Kindles, card catalogs, and the future of libraries: a collaborative digital humanities project

Anna L Neatour
J Willard Marriott Library, University of Utah, Salt Lake City, Utah, USA

Elizabeth Callaway
College of Humanities, Environmental Humanities, University of Utah, Salt Lake City, Utah, USA

Rebekah Cummings
J Willard Marriott Library, University of Utah, Salt Lake City, Utah, USA

Introduction

Librarianship is a profession that often inspires commentary both from practitioners within the profession and the public who use libraries. For librarians keeping up with the field, the literature is often engaged with predicting the effects of culture, policy or technology on libraries, sometimes with a great deal of hyperbole. For this article, two librarians and a digital humanities researcher formed a research team to determine if the digital humanities technique of distant reading through topic modeling would reveal interesting patterns in a Digital humanities project corpus of library-themed literature engaged in predicting the future and/or demise of libraries. There were two motivations behind this project. First, we wanted to determine if there were any trends in this body of literature on the future of the library. Questions considered included the following: Were there any trends of pessimism or optimism over the years? Who is engaged in publishing about this topic? Were there any cyclical moments in the way commentary about libraries engages with technology? Second, we prototyped a collaboration at our institution where librarians worked with digital humanities researchers as full partners on a digital humanities research project. Multiple colleges at the University of Utah have recently joined together to create an interdisciplinary space, the Digital Matters Lab, in the J. Willard Marriott Library, and this project also served as a model for cross-disciplinary collaboration in an emerging area of scholarship. We believe that direct experience with digital scholarship workflows and outcomes can better position all of the collaborators to facilitate further work in this area at our University.

Literature review

In December 2017, a 24/7 Wall Street article named libraries and archives as one of the “25 Dying Industries” in the USA with job attrition over the past ten years second only to the video tape- and disk-rental industry, the implication being that a future career in libraries was roughly as promising as a career at your local Blockbuster (Stebbins and Sauter, 2017). The response from the library community was swift and incredulous. American Library Association President Jim Neal tweeted that the data in the article, which claimed an 80 per cent decrease in employment from 2006-2015, were “grossly inaccurate, indeed laughable” (Neal, 2017). Inside Higher Ed identified the source of the
data inaccuracies and brought it to the attention of 24/7 Wall Street, who has since published a retraction (McKenzie, 2018). That 24/7 Wall Street did not question the remarkable nature of the data at first glance speaks to the persistent fantasy that books and libraries are being replaced by Google and free, ubiquitous digital content.

Forecasting the death of libraries is not a new phenomenon, in fact the death and demise of the library is a recurring theme in both the popular and academic presses. Despite the fact that both library visits and library employment have remained stable over the past decade, forecasting the “death of libraries” at the hands of technology has persisted as a common trope (Bruccoli, 2007; McTernan, 2016; MIT Technology Review Staff, 2005; Nichols, 2010; Shatzkin, 2011; Smith, 2014; Sorcinelli, 2016; Worstall, 2014). A 1992 New York Times article claimed “you will often hear it said that the print medium is a doomed and outdated technology, a mere curiosity of bygone days destined soon to be consigned forever to those dusty unattended museums we now call libraries” (Coover, 1992). Similarly, a 2007 article in College and Research Libraries warned that if librarians do not adapt “our campuses will invest in other priorities and the library will slowly, but surely, atrophy and become a little used museum of the book” (Lewis, 2007).

Other literature left the library intact, while predicting the death of particular services like collection building, reference or the library catalog (Anderson, 2015; Applegate, 2008; Liu, 2017).

Even within library literature, where the tone predictably bends towards optimism, there is a notable pattern of predicting either “The Death of the Library” or “The Library of the Future” (Schwartz, 2016). Library of the Future conversations in 2018 often include discussions about virtual reality, library-as-place, eliminating fines, research data services, makerspaces, gaming and civic engagement (“Symposium on the Future of Libraries | ALA Midwinter”, 2018). Previous Library of the Future conversations included the rise of the eBook, mass digitization, social media, online chat and compact shelving for physical collections (Darnton, 2009; Lewis, 2007). A second trend in library literature is documenting – and sometimes eulogizing – changes such as the decline of circulation statistics or the obsolescence of formats once heavily invested in such as the card catalog, microfilm and 35 mm slides (Applegate, 2008; Braunstein, 2013).

