Computational Topic Models of the *Library Quarterly*

Cody Hennesy and David Naughton

abstract: This case study demonstrates the application of an unsupervised topic modeling algorithm to 7,773 English-language articles published in the *Library Quarterly* from 1931 to 2015. The analysis of 85 years of the journal's output follows an exploratory data analysis framework to generate novel hypotheses about the history of LIS using topic modeling, a method for identifying clusters of co-occurring words within large collections of text. The paper closely examines two topics that suggest differences in gender representation in the journal to propose and support a new hypothesis regarding the historical inclusion of gendered objects of study in LIS literature.

Introduction

Scholars across the humanities and social sciences have actively probed the utility of computational text analysis methods in their fields of study over the last dozen years. One common theme finds researchers employing a statistical technique called topic models to examine trends in scholarly literature representative of their disciplines. Topic models use algorithms to discover patterns of related words within text documents and thus reveal latent themes. Topic modeling algorithms, such as latent Dirichlet allocation (LDA), provide computational methods for examining conceptual patterns across a volume of material that would otherwise be unreasonably time-consuming for an individual scholar to read in a conventional manner. This study applies an LDA model to identify 40 discrete topics in the full-text corpus of the *Library Quarterly* from 1931 to 2015. The authors attempt neither to map the breadth of topical shifts in the journal over time nor to summarize the journal or the field of LIS. Rather, they explore outputs of the model for evidence of hitherto underreported or undiscovered trends. Building upon Justin Grimmer and Brandon Stewart's assertion that "unsupervised methods are valuable because they can identify organizations of

portal: Libraries and the Academy, Vol. 22, No. 3 (2022), pp. 745–768.

text that are theoretically useful, but perhaps understudied or previously unknown," the study follows Lauren Klein's characterization of topic modeling as "a technique that stirs the archive."[1] This stirring prompts a closer analysis of two topics that show a unique preponderance of gendered third-person pronouns in the *Library Quarterly*. Topic models here provide a novel mechanism for teasing out relationships between clusters of articles related to "great men" and "women librarians." By combining close reading with methods from natural language processing, a branch of computer science encompassing the analysis of human-generated text, the paper straddles the domains of LIS, computer science, and data science. The study illustrates the potential for the application of topic models to generate hypotheses in LIS.

Literature Review

Developed in 2003, LDA uses the co-occurrence of words in documents to identify latent topics across a corpus.[2] In one study applying topic models to over 30,000 abstracts from articles on women's history, Sharon Block and David Newman unpack what it means to identify latent topics: "Topic modeling learns subject categories without a priori subject definitions . . . The content of the documents—not a human indexer—determines the topics collectively found in those documents."[3] One advantage of this approach is to explore a large amount of content efficiently—nearly 8,000 articles in the case of the *Library Quarterly* corpus. Another benefit is that, as an unsupervised learning method, LDA does not look for predefined topics, but develops topic word lists based on a mathematical model. An LDA model will not generate more accurate topics than a human could, but the algorithmic view of the literature may stretch what we know about the history of a discipline.

Scholars in both technical and nontechnical disciplines have applied topic models to a variety of textual corpora, such as newspaper articles, historical documents, and social media.[4] A subgenre of these studies has focused on the scholarly output of specific disciplines, seeking to quantify long-term historical trends evident in journals from those fields. One pioneering work in this area applied topic models to 30 years of computational linguistics literature. Subsequent studies explored trends in the literature of classics, German studies, and the philosophy of science.[5] While many of these papers apply some variation of an LDA model to a corpus of scholarly journal articles, the findings and precise methods vary widely. One common characteristic, however, which this study follows, is the embrace of a dual purpose: first, to explore the output of the computational model in view of the history of an academic field and, second, to reflect upon the potential applications of topic models as a scholarly method within the discipline.

A number of works have used unsupervised methods, and LDA specifically, to explore the history of LIS. One of the earliest studies, from 2010, mapped latent topics in the titles and abstracts of 3,121 doctoral dissertations from North American LIS programs, showing that topics shifted drastically from 1930 to 2009.[6] Informetrics studies have used citation metadata from LIS publications to track changes in research impact, author citation trends, and subject coverage over time.[7] One of the largest studies involving LIS citations applied LDA to 92,705 titles and abstracts from the Library and Information Science Abstracts (LISA) database from 1978 to 2014. It found 19 topics

across four general areas: "processes, information technology, library and specific areas of information application."[8] Of the works involving full-text articles, one research group mapped 30 topics from 1,648 full-text articles from five LIS journals, including the *Library Quarterly*, between 2000–2002 and 2015–2017.[9] Their findings included that the Internet became a foundation of LIS research and that the overall diversity of topics decreased from the earlier to the later period of study. In 2018, Micah Saxton provided a "gentle introduction" to topic modeling using the journal *Theological Librarianship* as a sample corpus, illustrating LDA methodologies.[10] Around the same time, Manika Lamba and Margam Madhusudhan used topic models to analyze a corpus of 393 full-text articles from the *DESIDOC* [Defence Scientific Information & Documentation Centre] *Journal of Library and Information Technology* from 2008 to 2017, reflecting on recent trends in the LIS literature of India.[11] The current study applies LDA across a larger corpus of full-text LIS literature than has previously been attempted, focusing on the potential to engage in closer readings of specific topic outputs for hypothesis generation, rather than attempting to map topical shifts across the field more generally.

While a full technical exposition of LDA is beyond the scope of this article, the method works by "iteratively assessing probability distributions of words within topics and of topics within documents." It operates under the assumption "that topics are usually strongly expressed by few words and that documents only express a few topics at a time."[12] David Mimno highlights one of the advantages in utilizing topic models such as LDA for large quantities of text, noting that "in practice, assignments of topics that maximize these criteria"—that each document contains relatively few topics and each topic contains relatively few distinct word types—"are close to human understandings of the underlying concepts and linguistic categories in the corpus."[13] In other words, the topics identified by LDA often map well to ideas that humans understand. The ability to quantitatively track these concepts over time and to correlate them with specific primary sources provides a powerful analytic tool in the history of ideas.

Curiously, however, it may be more helpful to avoid thinking of the output of LDA as "topics" at all, but instead to view the findings more strictly as clusters of terms that co-occur in documents. As Andrew Piper notes:

> It is probably best to think of topic modeling not as a way to test "topics" in your documents, but as a way of generating insights about particular semantic behavior within them. This is a slight difference, but the key is to see the latter exercise as a form of "exploratory" data analysis rather than "explanatory." Topic modeling can reveal patterns and initiate questions, but it is less appropriate for testing and confirming them.[14]

The current study applies a topic model as a means of generating and gathering evidence related to a specific hypothesis about the representation of men and women as objects of study in LIS literature but does not purport to confirm the hypothesis.

The presence of gendered third-person pronouns in certain key topics, otherwise largely absent from the topic model, was a jumping-off point for a closer examination of those topics. Previous work has used LDA models to explore the research of men and women authors in computer science and used semi-supervised topic models to create gendered topics (feminine, masculine, and nonbinary) to assist in detecting gender bias.[15] But little appears to have been written specifically about the appearance and meaning

of gendered third-person pronouns in LDA models. Ted Underwood notes that "in topic-modeling fiction I find it useful to get rid of at least the most common personal pronouns, because otherwise the difference between 1st and 3rd person point-of-view becomes a dominant signal that crowds out other interesting phenomena." He also observes, "This sort of thing is very much a critical judgment call; it's not a science."[16]

Previous studies have found that women authors are underrepresented in many LIS journals: in the *Journal of the American Society for Information Science*, Ben-Ami Lipetz finds that women authorship grew from 21 percent in 1955 to 34 percent in 1995.[17] A similar study of *College & Research Libraries* reveals something closer to parity, though still unrepresentative of the demographics of the field, with just over 51 percent women authors from 1989 to 1994, up from 22 percent between 1939 and 1944.[18] Lois Buttlar identifies significant variation between journals, noting that women were authors in 78 percent of *School Library Media Quarterly* articles between 1987 and 1989, while only 25 percent of the authors in *Libraries & Culture* were women during the same period.[19] There is less clear quantitative evidence, however, of women librarians as objects of study in LIS. The topic model described here provides preliminary evidence that the quality and quantity of coverage of women and their work in the *Library Quarterly* have differed significantly from the treatment of men.

