Doing More with Digitization

% of article	Most likely words in topic in order of likelihood	Human-added topic label
47	court assembly general county office law election judge year council city person justice…	LAW & COURTS
18	act person aforesaid within authority further thereof enacted hereby officer state…	LEGISLATION
9	right great people power law colony act without britain subject country America liberty…	POLIT IDEOLOGY
8	state government constitution law united power citizen people public congress…	GOVERNMENT

An introduction to topic modeling of early American sources

In the 1990s, for a research project on colonial sexual coercion, I read hundreds of microfilmed early American newspapers for references to rape trials. Partway through my research, Accessible Archives released CD-ROMs of the *Pennsylvania Gazette* that were fully text searchable. I remember waiting eagerly for the processing of each new *Gazette* folio so that a simple keyword search could generate a list (printed on a dot-matrix printer) of all occurrences of the word *rape* in that folio's years.

A mere decade later, I am disappointed when an article or source is *not* available at the stroke of a few keys. Despite justified concerns over this exploding technology (Who will have access to these documents? Does the digitization of select documents re-privilege certain kinds of history?), historical-document digitization has enormously expanded research capabilities. What used to require months of searching can now be accomplished in an afternoon.

Yet having this material available electronically and fully searchable has created some new problems. Already we are seeing the limitations of keyword searching. In any given set of documents, some keywords are too broad in their meaning, some are too narrow, and others have too many different meanings. The

results of keyword searches are quite often incomplete or full of “noise,” irrelevant results that make it hard to find what you are looking for. For searching to be effective, access needs to be supplemented by analysis.

One promising way to move beyond keyword searching is to program computers not only to find words in these huge document collections but also to analyze documents by grouping them in subject classifications. This would provide a comprehensive indexing system that would far surpass human-indexing capabilities, but, even better, it would give scholars a complete picture of how their particular subject related to other subjects in that collection of documents: How much relative print space did the colonists give to discussions of Indians, of crime, or of politics? How did the entire contents of an eighteenth-century newspaper change over time?

A new kind of data-mining technique called topic modeling does just this. Topic modeling is based on the idea that individual documents are made up of one or more topics. It uses emerging technologies in computer science to automatically cluster topically similar documents by determining the groups of words that tend to co-occur in them. Most importantly, topic modeling creates topical categories without a priori subject definitions. This may be the hardest concept to understand about topic modeling: unlike traditional classification systems where texts are fit into preexisting schema (such as Library of Congress subject headings), topic modeling determines the comprehensive list of subjects through its analysis of the word occurrences throughout a corpus of texts. The content of the documents—*not* a human indexer—determines the topics collectively found in those documents. Sound hard to wrap your mind around? Before the math-challenged historians stop reading, let me try to put this in more accessible terms.

The Concepts behind Topic Modeling

Remember the *$10,000 Pyramid* hosted by Dick Clark? It was a game show started in the 1970s in which minor celebrities shouted a series of words or phrases and their contestant partners tried to guess the category to which those words belonged. So “*dog…parrot…cat…goldfish…pot-bellied pig*” would be possible hints for the category “Animals you keep as pets!”

The *Pyramid* show was based on the notion that our brains tend to look for patterns and make connections between similar items. Similarly, topic modeling determines a topic from the words that make up texts. In anthropomorphic terms, it “reads” hundreds of thousands of documents and “figures out” what words various documents most have in common. It then makes interpretable topics by ranking the words that are most likely to appear in those grouped sources. The words most likely to appear in documents that contain a given topic appear at the top of the list. By looking at the top ten to fifteen words, a scholar can then title the list with a short topic heading.

For instance, the topic model might group together the following words as those most likely to appear in a particular subset of documents: indian fort men town party off killed people came letter day french… In this case, we can easily identify that list as a topic related to interactions between colonists and Native Americans—perhaps we might label it **I**NDIANS, for simplicity's sake. By having the topic model divide an entire collection of documents into these list-of-word topics, topic modeling effectively indexes an entire corpus of texts.

Much of this is less than intuitive: How does topic modeling figure out what topics to make from the documents? How can a computer analyze which documents relate to which topics if it doesn't already know the topics? Rather than going into a detailed explanation of the formulae, statistics, and probabilities the computer scientists use in topic modeling, I'll turn to some colonial-era examples to show, if not *how* it works, at least *that* it works.