To look at this type of long-ranging and ongoing conversation, we turn to the method of topic modeling. Our work participates in a recent flourishing of humanities research that uses topic modeling as a new method for exploring large corpuses of documents. For many years, topic modeling practice in the humanities was primarily disseminated through informal distribution platforms. In 2012, Elijah Meeks and Scott Weingart, stated about topic modeling:

the work is almost entirely represented in that gray literature [ . . . ] in order to follow research that deploys topic modeling for humanities inquiry you must read blogs and attend conference presentations and workshops (Weingart and Meeks, 2012).

However, since that time, and thanks in part to that special issue that Meeks and Weingart edited, several topic modeling articles have been published in peer-reviewed journals. In 2013, for example, Poetics: Journal of Empirical Research on Culture, the
Media, and the Arts ran its own special issue on topic modeling. This collection featured articles on violence in Quing China (Miller, 2013), the disciplinary development of demography (Marshall, 2013) and news articles in relation to elevated national threat levels (Bonilla and Grimmer, 2013), among other treatments of topic models. More recently, Christopher Schoch has pushed the method to see if topic modeling can capture concepts such as genre in French classical and enlightenment drama (Schoch, 2017). Within the library literature, the engagement with topic modeling has been more modest. A large scale survey of the library literature was completed by Carlos Figueroa and collaborators in 2017, where titles and abstracts of articles appearing in the database of Library and Information Science Abstracts (LISA) from 1978-2014 were modeled, revealing 19 topics across a large corpus of library and information science literature (Figuerola et al., 2017). A survey of the library literature in Korean with a similar focus on analyzing titles and abstracts of library literature articles can be found in the work of Park and Song (2013). Another approach can be found in the work of Moore (2017), who used topic modeling and sentiment analysis as a method of enhancing assessment of LibQual survey responses.

In addition to surveys, topic modeling has also been used multiple times by library science researchers as a technique to improve discovery for users. Cain (2016) used topic modeling methods to illustrate how librarians may enhance access to a collection of electronic government documents. Hagedorn et al. (2011) investigated the possibility of topic modeling providing interface enhancement to large-scale digital collections with an instance of HathiTrust as a corpus, generating topics along with labels, then testing the results of an experimental interface built with topics with digital library users at Yale University and the University of Michigan. Muresan and Harper (2004) also explored the idea of topic modeling and accessibility to large-scale document collections by running experiments with a large corpus of documents from the Financial Times and experimenting with topic modeling as a method of providing mediated access to the documents. The most thorough example of research in this area can be found in the work of David Newman and collaborators, where they systematically evaluated topic modeling techniques applied to different types of digital collections such as Hathi Trust and Internet Archive book collections, NY Times articles, NIH grant abstracts and metadata records of digital items from Yale and the University of Michigan. Topics generated for the various types of digital collections were then subject to human scoring and evaluation (Newman et al., 2010).

Topic modeling in the library literature is primarily concentrated on modeling abstracts and titles in a large corpus and as a means of improving access to digital collections through enhanced faceting or user interface design. We were unable to find any other scholarship that topic modeled the full-text of a curated sub-corpus of articles published in the LIS literature, suggesting that there is a great deal of work left to be done investigating the knowledge production coming out of the library field.

What is topic modeling?
To read this corpus of articles forecasting and lamenting the future of libraries, we used the digital humanities technique of topic modeling. Explaining what topic modeling is has become something of a genre in itself with many accessible and comprehensive explanations written by prominent digital humanists.

In simple terms, topic modeling is an automated way to examine the contents of a corpus of documents. It is a form of text mining in which we can use a class of computer programs to extract “topics” from texts. A “topic” is simply a group of words that tend to co-occur, but the model’s outputs often looks like a topic in the normal sense – a group of words that are linked by subject or concept. Therefore, feeding in a corpus of articles from the Associated Press[1] might give you topics like “military, defense, force, pentagon, army, navy” and “environmental, plant, waste, energy, pollution, epa.” Articles associated with the first topic would probably be about military news and articles from the second topic would be about environmental news. But in each case, these are just words that happen to co-occur within the same document. In topic modeling, word order, syntax and grammar are not considered, just co-occurrence. Each document is assumed to comprise multiple topics. Occurrences of words are assigned to a topic based on how often instances of that word occur with each topic and how common that topic is in the rest of the document. Take the process of sorting one instance of the word “seal”, for example. Based on probabilities of words we had already sorted into topics, we would know that seal has a chance of occurring in both the military topic and the environmental topic (Navy SEALs and baby seals would be examples of the same word in different topics). To decide on this occurrence of the word, we would also have to consider how common the environment topic is in the rest of the current document before we assign that instance of seal to the environmental topic[2]. In reality, every word is in every topic, just in varying probabilities. The entire vocabulary is in each topic over a probability distribution. “Military” is in the environmental topic, just with a much lower probability than words like “waste.”