Methodology

Data Collection

The *Library Quarterly* (hereafter, *LQ*) was not selected to wholly represent the field of LIS, but the journal does reflect a prominent set of perspectives in LIS scholarship over a long period. *LQ* is the oldest publication devoted to library research in the United States. The journal operated "under the supervision" of a leading library education program, the Graduate Library School of the University of Chicago, from its founding in 1931 until the school's demise in 1989.[20] Later issues were published in cooperation with the LIS graduate programs at Indiana University Bloomington and UCLA (University of California, Los Angeles). The journal has continued to represent established perspectives in the field, offering high-quality original research, commentary, and reviews.

Digital content from *LQ* from 1931 to 2015 was readily available for computational analysis via JSTOR's Data for Research platform.[21] At the time of writing, JSTOR plans to sunset the Data for Research service, though many of its functions have been migrated to, and greatly enhanced at, a pilot Constellate website (https://constellate.org/). The Data for Research platform enables access to the thousands of digitized scholarly journals and books in JSTOR by providing tools for repurposing JSTOR's organizational markup and metadata—digital artifacts about the texts themselves—such that we can examine the digital editions of a journal as a data set. Metadata and ngrams for 8,808 items from *LQ* were retrieved from Data for Research using the query *jcode:library* on June 18, 2019, and regenerated on January 6, 2020, to replicate the analysis. At the time of retrieval in 2019, the most recent available issue was volume 85, issue 4, from 2015. Data for Research provides highly structured data: an XML file with metadata for each article (for example, volume, issue, year, title, author, and page numbers), along with a

tab-delimited list of the ngrams for each article. The term *ngrams*, or *n-grams*, refers to any sequential number (*n*) of items from a text. Crucially, these metadata allow us to track the shifting of topics in *LQ* over time, to identify articles in which topics are prevalent, and to understand the shape and size of the corpus itself.

The ngram files provided by JSTOR and used for this analysis contain a list of each word or token—that is, each sequence of characters with a known meaning—from an article, along with the number of times each token appears. For example, the first 10 ngrams from a 1999 *LQ* book review—"Cora Wilson Stewart: Crusader against Illiteracy"—capture the most frequently appearing words from the article: *her* (25 occurrences), *stewart* (22), *literacy* (17), *she* (16), *work* (10), *also* (9), *national* (9), *s*[22] (9), *from* (8), and *wilson* (8). This "bag-of-words" representation of a text is not intended for direct human evaluation but enables the application of methods such as LDA topic modeling, in which word order does not matter and topics are observed as collections of frequently co-occurring words at the document level.

Data Cleaning

After the ngrams and metadata files were downloaded from JSTOR, Thomas Klebel's *jstor* R package was used to reformat the metadata XML files into a single table containing key metadata for the entire corpus.[23] The metadata was imported into a Python environment and combined with the ngrams for each article, where it was cleaned and formatted to allow for topic modeling.[24] The corpus was reduced from an initial size of 8,808 items to 7,773 by removing content that was not of interest for analysis: namely, articles with titles of "Front Matter," "Back Matter," "Cover Design," and "Volume Information." Ngrams in the corpus were stemmed, reducing each word to a base, using the Natural Language Toolkit's Snowball Stemmer.[25] This stemming algorithm replaces such terms as *government*, *governs*, and *governmental* with a single token, *govern*, for example. A small set of stop words—common terms, such as prepositions—are already excluded from the ngrams available to download via JSTOR. To further reduce the amount of noise in the text to be analyzed, this study also excluded words that appear in more than 70 percent or fewer than 10 percent of the documents. The latter list of "rare" words is helpful for removing optical character recognition (OCR) errors, acronyms, terms in non-English languages, and uncommon names and terminology (for example, *cixxxii*, *tislaveri*, *alifornia*, *markowitz*, *rila*, and *taciturn*). The most "common" terms include those that appear too frequently in the corpus to assist in the analysis (for example, *librari*, *univers*, *book*). While similar studies often include pronouns in stop word dictionaries, thus removing them from analysis, the presence of gender pronouns such as *he/him/his* and *she/her/hers* in specific topic word lists can reveal objects of study that skew masculine or feminine in the corpus, as this article will later examine, and therefore were retained. Using pronouns as a proxy for the presence or absence of gendered objects is an admittedly imprecise approach. While it provides a useful hook for the computational analysis of traditionally narrow gender representations (men and women), it does not account for gender-nonconforming or nonbinary individuals.

Generating Topics: Tools

A list of 40 topics was generated by implementing LDA on the cleaned *LQ* corpus using Python's scikit-learn packages of software for machine learning and statistical modeling, including classification, regression, and clustering.[26] This process was performed using several key tools. First, *CountVectorizer* from the *feature_extraction* module transformed the ngram list for each article into a document-term matrix, a table containing the number of times each word in a corpus appears in every document. The topic model was then run against the document-term matrix using *LatentDirichletAllocation* from the *decomposition* module. Because LDA requires manually assigning the number of topics to look for in a corpus, *GridSearchCV* from the *model_selection* module was implemented to find the "best performing" topic model and parameters, including the number of topics (40) to apply to the *LQ* corpus. Some scholars recommend a more subjective approach, looking at a variety of sets of topics and choosing the number that seems the most coherent.[27] This study, however, was interested in the "machine's view" of the corpus, rather than creating a list of topics a reasonable human would expect to see. The machine's view is not intended as a neutral or objective measure (because it is not one), but rather as a mathematical abstraction of text that has the potential to reveal patterns humans might not otherwise have expected to see, including topics that we may not even initially understand. Note, too, that this 40-topic model by no means represents the total number of substantive topics discussed in the pages of *LQ*.

Analyzing Topics: Outputs

Three primary outputs of the topic model were created to assist in analysis: (1) lists of topic words, (2) the five most prevalent articles for each topic, and (3) plots of the prevalence of the coverage of each topic in *LQ* over time. To better explain the methods for generating each output of the model, it will help to illustrate the outputs using a relatively straightforward topic from the findings, topic 17, *Bibliographic classification*. It is not the authors' intention here to analyze the topic itself, but merely to describe the methods in context.

While topics identified by LDA are often sensible to humans, they are not automatically labeled by the LDA tool itself, nor are they presented in any meaningful order (topic 1 is no more important than topic 40). Rather, the LDA model can be understood as a "topic word distribution . . . that represents the number of times word *j* was assigned to topic *i*."[28] The next step in the analysis of the model, then, consisted of the first author reviewing the top words for each topic and manually assigning labels for each. The top 10 tokens for topic 17 in the analysis of *LQ* are *classif*, *subject*, *system*, *class*, *scheme*, *classifi*, *general*, *number*, *arrang*, and *divis*. Based on these terms, one might manually assign a label, such as *Classification*. To label each topic as accurately as possible, however, it is also helpful to consult a list of the articles in which the topic words are most prevalent. To generate those lists, the authors applied the *LatentDirichletAllocation transform* function to the document-term matrix, creating a document topic distribution that shows the proportion of terms from each topic in every article. The articles in which topic 17 terms were most prevalent were "A Classification for Medical Libraries" (1936), "A System of Bibliographic Classification" (1936), "Classification for Works on Pure and Applied

Science in the Science Museum Library" (1937), "Classification for International Law and Relations" (1971), and "Colon Classification" (1934). Most of these articles are from a pre-digital era and appear to concern the classification of books, specifically, and so a more accurate label for the topic was determined to be *Bibliographic classification*.

