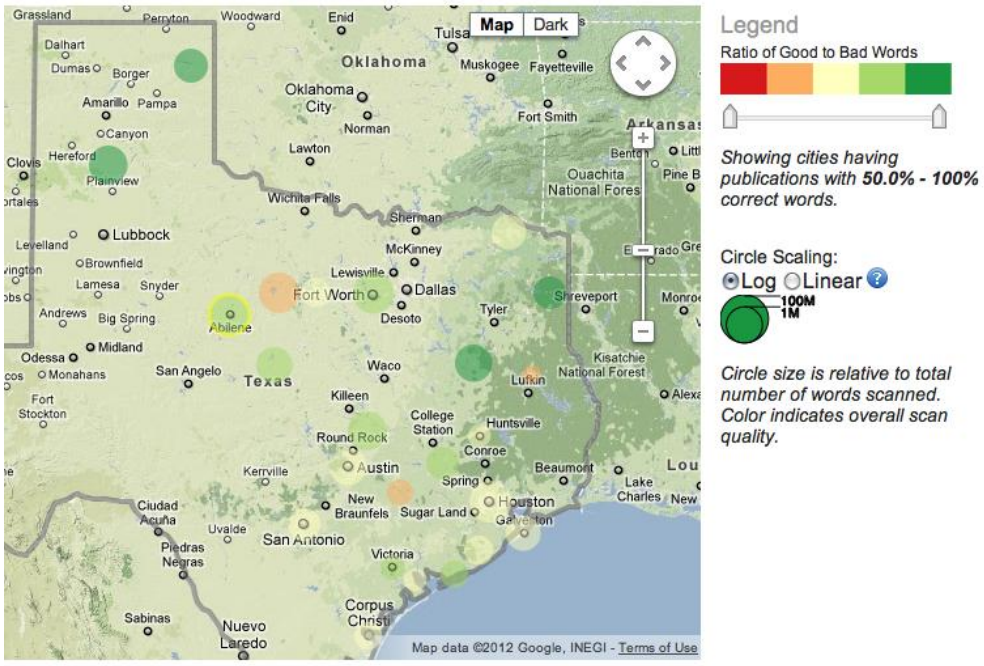
- Highlighting the Texas state borders using polygon data that could be easily swapped out for another state or region if desired for future re-use.
- Changing the time selection tool from an editable text field to a draggable slider, and later to a two handled slider that let the user select both a start and end date.



- Creating a “detail view” for cities selected on the map, showing their total ratio of good to bad words over time, and allowing a user to drill down into each individual publication in the each location, in the selected time period.



- Adding symbology to the map to enable at-a-glance information on (1) the size of a collection for a given city, and (2) the overall ratio of good to bad words in the collection. It was determined that using a circle sized to the relative quantity of pages and colored according to the ratio of good to bad could quickly impart basic information. And this symbology would update to reflect values changing according to the temporal range selected.