For this article, we used MALLET for R (McCallum, 2002), which uses latent Dirichlet allocation (LDA) style topic modeling developed by Blei et al. (2003). This is the form of topic model that continues to be the most popular among practitioners, though there is a wide variety of different styles of topic modeling. The outputs of the model are an ordered list of word probabilities relating to each topic (taking the top 10-100 most highly probable words from each topic, is what users tend to think of as the “topic” itself). We also used a document–topic matrix, which showed the distribution of each of the topics across each document. When run on our corpus, these topic modeling outputs allowed us to get a sense of what authors talk about when they write about the future of the library.

**Corpus and metadata gathering methods**

To build the corpus, a variety of sources were used, including Google, general academic and news databases, as well as databases such as LISA and Library, Information Science & Technology Abstracts. The authors deliberately sought out commentary on the future
Blog posts, newspaper articles and opinion pieces and scholarly articles were all considered to be in scope for the topic modeling project, as we were interested in tracking both academic and broad public conversations. Search methods in databases ranged from simply searching for articles on future and libraries to variations of the idea of death and/or doom combined with libraries, with variant keywords like decline, threat or crisis to get a variety of commentary. Articles selected for the corpus needed to be concerned with the future or death of libraries in general, so news articles about one particular library’s closure because of funding cuts, for example, were not included in the final set of articles analyzed for this project. Large-scale strategic exercises, like MIT’s Future of Libraries report, or exercises in prediction from the Association of College and Research Libraries (ACRL) using a library as a case study, were included. Ultimately, over 150 documents on the future or death of the library were selected and added to a collaborative box folder as plain text (.txt) files. Topic modeling was used to generate lists of the top five articles for each topic, and the researchers then engaged in close reading for those top articles to learn more about the topics.

The corpus was developed over a period of several months by the librarian members of the research team. For each document, 16 metadata elements were captured including standard elements such as author, title and date and less obvious fields such as whether the article primarily addressed the future of the library, the death of the library or both. Information about the authors was collected primarily from the biographical statements accompanying the articles in the corpus, or through publicly available information on author websites. Geographical regions for the authors were also collected, with close to 60 percent of the authors in the corpus residing in the USA; 20 percent from a variety of countries including South Africa, Canada and England; and 15 percent were of unknown national origin. For authors within the USA, male authors were represented at close to 60 percent, with 40 percent of the articles written by female authors. This is interesting, given the demographics of the library profession in the USA, which in 2011 was close to 83 percent women (AFL-CIO Department for Professional Employees, 2011). Institutional affiliation of the authors was overwhelmingly academic, with additional representation from news publications, corporations, blogs and nonprofit organizations. Authors with a connection to libraries made up over half of the corpus, with librarians (40.6 percent), professors of library and information science (10.3 percent) and library directors (2.1 percent), all represented. Professors in fields other than library and information science represented around 6 percent of the corpus, while reporters, writers and editors combined for close to 10 percent of the articles gathered. Although the researchers attempted to search broadly for both optimistic and pessimistic takes on the future of the library, close to 70 percent of the articles in the corpus were focused on the future of libraries, with over 20 percent concerned with the death of libraries in the future and 8 per cent of articles exploring both themes at once. Years of publication for the articles in the corpus ranged from 1935 to 2017, with an increased number of articles dealing with the future of libraries found from 2010 onward.
Topic modeling methods

To examine this corpus, we ran several instances of LDA-style topic modeling in R and decided to focus our analysis on a 15-topic run of the model. We were satisfied that at 15 topics, the topics themselves appeared both meaningful and discrete. In the following analysis, we will present the topics in two ways. First, we visualize the list of words associated with topics in word clouds with the weight of each word’s association with the topic corresponding to the size of the word. This provides a way to quickly demonstrate how evenly distributed a topic is among its top words – how much a topic is driven by particular word co-occurrence. Second, we have included some timelines of the average yearly “presence” of a topic in our corpus since 1980. In topic modeling, each document is conceptualized as consisting of each topic in different proportions (or alternatively probabilities), so that in a document–topic matrix, each document would have different percentages under each topic that all sum to 1 (100 percent of the document is divided up among different topics). To generate the timeline of a topic’s average presence, we simply summed the percentage of that topic in each document in each year then divided it by the number of documents in that year[3]. In our discussion, we analyze some of the most interesting trends uncovered, but we have also included Appendix 2 with the timelines for all of the 15 topics. However, it is important to remember when interpreting our timeline figures that our corpus is not even (Figure 1).