The group of articles from the 1930s that are associated strongly with topic 17 demonstrates a potential weakness of topic modeling. As Alan Beye Riddell explains: "LDA assumes that association of words with a topic does not vary over time. In other words, LDA assumes scholars are using the same collection of words to talk about folktales in the year 1940 as in the year 2000. We know this is wrong."[29] The collection of words that LIS scholars have used to talk about bibliographic classification have varied over time. Topic model studies sometimes address this by running a topic model on some temporal subset of the corpus (for example, each year of the journal) and then assigning labels that reflect the authors' understanding of the topics, considering changes in terminology. This study does not utilize this approach since the purpose is not to map the breadth of topics in the journal over time, but rather to explore underlying semantic patterns for hypothesis generation.

There is still value in visualizing the prevalence of topics in the journal over time, however. Plots were created to display the shifting coverage of each topic in *LQ*, mapping their prevalence relative to the total number of words in the journal each year. These plots offer timelines of the rise and fall of specific clusters of terms, not only reflecting the shifting terminology and jargon of LIS, but also highlighting particularly stable groups of terms in *LQ*. Returning to topic 17, for example, Figure 1 shows a significant drop in coverage of the *Bibliographic classification* topic from 1960 to present (see Figure 1). The conversation related to classification likely continued or increased during that time—using language related to search and information retrieval systems, for example—but those articles use terms that co-occur infrequently with the terms in Topic 17. The decreasing coverage observed in Topic 17 reflects a shift in the semantic expression of classification as a topic discussed in *LQ*. The terms that scholars use to discuss a field are important and worthy of study, and shifts in the co-occurrence of those words and phrases, as well as their relative stability, can signal interesting patterns in the scholarly conversation. To provide broader context about the corpus, the authors also generated a list of the most common terms in the entire *LQ* corpus and plotted the sums and means of *LQ* issues per year.

Findings and Discussion

The full corpus includes 342 issues of *LQ*. While an average of 91 articles were published per year across the entire collection, the number of articles in the journal declined steadily between 1993 and 2010, except for a significant bump in and around 1999 (see Figure 2). While *LQ* consistently published one issue per quarter throughout its history, the average number of articles per issue peaked in 1976 with 36 articles and reached a nadir in 2010 with only 9 articles per issue. Plots of specific topics over time (see Figures 1 and 3) reflect the total number of words per year and track prevalence relative to the overall word count.

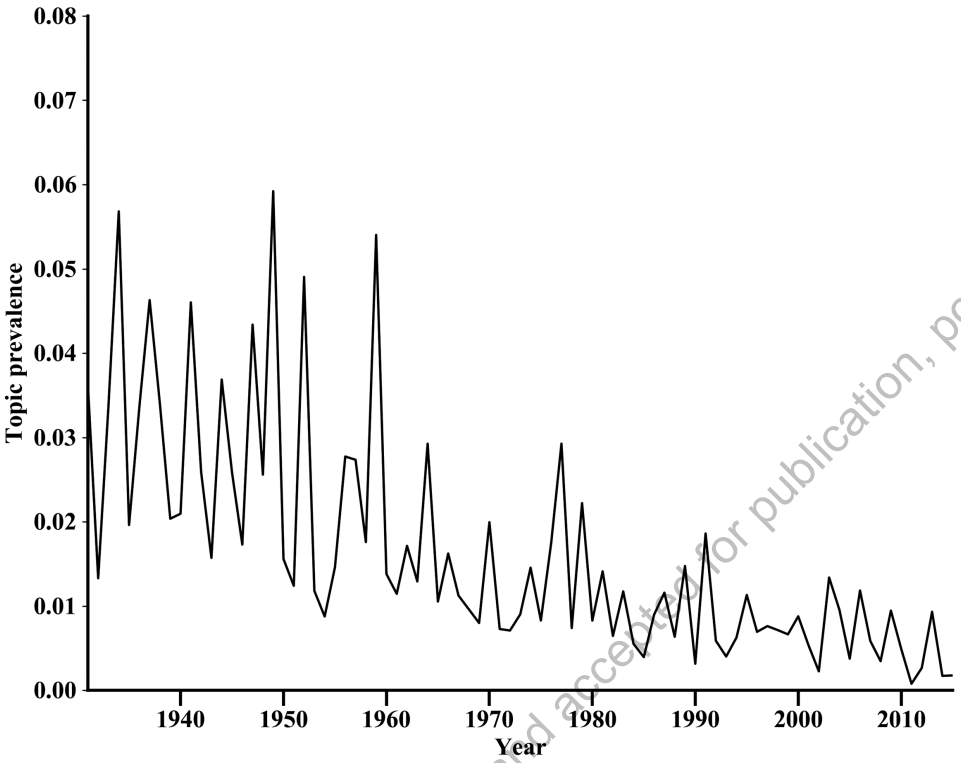

Figure 1. The prevalence of topic 17, *Bibliographic classification*, in the *Library Quarterly* by year from 1931 to 2015. Prevalence here is calculated by dividing the sum of the prevalence of topic words for topic 17 for each year (from the document topic distribution) by the total number of words in the journal for the same year.

While 40 topics do not represent the breadth or depth of *LQ*'s output, a careful examination of those 40 topics would still require a book-length analysis. The Appendix to this article includes a table of the labels and top 30 terms for all 40 topics, in the order of the topics' overall prevalence in the corpus. Overall, many general trends are unsurprising, especially when considering the growth or diminishment of the coverage of a topic over time.[30] Many topics related to print collections and books, for example, showed marked decreases: *Manuscripts and printing* (topic 1), *Circulation of library collections* (topic 3), *Reference materials* (topic 8), *Bibliographic classification* (topic 17), *Reading and readers* (topic 25), *Book lists* (topic 28), *Cataloging* (topic 33), and *Books for college students* (topic 39). Inversely, many topics with increasing prevalence relate to the introduction of new technologies: *Bibliometrics and citation analysis* (topic 2), *Information-seeking behavior* (topic 21), *Information technology* (topic 22), *Media and communications* (topic 26), and *Information retrieval tools* (topic 38). For both increased and decreased prevalence, these shifts align neatly with the changing vocabularies of library conversations over the course of the twentieth and early twenty-first century.