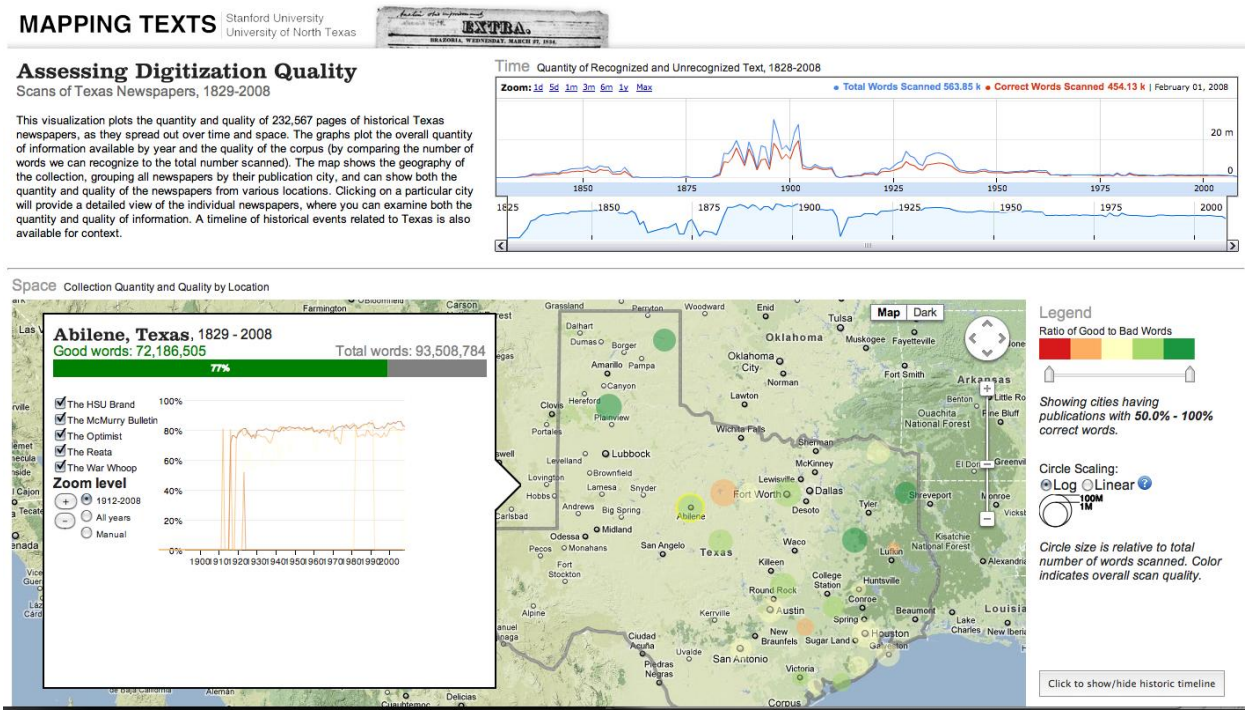


- Adding a timeline of Texas history to help less specialized users place the different time periods in context. To save space, it was decided that this extra timeline could be shown or hidden on demand.

The final version of this visualization (http://mappingtexts.org/quality) offers multiple ways to access and parse the quantity and quality of the digitized newspapers. Moreover, it contains an annotation layer of descriptive headlines, an introductory text, scales and labels.

Lastly, all of the onscreen texts are drawn from a simple text "configuration" document that could be easily edited to change the labeling, geographic or temporal context, or underlying data sets.

Here is the completed version:



- At the top is a timeline that plots the quantity of words (both in the complete corpus, and the "good words") over time, providing an overall sense of how the quantity of information ebbs and flows with different time periods. Users can also adjust the dates on the timeline or order to focus on a particular date-range in order to explore in more detail the quantity of information available.
    - And in our collection, visualizing the data on this timeline reveals that two time periods in particular dominate the information available in our collection: 1883-1911 and 1925-1940. Even though the entire collection represents 1829-2008, newspapers from those

two smaller eras vastly outnumber the representation from any other era. This finding makes sense, too, since these are eras that were targeted by the initial phases of the *Chronicling America* project, and therefore are most likely to be overrepresented in that dataset. For the moment, it seems that scholars of the Gilded and Progressive eras would be far better served by this database than scholars of other periods.

- Adjusting the timeline also affects the other major index of the content: an interactive map of Texas. For the visualization, all the newspapers in the database were connected by their publication city, so they could be mapped effectively. And so the map shows the geographic distribution of the newspaper content by city. This can be adjusted to show varying levels of quality in the newspaper corpus (by adjusting the ratio bar for "good" to "bad" words) in order to find areas that had higher or lower concentrations of quality text. The size of the circles for cities show their proportion of content relative to one another—which the user can switch from a logarithmic view (the default view, which gives a wider sense of the variety of locations) to a linear view (which provides a greater sense of the disparity and proportion of scale between locations).
    - Viewing the database geographically reveals that two locations dominate the collection: newspapers from Houston and Ft. Worth. Combined, those two locations outstrip the quantity of information available from any other location in Texas, which is interesting in part because neither of those locations became dominant populations centers in Texas until the post World War II era (and therefore well after the 1883-1911 and 1925-1940 time periods that compose the majority of the newspaper content). This would suggest that the newspapers of rural communities, where the

majority of Texans lived during the Gilded and Progressive eras, are underrepresented among the newspapers of this collection, and that urban newspapers—and therefore urban concerns—are likely overrepresented. While scholars of urbanization would be well-served, scholars interested in rural developments, it seems, would be advised to be wary of this imbalance when conducting research with this collection.