While we have 15 documents in the corpus from 2015, for example, we only have one document in each year from 1980 to 1984. So, seeing a spike in an early year means that one document registered the presence of a topic, but spikes in later years generally mean that a topic was present across a number of documents. While we tried to be comprehensive in gathering our corpus, and certainly did not exclude any articles that we found that were from the earlier years of our study, this type of corpus unevenness can be a common problem in assembling contemporary corpuses.
Figure 1: Articles per year in corpus

Despite the challenges of corpus unevenness, the topics produced by our model gave us meaningful results. We were encouraged to see some results that we expected and some results that granted us new insight into this body of literature. While part of the promise of topic modeling, as a method, is to spark new insights about a body of work, another part of how researchers know a model is working is that it produces results that make sense given their own expertise in a subject. Our model did both, producing topics that could be roughly sorted into three broad, expected, library categories: collections, services and technology. Below, we will “read” several of the topics that belong to these three categories. In analyzing the topics, we will examine the corpus on different scales, drawing upon the two forms of visualization that present information from multiple documents at once and our own reading of individual documents that our model indicated “contained” the highest proportion of each topic.

Collections

Topics dealing with the idea of collections approached both digital and physical collections, as well as the shifting nature of work inherent in the field of cataloging, where practitioners describe library collections of all kinds (Figure 2).
In both close and distant reading on the topic “books, google, book, copyright, paper”, a common assertion was that libraries must respond to a changing information landscape in collection development. Advances in technology challenge library collection managers to stay relevant and advance the goals of libraries (e.g. providing access to information, lifelong learning and enabling democracy), which remain unaltered and perhaps even more critical in the digital age. Articles associated with this topic focused on technological infrastructure and understanding policy as a role for modern librarians (Darnton, 2009; Darnton, 2012). Futurists writing on this topic forecasted the death of the book (Bruccoli, 2007) and encouraged librarians to shift their professional focus from libraries qua buildings and collections to librarian skills, knowledge and activities unmoored from the materiality of libraries (Plutchak, 2012). In contrast to commonly cited threats to the book in recent literature, one historical article from 1935 surfaced in our top five close readings that listed excessive dryness, excessive dampness, sunshine, gas, artificial light, sulfur dioxide and actinic light as the “enemy of the book.” (Liu, 2017; Gilchrist, 1935). Articles in this topic generally read as a call to action for librarians to embrace change, focus on current needs and stay informed and engaged in technical, political, legal, commercial and cultural information change (Figure 3).

A distant reading of this topic showed a noticeable upward trend in articles related to “books, google, book, copyright, paper” from 2002-2015, with a large spike around 2008. This result comports with our assumption that a topic including Google and copyright would see an upward trend over the past two decades. Google was founded in 1998 and firmly entrenched in our culture by the time of its initial public offering in 2004. The migration of mass amounts of content online during this same period resulted in increased concern over copyright violations in the library community, as well as a fear of being marginalized by online content and automated reference. One contribution to these spikes includes scrutiny of the Google Books Project in library literature, which started in 2002 when Google partnered with various research libraries to digitize mass amounts of copyrighted content. Our graph shows a small spike in 2005 when the Author’s Guild sued Google Books followed by a much larger spike in 2008, which corresponds with Google’s settlement with publishers and the formation of the HathiTrust.