The appearance of these recognizable topics and the confirmation of some of our expectations about the attention paid to them over time suggest some level of accuracy in

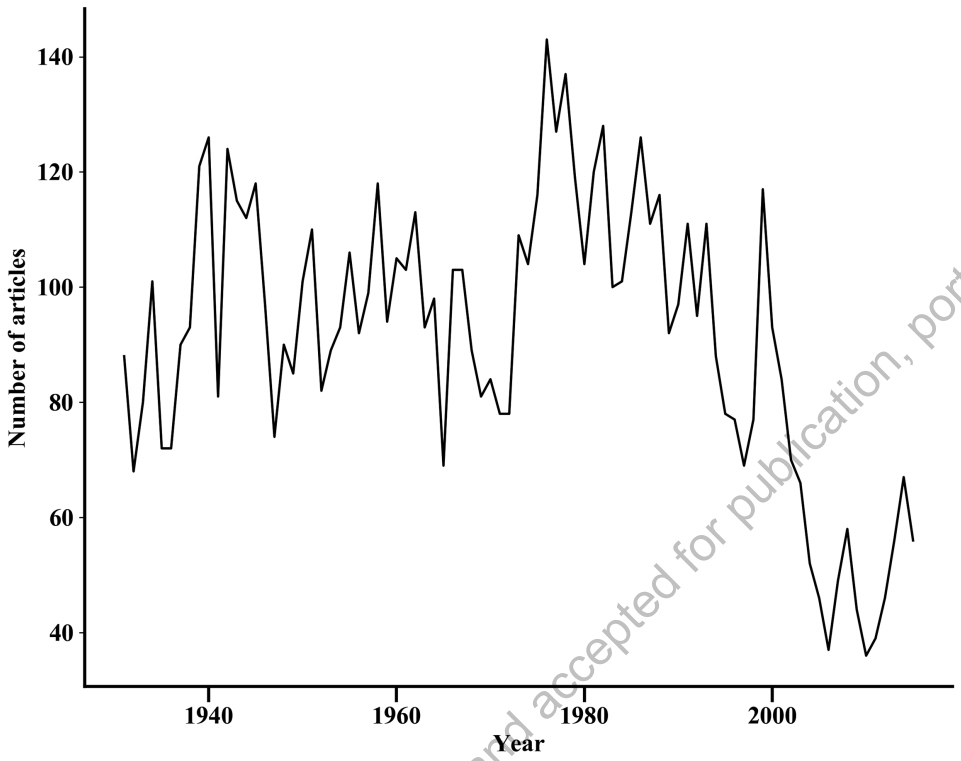

Figure 2. The number of articles published in *Library Quarterly* per year from 1931 to 2015.

the LDA process. Nan Z. Da, however, points to a tension in looking to "the obvious" to confirm the accuracy of computational methods in literary studies, noting, "The problem with computational literary analysis as it stands is that what is robust is obvious (in the empirical sense) and what is not obvious is not robust."[31] The 40 topics identified in the model, however, display a mix of the obvious with many less easily predicted or readily apparent clusters of semantic expressions.

Women Librarians and *Great Men*

The findings in this section focus on a closer evaluation of two topics that feature third-person pronouns and other gender signifiers in their topic word lists: topic 10 (*Women librarians*) and topic 37 (*Great men*). While the general trends suggested by these topics initially seem to fall into the category of the "obvious," closer examination reveals intriguing patterns that invite us to explore new hypotheses about the history of scholarship in *LQ*. Key outputs from the model for these two topics are shown in Table 1, which lists the top 30 tokens for each topic; Table 2, which includes five articles for each topic in which the topic words are most prevalent; and Figure 3, which charts the relative prevalence of topic word clusters in the *LQ* corpus over time.

Table 1.

Top words* for topic 10 (*Women librarians*) and topic 37 (*Great men*) in *Library Quarterly*, 1931 to 2015

Topic 10 (Women librarians)	Topic 37 (Great men)
her she women who miss men mari had	his he had were who year time him when after
librarian york when famili friend first illinoi	first did great could letter man later would
person while also posit although after would	them also made two mani during day own
own home were time where year career did	before john write dr

*The top words for each topic were created using the topic word distribution tool *LatentDirichletAllocation* module from the Python scikit-learn package.

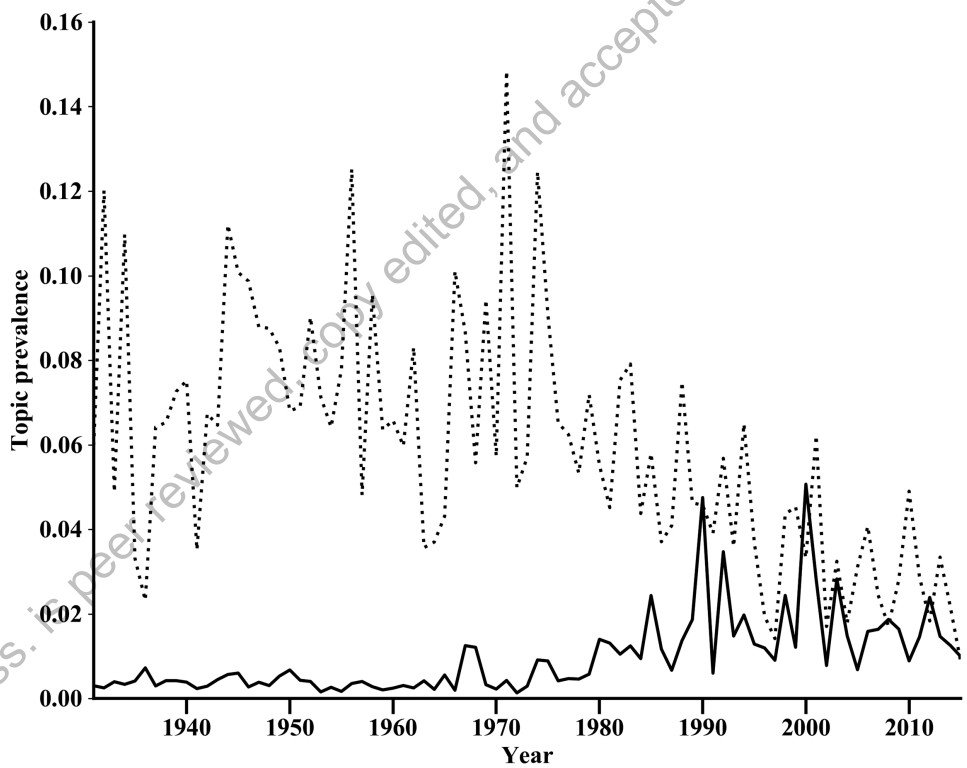

Figure 3. The prevalence of topic 10—*Women librarians*, shown by the solid line—and topic 37—*Great men*, indicated by the dotted line—in the *Library Quarterly* by year from 1931 to 2015. Prevalence here represents the sum of article prevalence scores for each year, divided by the total number of words in the journal for the same year. The maximum value of the x-axis in this figure is 0.16, whereas in Figure 1 it is 0.08, so these visualizations are at different scales.

Table 2.

Top articles for topic 10 (*Women librarians*) and topic 37 (*Great men*) in *Library Quarterly*, 1931 to 2015

Article title	Year	Type	Prevalence*
Topic 10 (*Women librarians*)			
"'No Philosophy Carries so Much Conviction as the Personal Life': Mary Wright Plummer as an Independent Woman"	2000	Article	0.420813901
"Fervent and Full of Gifts: The Life of Althea Warren"	1962	Book review	0.417992765
"Librarian/Library Educator: An Autobiography and Planning for the Future"	1988	Book review	0.384160036
"Remembering Connie Van Fleet"	2014	Article	0.36872196
"Cora Wilson Stewart: Crusader against Illiteracy"	1999	Book review	0.367725466
Topic 37 (*Great men*)			
"Francis Bacon and the King's Printer"	1952	Article	0.727749221
"Michael Joseph: Master of Words"	1987	Book review	0.709463877
"The Book-Peddling Parson"	1985	Book review	0.677837
"Mr. Hanson and His Friends"	1934	Article	0.664858114
"The Astonishing Mr. Scripps: The Turbulent Life of America's Penny Press Lord"	1993	Book review	0.648417759

**Prevalence* here refers to the "document topic distribution" provided by the *LatentDirichletAllocation* module's *transform* function. This essentially adds up the proportion of the top words for each topic that appear in every article in the *Library Quarterly* corpus.