- The third major window into the collection is a detail box that, for any given location (such as Abilene, Texas), provides a bar of the good-to-bad word ratio, a complete listing of all the newspapers that correspond to that particular location, and metrics on the individual newspapers. The detail box also provides access to the original newspapers themselves, as clicking on any given newspaper title will take the user to the originals on UNT's *Portal to Texas History* site (http://texashistory.unt.edu/).
    - Exploring the various geographic locations with the detail box reveals more useful patterns about the information available in the dataset. Although Houston and Ft. Worth represent the locations with the *largest* quantity of available data, they are not the locations with the *highest* quality of available data. The overall recognition rate for the OCR of Houston newspapers was only 66 percent (although this varied widely between various newspapers), and for Ft. Worth the overall rate was 72 percent. By contrast, the newspaper in Palestine, Texas, achieved an 86 percent quality rate, while the two newspapers in Canadian, Texas, achieved an 85 percent quality rate. At the lowest end of quality was the OCR for newspapers from Breckenridge, Texas, which achieved only a 52 percent recognition rate. Scholars interested in researching places like Breckenridge or Houston, then, would need to consider that anywhere between a

third to fully half of the words OCR'd from those newspapers were rendered unrecognizable by the digitization process. Scholars who decided to focus on newspapers from Palestine or Canadian, on the other hand, could rely on the high quality of the digitization process for their available content.

- One consistent metric that emerges from plotting the data in this visualization is that the quality of the OCR improved significantly with newspapers published after 1940. That makes sense because the typeface for post-World War II newspapers was often larger than that used in earlier newspapers (especially compared to nineteenth-century newspapers), and because the microfilm imaging done for later newspapers was often of higher quality. While newspaper from earlier eras were digitized in larger numbers, the quality of the digitization process was higher for post-1940 newspapers.

BUILDING A QUALITATIVE MODEL: ASSESSING LANGUAGE PATTERNS

Once we had completed our quantitative survey of the collection, we turned our attention to building a model for a qualitative assessment of the language patterns of our digitized newspaper collection. With this model, we wanted to experiment with ways for people to explore the dominant language patterns of the newspapers as they spread out across both time and space.

COMMON LANGUAGE METRICS

We chose to focus on three of the metrics most widely used by humanities scholars for surveying language patterns in large bodies of text, and use that for a visualization of the language patterns embedded in the collection:

(1) Word Counts. One of the most basic, and widely used, metrics for assessing language use in text has been word counts. The process is simple—run a script to count all the words in a body of text, and then rank them by frequency—and the hope is to expose revealing patterns by discovering which words and phrases appear most frequently. Such counts have become perhaps the most recognizable text-mining method, as word clouds (which typically show the most frequently appearing words in a text) have become popular online.

(2) Named Entity Recognition (NER) Counts. This is a more finely-grained version of basic word counts. In collecting NER counts, a program will attempt to identify and classify various elements in a text (usually nouns, such as people or locations) in a body of text. Once that has been completed, the frequency of those terms can then be tallied and ranked, just like with basic

word counts. The result is a more specific and focused ranking of frequency of language use in the corpus of text.

(3) Topic Modeling. This method of text-analysis has grown in popularity among humanities scholars in recent years, with the greater adaption of programs like the University of Massachusetts's MALLET (MAchine Learning for LanguagE Toolkit). The basic concept behind topic modeling is to use statistical methods to uncover connections between collections of words (which are called "topics") that appear in a given text. Topic modeling uses statistics to produce lists of words that appear to be highly correlated to one another. So, for example, running the statistical models of MALLET over a body of text will produce a series of "topics," which are strings of words (such as "Texas, street, address, good, wanted, Houston, office") that may not necessarily appear next to one another within the text but nonetheless have a statistical relationship to one another. The idea behind topic modeling is to expose larger, wider patterns in language use than a close-reading would be able to provide, and the use of these models has gained increasingly popularity among humanities scholars in recent years, in large measure because the statistical models appear to produce topics that seem both relevant and meaningful to human readers.

COLLECTING WORD AND NER COUNTS

Generating the dataset for the word counts was a simple process of counting word occurrences, ranking them, and then organizing them by newspaper and location.

Generating the Named Entity Recognition dataset was somewhat more complicated. There are a number of available programs for performing NER counts on bodies of text, and we

spent a fair amount of time experimenting with a variety of them to see which achieved the best results for our particular collection of historical newspapers. To determine the accuracy of the candidate parsers, we manually annotated a random sample of one hundred named entities from the output of each parser considered. To measure the efficiency (because scale, again, was a primary consideration), we also measured the time taken for the parser to label 100 documents.

Among those we tried that did not—for a variety of reasons—achieve high levels of accuracy for our collection were LingPipe, MorphAdorner, Open Calais, and Open NLP. We had a great deal more success with the Illinois Named Entity Tagger (http://cogcomp.cs.illinois.edu/page/publication_view/199). It was, however, the Stanford Named Entity Recognizer (http://www-nlp.stanford.edu/software/CRF-NER.shtml) that achieved the best parser accuracy while also maintaining a processing speed comparable with the other taggers considered. We, therefore, used the Stanford NER to parse our newspaper collections. We then ranked the NER counts by frequency and organized them by newspaper and year.