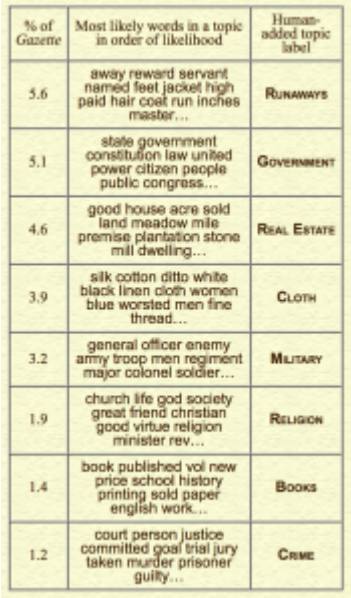
A Topic Model of the *Pennsylvania Gazette*

I recently entered into a joint project with computer scientists at the University of California, Irvine, to try newly developed topic-modeling techniques on historical documents. We've detailed the technical concepts and equations elsewhere, so here I will concentrate on a historian-friendly exploration of what one can do with a topic-modeled colonial newspaper. I'm using the topic-model algorithm made available by Padhraic Smyth and executed by David Newman, both of the Department of Computer Science, University of California, Irvine.

We began with the entire eighteenth-century run of the *Pennsylvania Gazette,* which includes approximately eighty-two thousand articles and advertisements from 1728-1800. Then we had to decide into how many topics the computer should divide the corpus of documents. We chose to do a forty-topic analysis of the *Gazette*. (While we specified the number of topics, the topic-model program determined the content of each topic without any human input.) The topic model then determined the forty topics and their prevalence in the *Gazette*.

Figure 1 shows a selection of some of the topics created from the *Gazette's* articles and advertisements. It lists between ten and thirteen of the most highly ranked words in each topic (depending on how many fit in the table), the prevalence of that topic in the *Gazette*, and a human-added title for each topic. Throughout this essay I've listed the most likely words that make up each topic in an alternative font, and the human-added topic labels in **S**MALL **C**APS to make identification of each easier. Readers may quibble over the exact topic label, but should be able to see the relationship between the words listed in each topic. Indeed, this transparency of topics is one of the

wonderful things about topic modeling. Users don't have to rely on what an indexer means by a particular subject label; they can look at the list of words in any given subject and decide for themselves how they understand the meaning of that collection of terms. And again, it is worth pointing out that these categories were not preselected before running the topic model; the computer used the *Gazette*'s contents to determine these topics.



% of Gazette	Most likely words in a topic in order of likelihood	Human-added topic label
5.6	away reward servant named feet jacket high paid hair coat run inches master…	RUNAWAYS
5.1	state government constitution law united power citizen people public congress…	GOVERNMENT
4.6	good house acre sold land meadow mile premise plantation stone mill dwelling…	REAL ESTATE
3.9	silk cotton ditto white black linen cloth women blue worsted men fine thread…	CLOTH
3.2	general officer enemy army troop men regiment major colonel soldier…	MILITARY
1.9	church life god society great friend christian good virtue religion minister rev…	RELIGION
1.4	book published vol new price school history printing sold paper english work…	BOOKS
1.2	court person justice committed goal trial jury taken murder prisoner guilty…	CRIME

Fig. 1. Sample of Gazette topic-model topics with human-suggested topic labels

Overall, our topic model shows that most *Gazette* articles and advertisements related to economics and politics. Despite the occasional poetry or fiction, the overall distribution of topics in a forty-topic run of the topic model confirms that the *Gazette* was fundamentally a newspaper about land, shipping, sales, and politics. Only about ten of the forty topics did not directly focus on economics or politics, and even these often indirectly related to those topics (e.g., topics on **I**NDIANS or **W**AR).

Beyond this picture of the overall content of the *Pennsylvania Gazette*, the topic model can be used to link topics to specific documents, to link specific words to the topics in which they are most likely to occur, and to track changes in topic prevalence over time. The rest of this article provides examples of such results and briefly suggests the research possibilities related to each finding.