Topic 10 contains more gender-specific terms than any other topic in the analysis: the three most common words in the topic are *her*, *she*, and *women*, and the word list also includes *miss* and *men*. Topic 10 is, in fact, the only topic that has feminine pronouns, which is particularly notable as third-person pronouns are high-frequency terms in the corpus overall. While gender clearly plays an important role in the topic, looking at the list of articles in which the topic is most prevalent reveals more about the meaning and coherence of the topic across the *LQ* corpus. Four of the five top items concern prominent women librarians (Mary Wright Plummer, Althea Warren, Connie Van Fleet, and Martha Boaz), and the fifth is a book review of a biography of educator Cora Wilson Stewart. The topic points toward a subgenre of biographic treatments of women librarians and educators that includes both book reviews and articles. Nine of the 10 articles most strongly associated with the topic are biographical. It is reasonable, then, to label the topic as *Women librarians*, though a more comprehensive label might be *Lives and careers of women librarians*. As one would suspect of biographical content, the topic shows a focus on accomplishments from the past. Top words associated with the topic include such past-tense verbs as *had*, *would*, *were*, and *did*.

Charting the prevalence of the topic in *LQ* over time, we observe a clear increase in coverage during the 1980s and 1990s, followed by a drop again after 2000, though not to levels as low as those in the first four decades of *LQ*'s publication (see Figure 3). The precipitous rise in scholarly recognition of the contributions of women in the field during the last two decades of the twentieth century maps to what we know of the growth of feminist perspectives in the academy and in LIS since the 1960s. But this trend, which we would expect to continue to the present, is not reflected in the coverage of the topic, which decreases after peaking around the year 2000. It is helpful to keep in mind that we are observing a trend related to the prevalence of clusters of specific terms associated with a topic that we have chosen to label as *Women librarians*. Coverage of women, women librarians, or both in *LQ* may have continued to steadily rise in the twenty-first century. What has decreased is the co-occurrence of these specific terms. This suggests that the words authors in *LQ* have used to discuss women librarians (and the frequency with which they discussed them) may have shifted significantly in the 1980s and 1990s. A closer look at other words in the topic—especially the presence of terms related to domesticity (*home*) and relationships (*famili*, *friend*)—suggests several possible hypotheses. The terms *home* and *friend* do not show up in any other topics in the LDA model, while *famili* only shows up again in a topic concerning *Children's literature*. Did late twentieth-century coverage of women librarians in *LQ* focus on issues of domesticity and relationships in ways that articles on other subjects (including men) did not? If so, has that kind of coverage decreased in the twenty-first century? What kinds of discussion of women's careers in LIS have replaced it? And how does this all fit into the many shifting scholarly conversations related to gender, feminism, leadership, representation, and affective labor in libraries?

The precipitous rise in scholarly recognition of the contributions of women in the field during the last two decades of the twentieth century maps to what we know of the growth of feminist perspectives in the academy and in LIS

While this paper will not answer these questions—the topic model does not provide sufficient evidence to do so—it is worth considering indications from the LIS literature and examining how they interact with the model. For instance, the increasing coverage of the *Women librarians* topic at the end of the twentieth century could be understood as a scholarly response to Roma Harris's injunction in 1990 that "librarians who wish to stop the erosion of their profession must stop shunning the female traditions of library work."[32] Many of the librarians celebrated in the five most representative articles for the *Women librarians* topic were high-ranking leaders in the field: two were American Library Association presidents, and one the president of the Association for Library and Information Science Education. We know that library leadership continues to display significant gender disparities and that gendered expectations for library leaders often diverge from the kinds of work performed in a profession largely composed of women.[33] Just as women have been underrepresented in library leadership and library information technology, the topic models suggest they have been scarce in the scholarly literature of LIS. A closer examination of a contrasting topic in *LQ*—topic 37 on *Great men*—illustrates key differences in how certain gender associations in the journal have developed.

We know that library leadership continues to display significant gender disparities and that gendered expectations for library leaders often diverge from the kinds of work performed in a profession largely composed of women.

Topic 37, which the authors have labeled as *Great men*, is the second-most prevalent across the entire *LQ* corpus, while topic 10, *Women librarians*, is the fourth lowest. In other words, the co-occurrence of the cluster of terms from the *Great men* topic is six times more common in *LQ* than those related to the *Women librarians* topic. While coverage of topic 37 has waned significantly in the last several decades (see Figure 3), the period of its lowest prevalence still roughly equals the height of coverage that *Women librarians* received in the 1980s and 1990s. Topic 37—which includes three masculine pronouns (*his*, *he*, and *him*) in its 10 most common words, as well as the word *man* and the given name *john*—is an important one to the *LQ* corpus.

Like the *Women librarians* topic, *Great men* displays a preponderance of terms related to the past (*had*, *were*, *year, time*, *when*, *after*, *did*, *could*, *later*, *would*, *made*, *during*, *before*), and the articles strongly associated with the topic are likewise largely biographical. A key difference, however, is that few of the articles associated with topic 37 are about librarians at all, but rather concern publishers, booksellers, bookbinders, and collectors. Only 1 of the 10 articles in which the topic terms are most prevalent is about an individual known as a librarian (James C. M. Hanson, in "Mr. Hanson and His Friends," 1934). More common are biographies of men associated with the European and American book trade from the seventeenth to the early twentieth century (for example, John Siberch, Mason Locke Weems, Michael Joseph, and Edward Scripps). Other articles associated strongly with the topic cover the book-collecting habits of famous scholars and politicians (for example, Francis Bacon, David Hume, and George Washington). One possible reading of this difference is that for women and their work to merit inclusion as objects of study

in *LQ*, they have more often required a direct connection to the field, while men from a wide variety of backgrounds are covered regularly and at greater length. This interpretation is supported by the fact that masculine pronouns appear in four topics—*Classics manuscripts* (topic 1), *Philosophy and organization of knowledge* (topic 16), *Bibliographies* (topic 23), and *Great men* (topic 37)—while feminine pronouns appear only in topic 10. Where masculine pronouns are found in these topics, they always occur in areas peripheral to library services and seem to rarely concern library workers who are men. The men under discussion more often are authors of classic manuscripts, publishers, philosophers, historians, politicians, and other scholars. Where previously women were rarely discussed in the pages of *LQ*, their inclusion at the turn of the twenty-first century appears to have depended on the performance of specific professional roles. Men from a wide range of professional and disciplinary roles continued to be discussed in *LQ*, while women were more often covered only if they were librarians, library administrators, or library educators. The model suggests that women were relevant to the field when they were librarians, while all "great" men were significant. A specific hypothesis based upon our reading of the topic models, then, is that the inclusion of women as objects of study in the LIS literature is more often contingent on their performance of library-related roles than that of men, who are included from a far wider range of professions.

Men from a wide range of professional and disciplinary roles continued to be discussed in LQ, while women were more often covered only if they were librarians, library administrators, or library educators.

Coverage of the *Great men* topic remains consistently high from the first issue of *LQ* through the 1970s, after which it begins a gradual but steady decline to the present. One explanation for this downturn is the diminishing importance of books and the history of books to the field of LIS. While the topic does not explicitly include terms related to the book trade and publishing, a closer examination of the articles show that the topic is more closely associated with books and publishers than it is with libraries or librarians. While print books have become less central to scholarly conversations in and about libraries than they were in the twentieth century, this pattern mirrors a shift in the scholarly recognition of individuals notable in and to the field. If *LQ* lionized men who were book publishers, printers, and collectors over the course of its first 40 years, whom does the field celebrate now? As LIS conversations in general shift toward library services, technologies, and communities, do we see a corresponding rise in biographies of the individuals performing leadership roles in those areas, who often emerge from beyond the disciplinary boundaries of LIS? And are there new gender differences arising in who is deemed worthy of inclusion and celebration in LIS literature?