TOPIC MODELING

For our topic modeling work, we decided to use the University of Massachusetts's MALLET package (http://mallet.cs.umass.edu/) for a number of reasons, the most prominent of which were that (a) it is well documented, and (b) other humanities scholars who have used the package have reported high quality results (see, for example, the work of Cameron Blevins at http://historying.org/2010/04/01/topic-modeling-martha-ballards-diary/ and Robert K. Nelson at http://dsl.richmond.edu/dispatch/pages/intro). MALLET also uses the probabilistic latent semantic analysis (pLSA) model that has become one of the most popular within the natural

language processing field of computer science, and so we decided to use the package for our experiments in testing the effectiveness of topic modeling on our large collection of historical newspapers.

We spent far more time working on and refining the topic modeling data collection than any other aspect of the data collection for this project. Much of that work concentrated on attempting to assess the quality and relevance of the topics produced by MALLET, as we ran repeated tests on the topics produced by MALLET that were then evaluated by hand to see if they appeared to identify relevant and meaningful language patterns within our newspaper collection. The results of those experiments resulted in a paper, "Topic Modeling on Historical Newspapers," that appeared in the proceedings of the Association for Computational Linguistics workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (ACL LATECH 2011). The paper is included as an appendix to this white paper. In short, our close examination of the topics produced by MALLET convinced us that the statistical program did, indeed, appear to identify meaningful language patterns in our newspaper collection. We therefore determined to process our entire newspaper corpus using MALLET.

We generated topics for every newspaper and location in the collection. There was, however, a challenge that emerged when it came to setting the time ranges for our topic models. With basic word counts and NER counts—which were tallied by year—we could easily recombine any given range of years and get accurate new results. So, for example, if we wanted to display the most frequently appearing words from 1840 to 1888, we could simply add up the word counts for all those years. Topic models, by contrast, are unique to every set of text that you run through

MALLET, which meant that we could not generate topics by individual years and then hope to combine them later to represent various date ranges.

We, therefore, decided to select historically relevant time periods for the topic models, which seemed the closest that we could get to building a data set of topic models that could be comparable and useable in context with the word and NER counts. The eras that we selected were commonly recognized eras among historians who study Texas and the U. S.-Mexico borderlands: 1829-1835 (Mexican Era), 1836-1845 (Republic of Texas), 1846-1860 (Antebellum Era), 1861-1865 (Civil War), 1866-1877 (Reconstruction), 1878-1899 (Gilded Age), 1900-1929 (Progressive Era), 1930-1941 (Depression), 1942-1945 (World War II), 1946-2008 (Modern Texas). For each of these eras, we used MALLET to generate a list of topics by location. We believe this kind of iterative conversation between history and other humanities disciplines, on the one hand, and information science and computer science, on the other, is an essential part of the process of designing, building, and using models such as the ones we constructed.

BUILDING THE VISUALIZATION

The interface for the textual analysis visualization presented some challenges not posed by the OCR quality visualization. The temporal and spatial dimensions were roughly the same, but this visualization needed to show the results from three separate types of natural language processing, not all of which could be sliced into the same temporal chunks.

The team decided that the best approach would be to repeat the time slider and map interface, and add a three-part display to present the individual results of the three NLP

techniques for the active spatial and temporal query. In essence, this meant three separate list views, each updating to represent changes in the spatial and/or temporal context:

- Word counts for any given time, place, and set of publications.
- Named entity counts for any given time, place, and set of publications.
- Topic models for any given era, and individual locations.

One other challenge presented by moving from an analysis of the data quality to drilling down into the collections themselves was the sheer scale of information. Even compressed into "zip" archives, the natural language processing results comprised around a gigabyte of data. A sufficiently "greedy" query of all newspapers, in all cities, in all years would – at least in theory – demand that this (uncompressed) gigabyte-plus of data to be sent from the server to the visualization. Fortunately, showing 100 percent of this information would require more *visual* bandwidth than three columns of word lists could accommodate. We therefore decided that only 50 most frequently occurring terms would be shown for the word counts and NER counts. Topic models, for their part, would display the 10 most relevant word clusters. This decision allows the interface to maintain a high level of response to the user's queries and questions, while also highlighting the most prominent language patterns in the text.