Figure 2: Word clouds generated with the top 50 words in each of the collections topics, word size denotes ration of the words in the topic
The topic “collections, digital, research, resources, materials” highlighted dramatic shifts in collection development, curation of materials and preservation of both analog and digital materials. Rising journal costs, declining circulation and shrinking library budgets have led to a deemphasis in monograph collection. Patron driven and “just in time” acquisition is challenging the traditional “just in case” model of collection building. Research universities are investing in their special collections with a local focus to prove their value, as “commodity” materials become more broadly available online or by request (Anderson, 2015). With the advent of digital collections, questions of preservation and permanence have risen as librarians ponder the concept of “forever” when stewarding and collecting digital materials (Teper, 2005; Evans, 2015). Earlier articles tend to be more hyperbolic about the change (Lewis, 2007) perhaps because calls
to action had to be more dramatic to facilitate change, while more recent instructive principles on collection management (Dempsey et al., 2014) tend to be more moderately toned and matter of fact. The 10-year span represented in the top articles on this topic all share the common thread of an ongoing response to technological change, decline of print collections and reference interactions, shrinking budgets, preservation challenges and discussion of the role and relevance of the library (Figure 4).

Figure 4: Timeline of average proportion of topic “catalog, cataloging, research, records, data” across documents in corpus within each year

A section of the corpus dealt with cataloging and technical services departments is expressed in the topic “catalog cataloging research records data”. While the entire field is affected by changes in technology, library automation went through a revolution in the 1980s and 1990s, with the introduction of the online public access catalogs. This period was also reflected in the proportion of the topic over time, which displayed greater activity in the early- to mid-1980s, with more scattered publications shown in the 2000s, as shown in Figure 4. The staffing changes required by this shift in cataloging and development of shared and consortial environments to manage the work of creating catalog records had a large impact on technical services departments. There were notable increases of articles for these topics in the early- and mid-1980s, which confirmed author expectations that changes in library automation would result in more concerns about the changing nature of cataloging.
Services

Three topics in the corpus seemed to be concerned primarily with academic library services. Broadly, two of these topics were engaged in trend forecasting and responses, while one topic dealt more critically with social justice and representation in the profession (Figure 5).

Figure 5: Word clouds generated with the top 50 words in each of the services topics, word size denotes ratio of the words in the topic.

Articles in the services category of clustered topics were focused on thinking about the future based on broad observable trends and relating them to services and plans libraries should be making. Envisioning the life of a researcher in the future and developing services to meet those needs was a strategy recommended for librarians in the “research, work, students, faculty” topic. Strategies outlined by these articles included future proofing, trends to be aware of and forecasting exercises. The idea of engaging in strategic planning to be more nimble in responding to future developments was centered in library leadership, with examples found in the MIT Future of the Libraries report (Ad Hoc Task Force on the Future of Libraries, 2016) and Cornell University’s Day in the Life of a Library project (Tancheva et al., 2016), as seen in the research, work, students and faculty topic. The idea of librarians serving as trend spotters and responding to those trends to stay relevant was also evoked in multiple articles. The importance of community and the library as a third place was a direction to explore to ensure that the library remains relevant in a world driven by changing technology.

The idea that academic libraries have weathered and continue to weather a significant amount of challenges but have adapted to a changing information environment by identifying new services and value was a primary theme in the “academic, students, reference, data” topic. Tenopir et al. (2015) discuss the emerging service of research data management, particularly in research libraries. Brindley (2009) writes a prescient and optimistic article about the way forward for academic libraries, focusing on areas of opportunity such as data services, Web 2.0, digitizing special collections, information
literacy instruction, digital preservation and long-term access and library as place. In 2016, Donald Barclay took a data-driven look at the state and use of public and academic libraries. While print materials and reference statistics are on the decline, the “gate count” of the 60 largest academic libraries has increased 39 per cent from 200 to 2012 (Barclay, 2016). The selection of articles in this topic highlighted the resilience and reimagining of libraries and how our major concerns evolve over time.

We read the topic “social future American white association diversity” as being broadly about social justice issues, which surfaced articles that dealt with the idea of the future or death of the library more critically, instead of just making predictions. As an idea of the future produced by American Library Association, Mirza and Seale (2017) point out that a major problem with the ALA Trend Library is that “Technocratic visions of the future of libraries aspire to a world outside of politics and ideology, to the unmarked space of white masculinity, but such visions are embedded in multiple layers and axes of privilege”. Morales et al. (2014) look at differences in demographics of the college students academic libraries serve and contrast that with the majority white and female characteristics of librarians, stating “We must ask ourselves what message such a stark lack of diversity communicates to patrons, to current and potential members of the profession, and to society at large.”

Technology

There were three topics that dealt with the future of the library in light of new technologies (Figure 6).