One simple metric, the list of the most frequent words in the *LQ* corpus, paints a distressingly clear picture of the genders of the people discussed in the journal. The third and sixth most common words in the corpus are *his* (45,805 occurrences) and *he* (40,997), while *her* (11,458) and *she* (9,783) are the 257th and 315th most common. Masculine pronouns are used four times as often in the *LQ* corpus as feminine pronouns, a disparity

that is especially striking given that women have long dominated the ranks of library workers, even bearing in mind the historical (and problematic) use of masculine pronouns as generic third-person pronouns. An analysis of United States Census records shows that women in 2010 comprised 83 percent of librarians and have predominated in the profession since the 1880s.[34] As Harris noted in 1993, "For more than 100 years, library work in North America has been women's work."[35] While there have been roughly four women librarians for every man throughout much of the twentieth century, *LQ* has used masculine pronouns four times as often as feminine pronouns.

While there have been roughly four women librarians for every man throughout much of the twentieth century, LQ has used masculine pronouns four times as often as feminine pronouns.

Limitations and Directions for Future Study

Many of the general limitations of using topic models to study scholarly literature have been noted earlier. The authors have paid special attention to topic models' lack of explanatory power to accurately or comprehensively map the subject content of any corpus and have followed instead an exploratory data analysis model. A fundamental limitation of topic modeling for this study, then, is that the topic model alone cannot sufficiently confirm the hypotheses for which it provides initial evidence. The data generated by the topic model provide strong evidence for a number of hypotheses, but further research using different methods is necessary to confirm or disconfirm them. One compelling direction for future research, then, would be to analyze whether the inclusion of women as objects of study in the LIS literature more often hinges on their performance of library-related roles than does that of men. Laura Nelson's computational grounded theory provides one possible framework for further research along these lines. By applying qualitative close readings of the primary sources—what Nelson refers to as "computationally guided deep readings"—as well as a final pattern confirmation step to test the hypothesis using such methods as supervised machine learning, researchers might strengthen or deny the initial hypothesis detected using the inductive topic models.[36]

Another important limit to acknowledge is that the corpus analyzed in this article was a single journal from LIS. Early tests applying topic models to multiple LIS journals revealed topics that mapped too neatly to the specific domains of each journal to be of interest (for example, topics related to teaching and learning prevailed in the *Journal of Education for Library and Information Science*, but not in other journals). Further research across a broader corpus in LIS is critical for both the generation and confirmation of hypotheses, if any broader claims about the field are to be made.

Finally, future research on the specific claims highlighted here would require expertise on LIS history, certainly, but also knowledge of computer and data sciences and gender and women's studies. These fields, like all academic disciplines, operate according to methods and modes with long histories, which often conflict with one another and with those of other fields of study. LIS scholars with an interest and aptitude for

applying computational methods may still face significant challenges in contextualizing their results across disciplinary boundaries.

Conclusion

Topic modeling the *Library Quarterly* greatly reduces the complexity of the corpus. This reduction, on the one hand, oversimplifies a rich literature full of ideas, shifts, currents, and undercurrents, almost all of which pass beyond the realm of algorithmic detection. The authors make no claim to have captured a topical summary of the journal or of LIS, but rather have used topic modeling to explore the corpus and to pull on a few threads related to the inclusion of gendered objects in the journal. In that respect, topic models provide a novel mechanism for generating new hypotheses about the history of LIS.

Close examination of the *Great men* and *Women librarians* topics did not suggest a tidy narrative of increasing gender equity in the pages of *LQ* over time. Rather, the articles associated with the topics, alongside plots of the topics' prevalence over time, provide initial evidence to support the hypothesis that in the 1980s and 1990s, women would more likely appear in the journal for their roles in libraries, while men from a wider professional range were represented. There are, of course, hundreds of articles from *LQ* that are exceptions to these apparent trends, and the topics in no way capture the full expression of viewpoints and perspectives offered in the pages of the journal or by its authors and editors. These patterns do exist, however, and may prompt us—as authors, editors, reviewers, and readers—to more closely examine the ways in which gender representation in LIS continues to reflect and refract systemic power structures and ongoing gender inequities in the profession and beyond.

. . . gender representation in LIS continues to reflect and refract systemic power structures and ongoing gender inequities in the profession and beyond.

An increasing number of platforms that enable computational access to specific digital archives have been launched in recent years. JSTOR's Constellate platform, which will replace Data for Research, provides built-in tutorials and example code for running in-depth analyses of "text as data." The platform provides access not just to JSTOR content but also to documents from the Portico digital preservation service; CORD-19, scientific papers on COVID-19; Chronicling America, a collection of historic newspapers; DocSouth, files related to Southern history, literature, and culture; and the South Asia Open Archives. The HathiTrust Research Center provides various levels of computational access to over 16 million volumes in the HathiTrust Digital Library, content digitized from research libraries. As these tools become more widely available, and as the tools become easier to implement with the availability of high-quality tutorials and other adaptable online modules, scholars and library workers have much to gain from a closer acquaintance with such methods as topic modeling. Understanding both the potential promise and pitfall of these methods, to use Grimmer and Stewart's formulation in the title of their 2013 article, could unlock future directions in scholarship as well as provide new insights into the history of LIS.

Acknowledgments

The authors would like to thank Danya Leebaw and Jenny McBurney at the University of Minnesota, Twin Cities, for reading and providing feedback on an earlier draft of this paper.

Cody Hennesy is the journalism and digital media librarian at the University of Minnesota, Twin Cities; he may be reached by e-mail at: chennesy@umn.edu

David Naughton is a software developer at the University of Minnesota, Twin Cities; he may be reached by e-mail at: naughton@umn.edu.

Appendix

Latent Dirichlet Allocation (LDA) Topic Model Output for the *Library Quarterly*

Topic	Topic label	Top words	Prevalence
16	Philosophy and organization	may he can his what so we mean any would should must point object does of knowledge question problem knowledg theori natur even fact seem differ make practic critic do idea term	0.062726
37	Great men	his he had were who year time him when after first did great could letter man later would them also made two mani during day own before john write dr	0.060197
14	"Cooperative" collections	li tion brari librarian problem need special re co oper should materi ing may would research must field con can general area larg mani de ment possibl plan develop made	0.054475
4	Us and them	we our what do so can about us them who like my you would know how peopl when very even make much out say mani up now good way time	0.052938
31	LIS book reviews	chapter author discuss review reader section present volum page includ inform librarian topic interest provid essay text detail exampl mani refer part edit deal isbn cover describ does read eac	0.041179
24	Library communities	communiti learn cultur develop focus research particip provid how also practic institut inform role can need divers project through issu includ technolog within activ support about organ process understand academ	0.037679
18	"Books received"	paper isbn press york edit american associ chicago ed vol washington state scienc publish compil bibliographi school cloth guid studi co educ servic john research histori refer ix viii inform	0.036956
36	Survey studies	were percent studi tabl data differ number test respond group level sampl had survey between signific result two indic report measur rate those total each high factor statist rank relat	0.035289
6	International exchange	nation union countri were state unit intern congress american war had world organ tion foreign year develop report institut south program govern exchang committe establish activ region oper confer office	0.035230

1	Classics and manuscripts	print press edit centuri text page type copi manuscript first two were his illustr origin volum earli appear date form letter also made design so line note scholar same mani	0.034139
30	Histories	social cultur histori polit societi world histor centuri human intellectu chang power life knowledg econom modern peopl scienc who idea movement institut tradit our between historian what scholar press time	0.033775
8	Reference materials	languag english volum german list includ histori music literatur publish refer index period author edit titl under articl bibliographi name first entri mani translat subject dictionari general section compil year	0.031033
13	School libraries and teaching	school student colleg program educ cours faculti teacher graduat train teach studi instruct high librarian year time requir who institut field need standard should degre class prepar experi scienc master	0.030563
32	Social science methods	research studi analysi method social model relat data behavior approach problem relationship structur theori measur evalu process between effect investig scienc methodolog result differ area develop factor tion influenc two	0.025179
22	Information technologies	inform knowledg need research seek health sourc user individu provid process behavior access about informa how resourc understand search practic model concept mation organ can role exampl care may activ	0.025009
3	Public library use	per cent were circul tabl year total number group citi two averag ii tion period figur increas show popul three li had non time brari over between general five larg	0.024873
11	Associations and administration	librarian report associ committe administr member board organ ala manag staff director institut meet posit depart profession presid plan personnel were activ offic respons had group council year confer servic	0.024484
33	Cataloging	catalog entri card rule name head titl author under subject form code congress would file refer list main should record practic corpor number descript case when chang bibliograph differ may	0.024399
27	“Notable materials”	collect archiv materi record manuscript histori museum preserv histor research institut includ document also hold report volum item acquir special state paper relat rare ad acquisit import mani about file	0.023864

Appendix. Continued.