The result is an interactive visualization (http://mappingtexts.org/language) that maps the language patterns of our newspaper collection over any particular time period and geography selected by the user:

MAPPING TEXTS Stanford University, University of North Texas

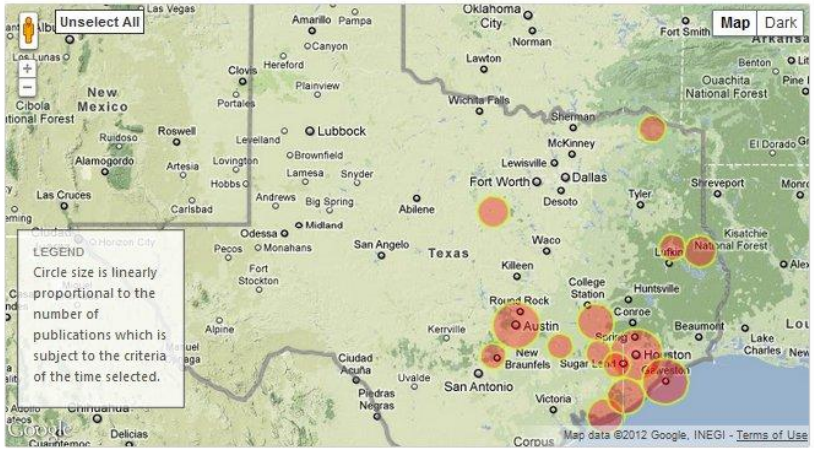
Assessing Language Patterns: A Look At Texas Newspapers, 1829-2008

This visualization plots the language patterns embedded in 232,567 pages of historical Texas newspapers, as they evolved over time and space. For any date range and location, you can browse the most common words (word counts), named entities (people, places, etc), and highly correlated words (topic models). [About Mapping Texts]

Time Period

Mexican Era	Republic of Texas	Antebellum Era	Civil War	Reconstruction
Gilded Age	Progressive Era	Depression	World War II	Modern Texas

1829 1847 1865 1883 1901 1919 1937 1955 1973 2008



Date Range 1836 - 1845

Publication By City:

- **La Grange** Deselect All
  - La Grange Intelligencer [view]
- **Houston** Deselect All
  - National Intelligencer [view]
  - The Houstonian [view]
  - The Weekly Citizen [view]
  - The Musquito [view]
  - The Morning Star [view]
  - Telegraph And Texas Register [view]
- **Nacogdoches** Deselect All
  - Texas Chronicle [view]
- **Columbia** Deselect All

Total Word Counts

list | tag cloud

copy info

texas	36,287
county	27,515
country	21,442
time	21,034
houston	20,809
land	20,562
court	20,399
sale	20,309
law	19,160
house	18,174
city	17,697
day	17,543
ihe	17,533
government	16,982
republic	16,731
tho	16,359
public	15,947
office	15,672
notice	15,179
received	14,692
john	13,831
united	13,761
persons	13,024
act	12,284
congress	11,644

Named Entity Counts

copy to clipboard

list | tag cloud

copy info

texas	26,142
houston	15,540
galveston	9,203
united states	8,576
mexico	8,439
congress	7,438
austin	6,340
republic	5,659
washington	4,941
mexican	4,805
brazoria	4,694
american	4,432
new york	4,012
matagorda	3,993
english	3,733
house	3,541
french	3,490
senate	3,482
messrs	3,310
mr	3,291
new orleans	3,098
england	3,005
indians	2,596
thai	2,522
spanish	2,430