Figure 6: Word clouds generated with the top 50 words in each of the technology topics, word size denotes ratio of the words in the topic

Broadly, these topics were about information technology shifts, Web 2.0’s influence on libraries and eBooks. Surprisingly, articles dealing with information technology and the Web generally did not decry either as the end of the library. The “information technology” topic was broadly concerned with assessing new information landscapes and delivery systems and detailing how the library could incorporate or adapt to them. The topic was not centered around a single shift, but rather dealt with wide-ranging changes to how we access and explore information. Articles included information technologies like augmented reality (Zak, 2014), information retrieval given the rise of free-text
searching (Liu, 2017), information explosion, access expansion (Schwarzwalder, 2000), new reference databases like Google Scholar (Noh, 2015) and even a summary of predictions of library changes because of the advent of the computer itself (Coyle, 2007). In each case, a shift in information access or retrieval prompted authors to reflect actively on the best and most useful library response.

Articles that were associated with the “web, users, services” topic were also generally positive about the library’s future. These articles tended to be about directions libraries should take given changes in internet technology and use (i.e. Web 2.0 and 3.0). Given the new focus on interaction, user-generated content and virtual communities, these authors ask themselves what library 2.0 (and beyond) might look like. Rather than focus on a series of shifts in information technology, as the previous topic, “web, users, services” focused more generally on the ethos of user-centered, user-generated experience and content. One might think that the easy accessibility of reference information on the Web would create more anxiety within library commentary, but the articles in this topic framed the Web as an inspiration rather than a threat to libraries. In addition, several of these articles referred to library patrons universally as “users”, indicating a specific relationship to those who take advantage of library services that is perhaps aligned with that of tech companies.

The documents that are associated with the “books, book, reading” topic were decidedly less optimistic in tone. At first glance at the list of words that make up this topic, it is not evident that this topic is about eBooks and ereading. The biggest word in the topic is simply “book.” However, all of the top five articles associated with this topic have to do with eBooks. When articles highly associated with this topic spoke about “books,” it was always in the context of eBooks mostly as a comparison between eBook and paper book purchase in Pew data summaries. But sometimes these articles also referred to “electronic books” which was separated into two tokens in corpus preparation, demonstrating the limits of the bag-of-words approach to text mining. In this case, the list of top-words-in-topic is not fully indicative of what theme the documents in the topic really have in common.

Ereaders and elibraries are much more of a contentious topic than information technology and the World Wide Web. The articles associated with this topic ran the gamut from neutral summaries of Pew data about eBook purchases (Rainie et al., 2012; Horrigan, 2016) to e-readers as harbingers of the End of Days for libraries (Anthony, 2011) to early incorporation of e-reading content as one of the indications that libraries will not go extinct at all (Sorcinelli, 2016). One article even goes further, urging libraries to more quickly and exhaustively transition to eBooks and electronic collections (Lewis, 2007). While ereading is on the rise according to Pew research data, Sorcinelli declares this is no threat to libraries. Anthony, on the other hand, in 2011, predicts that “Amazon Digital Library would kill the public library system.” And there does seem to be a spike in this “ebook” topic in 2010 and 2011, right after the release of the Kindle Fire (Figure 7).

Gaiman’s (2013) article, which is also associated with this topic, though not in the top ten most highly associated documents, elegantly argues the opposite of the doom-and-gloom articles of the early 2010s – that paper books are here to stay, after all they are
“tough, hard to destroy, bath-resistant, solar-operated, feel good in your hand”. Articles in this topic also pointed out that possible death by ereader is not merely an issue for nostalgic librarians, but fraught with social justice issues. Several of the pieces discussed the problem of access: not everyone can afford an e-reader, so if e-readers kill the library it means abandoning the people who actually need free access to books. The great public commitment to free and available information and knowledge would be over, and the poor and minorities will primarily be the ones paying the price (Mott, 2013; Sorcinelli, 2016).