Topic	Topic label	Top words	Prevalence
34	Library services and programs	servic program communiti area plan branch staff provid survey serv need resourc agenc center collect region popul develop local level central evalu materi refer peopl mani educ activ can standard	0.023926
9	Collection use	cost time number per circul year statist each valu would total period loan estim figur measur volum hour item can model data rate may averag request materi increas unit budget	0.023000
15	College literary societies	american build were colleg societi room volum america centuri state harvard york boston plan first year histori two open li north earli institut space brari unit found collect mani place	0.020629
5	Library science education	educ librarian librarianship profession profess school american develop associ studi state academ scienc servic higher year train program graduat special status degre field colleg nation job role research career continu	0.019289
29	Book publishing	publish paper market price trade busi industri product econom press print firm compani distribut hous can may mani produc year increas cost copi small import would qualiti newspap profit high	0.018980
7	Government, law, and policy	state govern local citi feder agenc fund support unit educ offic grant report system board year commiss servic financi nation sourc general under aid establish polit depart school act regio	0.018054
20	Reference interviews	refer question librarian user answer patron you ask about what how help inform find student particip need were do would who when respons sourc know type studi correct time did	0.017309
17	Bibliographic classification	classif subject system class scheme classifi general number arrang divis materi order scienc section organ head refer each group special form tion practic under method relat main tabl period knowledge	0.016488
39	Recommended books for libraries	titl review list select number each were subject hold includ publish collect check period colleg note two three would chicago small re studi held four indic purchas view repres librarian	0.016101
21	Information-seeking behavior	inform knowledg need research seek health sourc user individu provid process behavior access about informa how resourc understand search practic model concept mation organ can role exampl care may activ	0.015681

2	Bibliometrics and citation analysis	journal scienc articl research citat scholar literatur scientif cite studi field publish author medic scientist disciplin review human inform sourc editor refer abstract period also physic serial academ societi number	0.015186
0	Children's literature	titl review list select number each were subject hold includ publish collect check period colleg note two three would chicago small re studi held four indic purchas view repres librarian	0.014813
25	Readers and reading	read reader interest group subject materi who peopl educ fiction select what may studi men about newspap magazin topic person popular appeal class author place differ each mani general time	0.014391
38	Information retrieval (indexes and thesauri)	index document term search subject relev user inform refer categori languag word abstract field exampl text system topic human each common relat can specif bibliograph number precis proper judg ment	0.012826
28	Booklists	british list london catalogu issu mr print publish american edit ii pamphlet year associ museum compil under copi bibliographi volum great vol societi seri may reprint titl supplement england annual	0.012055
35	Copyright and First Amendment	law right polici legal case act issu author access intellectu state civil materi any would copi should govern person free provid decis american against may limit restrict howev view general	0.011800
23	Bibliographies	bibliographi bibliograph art sourc studi critic literari subject graphic literatur scholar his period materi scholarship histor annot fine tool histori monograph guid catalog descript index list compil he primari publish	0.009860
10	Women librarians	her she women who miss men mari had librarian york when famili friend first illinoi person while also posit although after would own home were time where year career did	0.009526
19	Romance languages	de la der ii et french see die franc iv iii en al german un volum vol translat lectur im au document robert vi paul catalogu also part tion number	0.008596
26	Media and communications	media communic corpor center mass content specialist activ select effect industri compet materi product audienc concept intellectu time group set compani newspap organ firm statement tion cooper process contribut format	0.004308
12	Books for young adults (HIV/AIDS)	adult aid map age young older charact popul peopl york educ about increas includ relationship provid year also concern group what geograph discuss number current can interest who through begin	0.003824

Notes

1. Justin Grimmer and Brandon M. Stewart, "Text as Data: The Promise and Pitfall of Automatic Content Analysis Methods for Political Texts," *Political Analysis* 21, 3 (2013): 267–97, https://doi.org/10.1093/pan/mps028; Lauren Klein, "The Carework and Codework of the Digital Humanities," June 8, 2015, http://lklein.com/archives/the-carework-and-codework-of-the-digital-humanities/.
2. David M. Blei, Andrew Y. Ng, and Michael I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research* 3 (2003): 993–1022, http://dl.acm.org/citation.cfm?id=944919.944937. The term *Dirichlet* in LDA derives from Peter Gustav Lejeune Dirichlet (pronounced *dee ruh KLAY*), a German mathematician who made many contributions to number theory.
3. Sharon Block and David Newman, "What, Where, When, and Sometimes Why: Data Mining Two Decades of Women's History Abstracts," *Journal of Women's History* 23, 1 (2011): 81–109, https://doi.org/10.1353/jowh.2011.0001.
4. Paul DiMaggio, Manish Nag, and David Blei, "Exploiting Affinities between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of U.S. Government Arts Funding," *Poetics* 41, 6 (2013): 570–606, https://doi.org/10.1016/j.poetic.2013.08.004; Lauren F. Klein, Jacob Eisenstein, and Iris Sun, "Exploratory Thematic Analysis for Digitized Archival Collections," *Digital Scholarship in the Humanities* 30, suppl. 1 (2015): i130–41, https://doi.org/10.1093/llc/fqv052; Kyle W. Prier, Matthew S. Smith, Christophe Giraud-Carrier, and Carl L. Hanson, "Identifying Health-Related Topics on Twitter," in *Social Computing, Behavioral-Cultural Modeling and Prediction*, John Salerno, Shanchieh Jay Yang, Dana Nau, and Sun-Ki Chai, eds. (Berlin: Springer, 2011), 18–25.
5. David Hall, Daniel Jurafsky, and Christopher D. Manning, "Studying the History of Ideas Using Topic Models," in *EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing* (Stroudsburg, PA: Association for Computational Linguistics [ACL], 2008), 363–71; David Mimno, "Computational Historiography: Data Mining in a Century of Classics Journals," *Journal on Computing and Cultural Heritage* 5, 1 (2012): 1–19, https://doi.org/10.1145/2160165.2160168; Allen Beye Riddell, "How to Read 22,198 Journal Articles: Studying the History of German Studies with Topic Models," chap. 3 in *Distant Readings: Topologies of German Culture in the Long Nineteenth Century*, Matt Erlin and Lynne Tatlock, eds. (Rochester, NY: Camden House, 2014), 91–114; Christophe Malaterre, Jean-François Chartier, and Davide Pulizzotto, "What Is This Thing Called *Philosophy of Science*? A Computational Topic-Modeling Perspective, 1934–2015," *HOPOS: The Journal of the International Society for the History of Philosophy of Science* 9, 2 (2019): 215–49, https://doi.org/10.1086/704372.
6. Cassidy R. Sugimoto, Daifeng Li, Terrell G. Russell, S. Craig Finlay, and Ying Ding, "The Shifting Sands of Disciplinary Development: Analyzing North American Library and Information Science Dissertations Using Latent Dirichlet Allocation," *Journal of the American Society for Information Science and Technology* 62, 1 (2011): 185–204, https://doi.org/10.1002/asi.21435.
7. Erjia Yan, "Research Dynamics: Measuring the Continuity and Popularity of Research Topics," *Journal of Informetrics* 8, 1 (2014): 98–110; Kun Lu and Dietmar Wolfram, "Measuring Author Research Relatedness: A Comparison of Word-Based, Topic-Based, and Author Cocitation Approaches," *Journal of the American Society for Information Science and Technology* 63, 10 (2012): 1973–86, https://doi.org/10.1002/asi.22628; Erjia Yan, "Research Dynamics, Impact, and Dissemination: A Topic-Level Analysis," *Journal of the Association for Information Science and Technology* 66, 11 (2015): 2357–72, https://doi.org/10.1002/asi.23324.
8. Carlos G. Figuerola, Francisco Javier García Marco, and María Pinto, "Mapping the Evolution of Library and Information Science (1978–2014) Using Topic Modeling on LISA [Library and Information Science Abstracts]," *Scientometrics* 112, 3 (2017): 1507–35, https://doi.org/10.1007/s11192-017-2432-9.