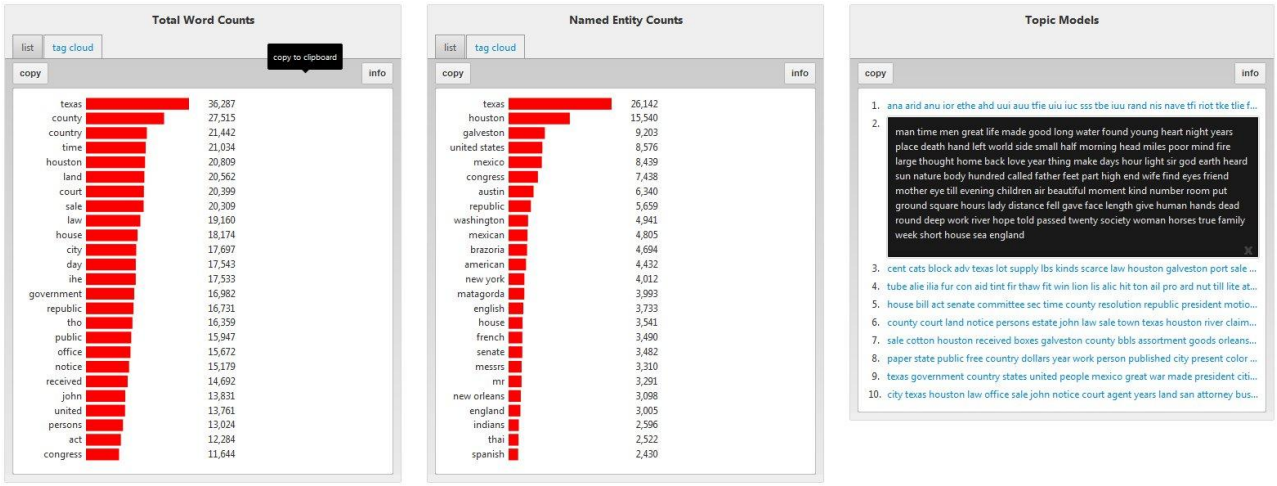
Topic Models

copy info

1. ana arid anu ior ethe ahd uui auu tfie uiu iuc sss tb...
2. man time men great life made good long water found young heart night years place death hand left world side small half morning head miles poor mind fire large thought home back love year thing make days hour light sir god earth heard sun nature body hundred called father feet part high end wife find eyes friend mother eye till evening children air beautiful moment kind number room put ground square hours lady distance fell gave face length give human hands dead round deep work river hope told passed twenty society woman horses true family week short house sea england
3. cent cats block adv texas lot supply lbs kinds scarc...
4. tube alie ilia fur con aid tint fir thaw fit win lion lis ...
5. house bill act senate committee sec time county r...
6. county court land notice persons estate john law s...
7. sale cotton houston received boxes galveston cou...
8. paper state public free country dollars year work p...
9. texas government country states united people m...
10. city texas houston law office sale john notice cour...

Just as with our quantitative model, the user can select any time period from 1829 through 2008. For several reasons, we have also included pre-set buttons for historically significant eras in Texas and U. S.-Mexican borderlands history (Mexican Era, 1829-1836; Republic of Texas, 1836-1845, and so on) which, if clicked, will automatically reset the beginning and end points on the time slider to those particular eras. Once the user has selected a time frame, they can also customize the geography they want to examine. Based on the timeline selection, the map populates so that the user sees all the cities that have publications from the time period they selected. The user, then, can choose to examine all the newspapers relevant to their time period, or they could customize their selection to particular cities or even particular newspaper titles. If, for example, someone wanted to know about the language patterns emanating from Houston during a particular era, they could focus on that. If the user wanted to burrow as far down as a single publication, they can do that as well.

Once a user has selected a time frame and geography, they can then examine the three major language patterns which are listed below the map in their own "widgets":



In the word counts and named entity counts widgets, there are two ways to look at the language data: (1) as a ranked list—with the most frequently appearing words at the top

followed by a descending list—that reveals the most frequently used terms in the collection, and (2) a word cloud that provides another way to look at the constellation of words being used, and their relative relationship to one another in terms of frequency. The word cloud has become one of the most common and popular methods of displaying word counts, and we see a great deal of value in its ability to contextualize these language patterns. But we have also found that our ranked list of these same words to be highly effective, and perhaps a great deal more transparent in how these words relate to one another in terms of quantification.

In the topic model widget, the user is offered the top ten most relevant "topics" associated with a particular date range. Within each topic is a list of 100 words that have a statistical relationship to one another in the collection, with the first word listed being the most relevant, the second being the second-most relevant, and so on. The 100 words are truncated for display purposes, but clicking on any given topic will expand the word list to encompass the full collection, which allows the user to parse and explore the full set of topic models.

Each topic's collection of words is meant to expose a theme of sorts that runs through the words in the newspapers selected by the user. Sometimes the topic is a collection of nonsensical words (like "anu, ior, ethe, ahd, uui, auu, tfie" and so on), when the algorithm found a common thread among the "noise" (that is, words that were jumbled by the digitization process) and recognized a commonality between these non-words, which it then grouped into a "topic." More often, however, the topic models group words that have a clear relationship to one another. If, for example, the user were to select all the newspapers from the Republic of Texas era, one of the topic models offered includes "Texas, government, country, states, united, people, mexico, great, war . . . " which seems to suggest that a highly relevant theme in the newspapers during this era

were the international disputes between the United States and Mexico over the future of the Texas region (and the threat of war that came with that). That comports well, in fact, with what historians know about the era. What is even more revealing, however, is that most of the other topic models suggest that this was only one—and perhaps even a lesser—concern than other issues within the newspapers of 1830s and 1840s Texas, such as matters of the local economy ("sale, cotton, Houston, received, boxes, Galveston"), local government ("county, court, land, notice, persons, estate"), and social concerns ("man, time, men, great, life"), which have not received nearly as much attention from historians as the political disputes between the United States and Mexico during this period.