Figure 7: The average proportion of “ebooks” topic presence in each year across the corpus

Process reflections

The process of developing a text corpus was illuminating for the librarians who had not engaged in a topic modeling project before. Even though we budgeted several weeks for the process of finding articles and recording article metadata, we ran into issues with quality control when working on article citations in a shared spreadsheet. Often, when consulting with non-library faculty about digital projects, we emphasize the importance of consistent metadata, so it was humbling to confront our own metadata inconsistencies and deal with clean-up as we noticed issues with our own project. Consistency in file naming and sharing was important, as we were building the corpus. The varying periods for the articles we were adding to the corpus also created problems with OCR and pdfs, as we were converting all articles in the corpus to text files. At times, we were able to add articles to the corpus easily when they had text fully embedded in the PDFs, but articles
that were scanned as image only required another round of OCR before adding them to the corpus for topic modeling. This required an additional workflow for text conversion. The context of this project provided a unique opportunity to reflect on the collaborative process between library faculty and researchers from other colleges and departments. As previously noted in our introduction, four colleges at the University of Utah – the College of Humanities, College of Fine Arts, J. Willard Marriott Library and the College of Architecture and Planning – have partnered to form an interdisciplinary space, the Digital Matters Lab, to support and enable digital scholarship. As is common at other institutions, librarians at the Marriott Library inhabit the dual role of both supporting digital humanities projects and conducting our own original research. Yet, as Trevor Muñoz and others have noted, “there is an ‘othering’ of research as a domain belonging to ‘the faculty’ (regardless of the fact that librarians at many institutions hold some kind of faculty status)” (Schoch, 2017; Muñoz, 2016). Muñoz notes that even the ACRL statement on digital humanities relegates librarians into a support position to research that primarily occurs outside of the library:

    Academic libraries can play a key role in supporting humanities faculty in their research by creating partnerships and collaborations and helping to connect with other campus units needed to implement and carry out digital humanities research (Muñoz, 2016).

This project was a strategic opportunity to model librarian–faculty collaboration where the research partners came in on equal footing and each brought their respective strengths to the project. The model explored by this project effectively inverted the typical dynamic where librarians are often recruited as support for a project. In this case, the librarians formed the initial idea, designed the project and identified additional skill sets that they needed to draw on among collaborators. In this project, the librarians took a leadership role, but having a third collaborator with a strong digital humanities background helped us develop a workflow, select metadata fields and repurpose existing R code to run topic models. Working closely with a digital humanities post-doctoral fellow helped us understand how non-library faculty approach their research and how to pave the way for future librarian–faculty collaboration.

**Assessment of topic modeling as a digital humanities technique**

After reading a few too many Future or Death of the Library articles in recent years, we were ready to take a step back from individual articles and engage the text corpus with a fresh approach. After running several topic models and exploring one run in detail, we conclude that topic modeling is a powerful tool for exploring a large, unstructured corpus of documents. Its division of the documents into topics could be useful for directing researchers directly to only those documents that interest them. For example, if a scholar was interested in social justice and libraries, the “social, future, American, white, justice, diversity” topic can help her/him pinpoint where to read more deeply by focusing on the top five documents associated with this topic.

For this article, while we explored particular topics that interested us, we also wanted to generate a general description of what this corpus covers. For this purpose, the topic
model also worked well, if in an understated way. While we were (perhaps naively) hoping that the topic model might reveal to us deep but hidden truths about our corpus, surprising us with flashes of insight, the model runs mainly confirmed what we expected to see in this collection of documents. Our anecdotal experience with library discourse is that it is both persistent and circular – new technologies arise, followed by predictable responses of fear and opportunity. That the corpus did not reveal a perceptible sea-change in the discourse over time comported with our personal experience. There was no detectable surge in pessimism among these articles with the economic downturn of 2008, for example. Both the death and the flourishing future of the library are written about in every year, with no major consensus building.

Conclusion

As a collaborative exercise between a digital humanities researcher and two librarians, this project provided an effective way for librarians to learn more about doing digital humanities. The librarians will be more effective collaborators with digital humanists in the future as a result of this experience. As librarians who support digital scholarship projects, the two librarians involved in this project feel better positioned to advise on corpus creation, metadata gathering and topic modeling workflows. The digital humanities researcher was able to explore topic modeling more critically as a technique by applying it to a corpus that was previously unfamiliar. Topic modeling can be useful as a method for reviewing a body of literature, alongside or in place of other traditional methods of analysis such as a traditional lit review or annotated bibliography. While we might have naturally arrived at the clusters of topics if we had read all the articles in the corpus, seeing a list of words associated with each topic along with the top articles for each topic allowed us to be more directed and efficient in our close reading.