9. Yosuke Miyata, Emi Ishita, Fang Yang, Michimasa Yamamoto, Azusa Iwase, and Keiko Kurata, "Knowledge Structure Transition in Library and Information Science: Topic Modeling and Visualization," *Scientometrics* 125, 1 (2020): 665–87, https://doi.org/10.1007/s11192-020-03657-5; Keiko Kurata, Yosuke Miyata, Emi Ishita, Michimasa Yamamoto, Fang Yang, and Azusa Iwase, "Analyzing Library and Information Science Full□Text Articles Using a Topic Modeling Approach," *Proceedings of the Association for Information Science and Technology* 55, 1 (2018): 847–48, https://doi.org/10.1002/pra2.2018.14505501143.
10. Micah D. Saxton, "A Gentle Introduction to Topic Modeling Using Python," *Theological Librarianship* 11, 1 (2018): 18–27, https://doi.org/10.31046/tl.v11i1.506.
11. Manika Lamba and Margam Madhusudhan, "Author-Topic Modeling of DESIDOC [Defence Scientific Information & Documentation Centre] Journal of Library and Information Technology (2008–2017), India," *Library Philosophy and Practice* 2593 (2019), https://digitalcommons.unl.edu/libphilprac/2593/.
12. Malaterre, Chartier, and Pulizzotto, "What Is This Thing Called *Philosophy of Science*?" For an in-depth yet accessible technical introduction to topic modeling and latent Dirichlet allocation, see Jordan Boyd-Graber, Yuening Hu, and David Mimno, "Applications of Topic Models," *Foundations and Trends in Information Retrieval* 11, 2–3 (2017): 143–296, https://doi.org/10.1561/1500000030.
13. Mimno, "Computational Historiography."
14. Andrew Piper, "What Are Some of the Problems with Topic Modeling?" *The Fish and the Painting* (blog), https://r4thehumanities.home.blog/what-are-some-of-the-problems-with-topic-modeling/.
15. Adam Vogel and Dan Jurafsky, "He Said, She Said: Gender in the ACL Anthology," in *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries* (Stroudsburg, PA: ACL, 2012): 33–41, https://www.aclweb.org/anthology/W12-3204.pdf; Hannah Devinney, Jenny Björklund, and Henrik Björklund, "Semi-Supervised Topic Modeling for Gender Bias Discovery in English and Swedish," in *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, Marta R. Costa-jussà, Christian Hardmeier, Will Radford, and Kellie Webster, eds. (Stroudsburg, PA: ACL, 2020), 79–92.
16. Ted Underwood, "Topic Modeling Made Just Simple Enough," *The Stone and the Shell* (blog), April 7, 2012, https://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/.
17. Ben-Ami Lipetz, "Aspects of *JASIS* Authorship through Five Decades," *Journal of the American Society for Information Science* 50, 11 (1999): 994–1003.
18. James L. Terry, "Authorship in 'College & Research Libraries' Revisited: Gender, Institutional Affiliation, Collaboration," *College & Research Libraries* 57, 4 (1996): 377–83, https://doi.org/10.5860/crl_57_04_377.
19. Lois Buttlar, "Analyzing the Library Periodical Literature: Content and Authorship," *College & Research Libraries* 52, 1 (1991): 38–53, https://doi.org/10.5860/crl_52_01_38.
20. Howard W. Winger, "A Salute to Past Editorial Boards," *Library Quarterly* 60, 4 (1990): 289–99, https://doi.org/10.1086/602262.
21. JSTOR, "JSTOR Data for Research," 2021, https://www.jstor.org/dfr/.
22. JSTOR's ngrams do not assign a part-of-speech tag based on a contraction's underlying grammatical form but rather capture *s* as present in such contractions as *she's* or *let's*.
23. Thomas Klebel, "jstor: Import and Analyse Data from Scientific Texts," *Journal of Open Source Software* 3, 28 (2018), https://doi.org/10.21105/joss.00883.
24. The complete code used for data cleaning and analysis is available at: https://github.com/chennesy/dfr_lq.
25. Steven Bird, Ewan Klein, and Edward Loper, *Natural Language Processing with Python* (Sebastopol, CA: O'Reilly Media, 2009).
26. Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al., "Scikit-Learn: Machine Learning in Python," *Journal of Machine Learning Research* 12 (2011), 2825–30, http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.

27. DiMaggio, Nag, and Blei, "Exploiting Affinities between Topic Modeling and the Sociological Perspective on Culture"; Grimmer and Stewart, "Text as Data."
28. scikit-learn developers, "sklearn.decomposition.LatentDirichletAllocation," 2020, https://scikit-learn.org/stable/modules/classes.html.
29. Riddell, "How to Read 22,198 Journal Articles."
30. A document including the top articles associated with each topic and plots of each topic's prevalence over time is available in the GitHub repository for this project: https://github.com/chennesy/dfr_lq/blob/main/output/topic_outputs.pdf.
31. Nan Z. Da, "The Computational Case against Computational Literary Studies," *Critical Inquiry* 45, 3 (2019): 601–39, https://doi.org/10.1086/702594.
32. Roma M. Harris, "Gender, Power, and the Dangerous Pursuit of Professionalism," *American Libraries* 24, 9 (1993): 874–76.
33. Jessica Olin and Michelle Millet, "Gendered Expectations for Leadership in Libraries," *In the Library with the Lead Pipe*, November 4, 2015, http://www.inthelibrarywiththeleadpipe.org/2015/libleadgender/; Christina Neigel, "LIS Leadership and Leadership Education: A Matter of Gender," *Journal of Library Administration* 55, 7 (2015): 521–34, https://doi.org/10.1080/01930826.2015.1076307; Melissa Lamont, "Gender, Technology, and Libraries," *Information Technology and Libraries* 28, 3 (2009): 137–42, https://doi.org/10.6017/ital.v28i3.3221.
34. Andrew A. Beveridge, Susan Weber, and Sydney Beveridge, "Librarians in the United States from 1880–2009," *OUPblog*, June 20, 2011, https://blog.oup.com/2011/06/librarian–census/.
35. Harris, "Gender, Power, and the Dangerous Pursuit of Professionalism."
36. Laura K. Nelson, "Computational Grounded Theory: A Methodological Framework," *Sociological Methods & Research* 49, 1 (2017): 3–42, https://doi.org/10.1177/0049124117729703.