The volume of information available here for processing is absolutely enormous, and so we are continuing our work in sifting through all of this language data by using this visualization interface. What we have seen, however, are numerous examples (such as the one detailed above) that expose surprising windows into what the newspapers can tell us about the eras they represent, which we hope will open new avenues and subjects for historians and other humanities scholars to explore.

PRODUCTS

The following are the main products produced thus far by this project, all of which are detailed in the preceding white paper:

- *MappingTexts* project website (http://mappingtexts.org), which documents the project's work and provides access to all its major products.

- "Assessing Newspaper Quality: Scans of Texas Newspapers, 1829-2008" (http://mappingtexts.org/quality)

- "Assessing Language Patterns: A Look at Texas Newspapers, 1829-2008" (http://mappingtexts.org/language)

- Tze-I Yang, Andrew J. Torget, Rada Mihalcea, "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 2011), June 2011, pp. 96-104.

- The source code for our work will soon be posted in a GitHub repository for downloading and modifying by groups interested in using the interface.
  (see http://mappingtexts.org/data).

CONCLUSIONS AND RECOMMENDATIONS

The following are our main conclusions and recommendations for text-mining work with digitized historical newspapers:

- The need for data transparency: one of the most pressing challenges facing humanities scholars in the digital age is the tremendous need for greater transparency about the quantity and quality of OCR data in databases of historical newspapers.
    - OCR recognition rates are, we have determined, one of the most important metrics to identify about a particular collections of digitized historical newspapers in assessing the collection's utility for humanities research (that is, what research questions can and cannot be answered by the dataset).
    - This underscored, in turn, the need for a standardized vocabulary for measuring and evaluating OCR quality (which we attempted to do with our visual scales for ratios of "good" compared to "bad" words in the corpus).
- Programmatic "scrubbing" of a collection of historical newspapers (as we documented above) can improve the quality of the available set of texts. While this has to be done with great care, it can yield an improved and cleaner corpus. We recommend the GNU Aspell dictionary for this sort of work.
- The use of language models—such as unigram and bigram probabilistic models—for correcting spelling errors introduced by the OCR process show great promise. We were unable to implement the use of these on our full dataset because of the problems we ran into with scale, although we nonetheless recommend that future researchers explore this method as a promising avenue for such work.

- Topic modeling shows significant promise for helping humanities scholars identify useful language patterns in large collections of text. Based on our extensive experimentation, we recommend the University of Massachusetts's MALLET program (http://mallet.cs.umass.edu/).
- For Named Entity Recognition work, we recommend the Stanford Named Entity Recognizer (http://www-nlp.stanford.edu/software/CRF-NER.shtml) based on the high level of accuracy it achieved, as well as its ability to cope successfully with scale.

The following are our main conclusions and recommendations for visualization work with digitized historical newspapers:

- In designing the visualizations, our team hopes that our efforts to modularize and simplify the design and functionality offer the possibility of further return on investment in the future, be it for similar text-quality visualizations, or for other spatio-temporal datasets. The source code is posted in a GitHub repository for downloading and modifying by groups interested in using the interface.
- Although the use of open source and commonly available widgets saves time and effort, it has some drawbacks, including lack of customization options in terms of design or deeper functionality, and dependence on the stability of the underlying APIs (Application Programming Interfaces). Already, Google Maps has gone through some dramatic revisions "under the hood," as well as introducing metering for high-volume users. The Google Finance widget, on the other hand, having been built in Flash, is not usable on mobile devices like phones or tablets running Apple's iOS. Still, we were able to produce a

moderately sophisticated information visualization spanning a large quantity of underlying data by relying almost entirely on freely available toolkits and widgets.

- The potency of data visualization as an analytical and explanatory tool was apparent very early on, from the moment that the team first passed around histograms showing the "shape" of the collection, from its overall peaks in quantity in the late 19th century and early to mid 20th, and how the rate of OCR quality climbed steadily, then dramatically, from the 1940s on. Moreover, the collection's spatial orientation was immediately apparent when we plotted it on the map. Interestingly, the size of the collections in a given city did not always track with that city's overall population, especially given the large collections of college newspapers.