Linking Topics and Documents

Because each set of topic words can be linked to the documents that most highly correlate to that topic, users can find individual documents on those topical subjects. Those documents that most exclusively focus on a topic are that

topic's most highly ranked. For instance, in the topic I've labeled **M**ILITARY above, the top-ranked *Pennsylvania Gazette* article is on the "OPERATIONS of the Allied Armies of France and America" (October 31, 1781), and the second-ranked article focuses on the European war theater (July 16, 1794)—both clearly documents that are primarily about a **M**ILITARY subject. In the category of **C**LOTH above, the most highly ranked documents include lists of a "great variety of plain and changeable mantuas" recently imported from Europe (April 14, 1773) and lists of broadcloths, flannels, swanskins, velvet, silk, and camblets (September 27, 1758). Beyond using this as a subject-based finding aid, researchers can get a better sense of a topic's definition from a ranked list of the documents most likely to contain a given topic.

The topic model allows users to see the multiple topics that a document simultaneously contains. Figure 2 shows how the topic model analyzed the reprint of the 1790 Constitution of the State of Pennsylvania in the September 8, 1790, *Gazette.* The topic model determined that the Constitution was focused primarily on a topic we might call **L**AW & **C**OURTS, secondarily on **L**EGISLATION, and more minimally on topics related to **P**olitical **I**deology and **G**overnment.

% of article	Most likely words in topic in order of likelihood	Human-added topic label
47	court assembly general county office law election judge year council city person justice…	LAW & COURTS
18	act person aforesaid within authority further thereof enacted hereby officer state…	LEGISLATION
9	right great people power law colony act without britain subject country America liberty…	POLIT IDEOLOGY
8	state government constitution law united power citizen people public congress…	GOVERNMENT

Fig. 2. Proportion of main topics contained within the 1790 Constitution of the State of Pennsylvania
with human-added topic labels

The topics contained in the 1790 Constitution seem appropriate and thus can confirm the accuracy of the topic model; if the model determined that the Constitution was primarily about **C**LOTH, for example, we would wonder about the model's precision. Other topical categorizations are equally accurate but perhaps not so immediately apparent—and as such, can suggest research possibilities. For instance, a poem published in the *Gazette* on July 28, 1737, titled "Women's Prerogative," correlates most strongly (over fifty percent of the article) to a topic that seems to be associated with the ideologies of Revolution (country men people liberty friend let man world god ever life mind virtue…), reminding us that the rhetoric of Revolution grew from existing vocabularies, including ones related to gender ideologies. Thus, scholars may be able to use topic modeling to trace how specific language was put to different uses across time and subject matter.

Linking Topics and Words

Topic modeling can also show users the most likely topics associated with particular words—type a word into a search box, and you can get a list of the most likely topics in which that word appears. Analyzing the significance of the appearance of words in various topics takes more careful contextual analysis than I can do here. But the results suggest the uses of this new approach to digital documents. Here are just a few examples.

Throughout the eighteenth-century *Gazette*, the word *slavery* is most highly associated with two topics that are both related to Revolutionary ideals and government forms: 1) COUNTRY MEN PEOPLE LIBERTY FRIEND LET MAN WORLD GOD EVER LIFE MIND VIRTUE … and 2) RIGHT GREAT PEOPLE POWER LAW COLONY ACT WITHOUT BRITAIN SUBJECT COUNTRY AMERICA LIBERTY … These associations suggest that readers of the *Pennsylvania Gazette* were far more likely to see discussions of the concept of slavery in relation to the rhetoric of political enslavement than in relation to the actual enslavement of Africans.

We can also use this technology to ask how colonists talked about firearms in the *Gazette*. The word *gun* is most highly related (over eighty percent of the time) to discussions relating to sea-going vessels that often carried an array of weapons (CAPT SHIP TAKEN ARRIVED FRENCH PRIVATEER GUN MEN VESSEL FLEET WAR SAIL …). About ten percent of the time, guns were discussed in relation to conflicts with Native Americans (INDIAN FORT MEN TOWN PARTY OFF KILLED PEOPLE CAME LETTER DAY FRENCH …), and only about six percent of the time did a mention of *gun* relate to a topic associated with advertisements for goods (IRON BRASS SILVER SORT DITTO LARGE WATCH STEEL PLATE SMALL POT GOLD …). Gun-related words (*gun*, *firearms*, *musket*, *pistol*) did not relate strongly to topics about crime or disasters, perhaps providing another approach to questions about the popularity and use of firearms in early America.