Analyzing the corpus of library literature through topic modeling produced a few surprises. Some of the researchers expected the corpus to be much more pessimistic about the future of libraries, perhaps because doom and gloom articles loom larger in the memory. Topics like the one on whiteness and social justice and the future of libraries caused us to read and appreciate the library literature in a new light, by indicating an area where there is more nuanced critique of the idea of the future of libraries. Indeed, the subject of social justice and equity in the context of this project allowed us to reflect on the inherent privilege of those forecasting the death of the library as if Internet access, computers, and mass quantities of copyrighted content could be easily attained by anyone at any time. To say “we no longer need libraries” ignores the reality that 25 per cent of Americans do not have internet at home, and for many, the library is how they bridge that digital divide (Pew Research Center, 2018).

We were also surprised at times at how our topic organized in unexpected ways. Under “services” we expected to see clusters of services such as “reference, collection development, data, academic”. What we found, however, was that several topics focused on, not what library services were, but on who is being served (e.g. students and the homeless). Overall though, the combination of topic modeling and close reading of the library literature confirmed our general expectations when initially embarking on the
project. When people talk about the future of the library, they talk about new technologies, they talk about services and they talk about collections. Academic libraries talk about students and faculty. Changes in library automation cause librarians in technical services and cataloging to reflect on the future.

We have found that topic modeling is less of a static result and more of a process that leads to increased knowledge. Rather than consisting of the final word-clouds and graphs, modeling is a coming-to-know through toggling among methods, scales, and disciplinary perspectives. What we learned about this corpus, we learned from viewing it in different ways and with different granularities. We have left this experience with both a better sense of how digital humanities collaborations can function and a great confidence in the future of libraries. Seeing how deliberately the literature in the field examines the conditions and future of libraries has caused us to be more optimistic about the future. On the whole, we were encouraged by how self-reflective this corpus showed libraries to be. Library professionals are actively grappling with technology changes, equitability, access, relevance and how best to engage with the people they serve. These characteristics testify to libraries’ resilience in times of great change. While we feel more confident than ever that libraries as an institution will persist, we are equally sure that the conversation on what form the library will take will continue with various levels of optimism and fatalism.

Notes
1. We did run a quick model on AP articles, and these examples are two actual topics that came out.
2. This example is inspired by Ted Underwood’s own example of the word “lead” in his explanation of topic modeling, which is a great place to start if one is interested in topic modeling (Underwood 2012).
3. There are multiple ways to average the presence of a topic across documents. For our purposes, a simple average worked best, although one could do a weighted average as well. But we decided to give each article the same weight because our articles do not vary tremendously in length, and we believed that each article has similar impact. In our averages a short article on technology in the library is equal to a long article on technology in the library. We are looking at the number of articles in a year that are “about” technology not the number of words in a year.

Acknowledgements

The authors wish to thank Dr William R.L. Anderegg for statistics support.
References


Appendix 1

Topics produced:

- books google book copyright paper
- Topic 2: academic students reference data
- Topic 3: catalog cataloging research records data
- Topic 4: collections digital research resources materials
- Topic 5: books, people, reading, read, public
- Topic 7: public, community, learning, services, communities
- Topic 8: information technology user search digital reality
- Topic 9: people, thinking, space, students, work, play
- Topic 10: research, work, students, faculty, mit
- Topic 11: access books make time public people
- Topic 12: web users services google information accessed
- Topic 13: information future services change academic research
- Topic 14: school games video students learning
- Topic 15: social future American white association diversity
Appendix 2: Presence of topics in text corpus, plotted across time
Appendix 3: Articles Included in text corpus with first author listed


Barclay, Donald A. "Has the Library Outlived its Usefulness in the Age of the Internet? You’d be Surprised." The Conversation. 2016.


Denning, Steve. "Do We Need Libraries?." Forbes. 2015.


Foote, Carolyn. "Think Like a Futurist; When we envision a future for our libraries, we shape the present." School Library Journal. 2014.


Graboyes, Robert. "Column: Do we need librarians now that we have the internet?" PBS Newshour. 2016.


Hirsh, Kenneth J. "Like Mark Twain: the death of academic law libraries is an exaggeration." Law Library Journal. 2014.


Scheinfeldt, Tom. "Nobody cares about the library: How digital technology makes the library invisible (and visible) to scholars." Found History. 2012.


Shatzkin, Mike. "It will be hard to find a public library 15 years from now." The Idea Logical Company. 2011.


Staff Editorial. "Sound Off; In this digital age, do libraries still matter?." Florida-Times Union. 2015.