APPENDIX 1: LIST OF DIGITIZED HISTORICAL NEWSPAPERS USED BY THE PROJECT

The following digitized newspapers—organized by their publication city—made up the collection used in this project, all of which are available on the University of North Texas's *Portal to Texas History* (http://texashistory.unt.edu/) and were digitized as part of the National Digital Newspaper Project's *Chronicling America* project (http://chroniclingamerica.loc.gov/):

Abilene

The Hsu Brand
The McMurry Bulletin
The Optimist
The Reata
The War Whoop

Austin

Daily Bulletin
Daily Texian
Intelligencer-Echo
James Martin's Comic Advertiser
Point-Blank
South and West
South-Western American
Temperance Banner
Texas Almanac -- Extra
Texas Real Estate Guide
Texas Sentinel
Texas State Gazette
The Austin City Gazette
The Austin Daily Dispatch
The Austin Evening News
The Daily State Gazette and General Advertiser
The Democratic Platform
The Free Man's Press
The Plow Boy
The Rambler
The Reformer
The Scorpion
The Sunday Herald
The Texas Christian Advocate
The Texas Democrat

The Texas Gazette
The Weekly Texian
Tri-Weekly Gazette
Tri-Weekly State Times

Bartlett

The Bartlett Tribune
The Bartlett Tribune and News
Tribune-Progress

Brazoria

Brazos Courier
Texas Gazette and Brazoria Commercial Advertiser
Texas Planter
The Advocate of The People's Rights
The People
The Texas Republican

Breckenridge

Breckenridge American
Breckenridge Weekly Democrat
Stephens County Sun
The Dynamo

Brownsville

El Centinela
The American Flag
The Daily Herald

Brownwood

The Collegian
The Prism
The Yellow Jacket

Canadian

The Canadian Advertiser
The Hemphill County News

Clarksville

The Northern Standard

Columbia

Columbia Democrat
Democrat and Planter

Telegraph and Texas Register
The Planter

Corpus Christi
The Corpus Christi Star

Fort Worth
Fort Worth Daily Gazette
Fort Worth Gazette

Galveston
Galveston Weekly News
The Civilian and Galveston Gazette
The Galveston News
The Galvestonian
The Texas Times
The Weekly News

Houston
De Cordova's Herald and Immigrant's Guide
Democratic Telegraph and Texas Register
National Intelligencer
Telegraph and Texas Register
Texas Presbyterian
The Houston Daily Post
The Houstonian
The Jewish Herald
The Morning Star
The Musquito
The Weekly Citizen

Huntsville
The Texas Banner

Jefferson
Jefferson Jimplecute
The Jimplecute

La Grange
La Grange Intelligencer
La Grange New Era
Slovan
The Fayette County Record
The Texas Monument

The True Issue

Lavaca

Lavaca Journal
The Commercial

Matagorda

Colorado Gazette and Advertiser
Colorado Tribune
Matagorda Bulletin
The Colorado Herald

Nacogdoches

Texas Chronicle

Palestine

Palestine Daily Herald

Palo Pinto

The Palo Pinto Star
The Western Star

Port Lavaca

Lavaca Herald

Richmond

Richmond Telescope & Register

San Antonio

The Daily Ledger and Texan
The Western Texan

San Augustine

Journal and Advertiser
The Red-Lander
The Texas Union

San Felipe

Telegraph and Texas Register

San Luis

San Luis Advocate

Tulia

The Tulia Herald

Victoria

Texas Presbyterian

Washington

Texas National Register
Texian and Brazos Farmer
The National Vindicator
The Texas Ranger

APPENDIX 2: TOPIC MODELING ON HISTORICAL NEWSPAPERS

The following paper appeared as: Tze-I Yang, Andrew J. Torget, Rada Mihalcea, "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 2011), June 2011, pp. 96-104.

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

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

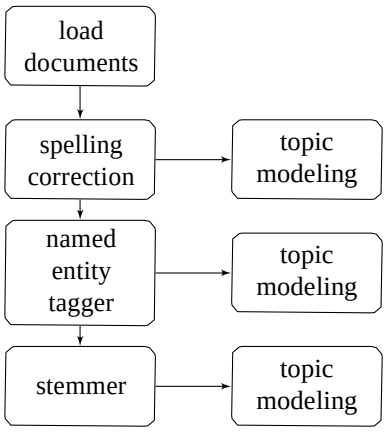


Figure 1: Work flow

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 company

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: $\text{Accuracy(topics)} = \frac{\text{\# of relevant topics}}{\text{total \# of topics}}$ $\text{Accuracy(terms)} = \frac{\text{\# of relevant terms in all topics}}{\text{total \# of terms in all topics}}$. 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 has 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.