Relating multiple words to topics can also raise an array of research possibilities. As figure 3 shows, looking up the most likely topics in which the words *Cherokee* and *Negro* appear shows the literal marginalization of these groups from general colonial commentary. The word *Cherokee* appears in the newspaper only in discussions of Indian-related topics (INDIAN FORT MEN TOWN PARTY OFF KILLED PEOPLE CAME LETTER DAY FRENCH …). The word *Negro* appears primarily in advertisements that discuss servants and slaves (YEAR NEGROE MAN SHE WELL SERVANT AGE COUNTRY LIKELY ENQUIRE MASTER SOLD …), although the word secondarily appears in a category that describes disasters (GREAT FIRE SHE MANY DOWN TOWN HEAR NIGHT FOUND CITY DIED POOR …). *Woman*, on the other hand, appears in somewhat more varied categories, including one for runaway servants (AWAY REWARD SERVANT WHOEVER NAMED JACKET OLD HAIR SECURE PAID RUN PAIR …), yet significantly overlaps the topical appearances of *Cherokee* and *Negro*. Like *Negro*, *woman* tends to be mentioned as a descriptor in advertisements for laborers and in tales of disasters. *Woman* also appears in documents focused on the topic of Indians—perhaps suggesting the importance to colonists of gendered

interactions with Native Americans. Overall, the narrow subjects in which these words appeared may suggest new ways to think about how print discourses operated to marginalize particular groups in early America.

Word	Topic 1	Topic 2	Topic 3	Topic 4
Cherokee	Indian (100%)			
Negro	Servant/Slave (96%)	Disaster (4%)		
Woman	Servant/Slave (82%)	Disaster (10%)	Indian (4%)	Runaway (3%)

Fig. 3. Are marginalized groups marginalized in Pennsylvania Gazette topics?

These examples are meant to suggest the possibilities of topic modeling, not to provide conclusive determinations about these subjects. Rather, they show that the ability to quickly categorize the ability to quickly categorize the thematic appearance of various words can open new directions for investigation.

Tracking Topics over Time

Topic modeling can also chart the changing prevalence of each topic over time. Not surprisingly, a topic related to the kinds of political issues discussed at the founding of the United States (STATE GOVERNMENT CONSTITUTION LAW UNITED POWER CITIZEN PEOPLE PUBLIC CONGRESS RIGHT LEGISLATURE …) increased in prevalence when it is supposed to: in the Revolutionary and early national eras (fig. 4).

Fig. 4. Increase in commentary on **G**OVERNMENT topic in Revolutionary and early national eras

The ability to trace subject categories as they occur over time means that

scholars can track changing relationships of various topics. For instance, as figure 5 shows, the 1750s zenith in cloth advertisements occurs at approximately the same time as the nadir of religious content in the *Gazette*. Were colonists (or at least *Gazette* editors) choosing consumption over spirituality during those years? The topic model also shows similar (though not exactly parallel) increases in early national **R**ELIGION and **C**RIME topics (also fig. 5). Does this suggest heightened concerns over morality and civic virtue in the new nation? Again, only careful research into the details of these shifts can explain such trends, but topic modeling is already doing scholars a service by mapping changes over time that might not otherwise be obvious or easily accessible.

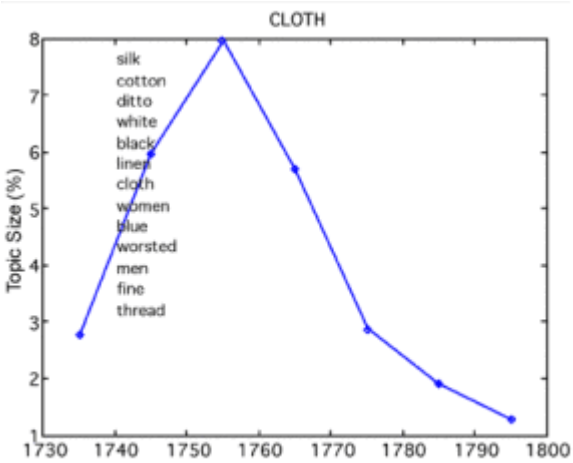


Fig. 5

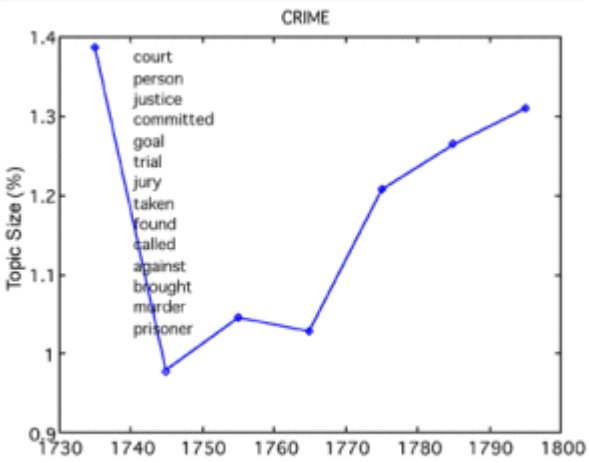


Fig. 5

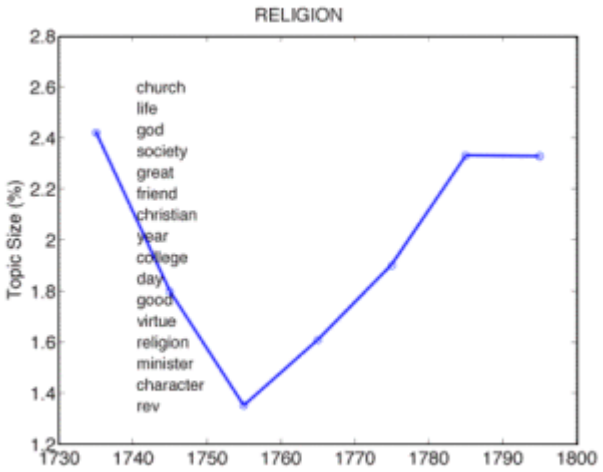


Fig. 5 Changing percentages of **C**RIME, **R**ELIGION, and **C**LOTH topics appearing in the Gazette over time

Conclusion

I hope that this brief foray into the possibilities offered by topic modeling shows how we can move beyond keyword searches to consider new methodologies that take advantage of the growing body of full-text resources. Topic modeling can provide a valuable sense of the contents of enormous sets of documents and can suggest answers to an array of questions about the relationships of words, texts, and historical subjects.

By combining the promise of digitization with cutting-edge topic-modeling technologies, we are no longer limited by the number of items that individual researchers can analyze. For instance, Charles Clark and Charles Wetherell's excellent (and undoubtedly painstakingly time consuming) 1989 analysis of the contents of the *Pennsylvania Gazette* sampled less than ten percent of the total number of articles in just a thirty-three-year period.

While I am a huge fan of the possibilities of this kind of topic modeling, I present it with one important caveat: topic modeling is only a tool. It requires historians' knowledgeable input and analysis. Unfortunately, it is also a tool that requires not only access to the text (rather than page images) of documents but the cooperation of a computer scientist. Should you not have a computer scientist on call (as I have lucked into), there are some indications that digital-document providers are beginning to consider the usefulness of topic modeling for their digital content. California Digital Library, for example, is currently working with seven partner institutions to create a central Website to access documents on the American West with subject headings created through topic modeling. Although much of this material is chronologically beyond the interests of some *Common-place* readers, this important project is already yielding new ways for scholars and educators to

use digital documents.

I am under no illusions that topic modeling will change historical research as we know it. Archives will still be visited and documents will still be read. But this new technique may allow scholars to use digital archives not just to *access* increasing numbers of documents but to *analyze* those documents in entirely new ways. I hope that early Americanists will be among the first to consider embracing such new techniques.

This work was supported by a University of California, Irvine, Academic Senate Council on Research, Computing, and Library Resources Grant.

Further Reading:

For an exemplary human analysis of the *Gazette,* see Charles Clark and Charles Wetherell, "The Measure of Maturity: *The Pennsylvania Gazette*, 1728-1765," *William and Mary Quarterly* 46 (1989), 279-303. For technical details on topic modeling, see David J. Newman and Sharon Block, "Probabilistic Topic Decomposition of an Eighteenth-Century Newspaper," *Journal of the American Society for Information Science and Technology,* forthcoming 57:5 (March 2006).

This article originally appeared in issue 6.2 (January, 2006).

Sharon Block is an associate professor of history at the University of California, Irvine. Her book *Rape and Sexual Power in Early America* will be published in 2006 by the Omohundro Institute of Early American History and Culture at the University of North Carolina Press. She is using topic modeling for her current project on colonial notions of beauty.