Qualitative and Quantitative Methods in Libraries (QQML) 7: 23 –30, 2018

An Introduction to Text Mining: How Libraries Can Support Digital Scholars

Niamh Wallace[1] and Mary Feeney[2]
[1]University of Arizona Libraries
[2]University of Arizona Libraries

Abstract: Scholars in the digital humanities and other disciplines utilize a variety of qualitative and quantitative methods in their research. One of these methodologies – textual analysis – can be facilitated by the use of text mining digital tools, such as Voyant, ATLAS.ti, JSTOR Data for Research, and Google's Ngram Viewer. Supporting digital scholarship is an important and expanding role for academic libraries, from providing resources for text mining to collaborating with digital scholars. This paper gives a brief overview of text mining, explores some of the available tools, and discusses limitations and challenges. It also shares examples of text mining projects and examines how libraries can offer expertise and services to support this research, including the provision of corpora for text analysis and the integration of digital tools in courses.

Keywords: Text Mining, Text Analysis, Digital Scholarship, Digital Humanities, Academic Libraries

1. Introduction

Text mining is a method for the "computational analysis of text" (Underwood 2015). King (2015) explains that it "leverages computational methods for the analysis of large digital collections of texts and images. It is an umbrella term for an array of tools that enables us to go beyond the capabilities of keyword or full-text searches." Text mining is used by researchers to count word frequencies, find patterns of word usage, identify n-grams (sequences of co-occurring words), examine large-scale patterns and trends, and explore topics. Analysis of corpora (sets of texts) through text mining may offer new insight about the texts, and data generated from text mining may be used for additional textual analysis. This paper provides an introduction to text mining for librarians to facilitate their support of or partnership with scholars engaged in text mining research.

Received: 14.3.2017 / Accepted: 19.12.2017 ISSN 2241-1925

2. Literature Review

Text mining is not a new method, but has broadened beyond its origin in computer science, where it is generally understood as a "subfield [of data mining] devoted to the extraction of knowledge from unstructured text" (Jockers and Underwood 2015). Within the digital humanities, text mining is "an interdisciplinary endeavor that also borrows freely from corpus linguistics and computational linguistics, as well as social-scientific traditions like social network analysis" (Jockers and Underwood 2015).

The emergence of text mining as a scholarly practice that crosses disciplinary boundaries has not escaped the notice of academic libraries, who play key roles in supporting faculty research. As Reilly (2012) notes, "the growing application of text mining techniques and technologies in many fields of research has implications that are beginning to be felt by libraries." Under the larger umbrella of digital scholarship, academic libraries are particularly well positioned to support text mining, thanks to their crucial role in licensing electronic content – the "high-quality" data sets (Orcutt 2015) needed by researchers engaged in this type of analysis. As such, much of the literature on text mining support in academic libraries centers on securing access to library-licensed content for mining purposes (Cheney 2013; Orcutt 2015; Reilly 2012; Stewart, Secker, Morrison, and Horton 2016; Williams, Fox, Roeder, and Hunter 2014). Negotiating licensing agreements, understanding copyright limitations, obtaining usable data, and working with publishers at scale are key elements of "building the right kinds of conditions and terms for computer-assisted processing and analysis of commercial database content" (Reilly 2012).

In addition to procuring access to data sets for researchers, academic libraries can support text mining in other ways. For scholars interested in the application of text mining, but lacking in technical expertise, "librarians have opportunities to lower the technological barriers to text mining by offering functional skill sets, developing easy-to-use tools, and offering occasions for education, training, and skill development" (Anderson and Craiglow 2017). This may require facilitating partnerships with specialists in the library or at the institution.

The "computationally intensive" (Anderson and Craiglow 2017) work of text mining can be a challenge for librarians as well as novice researchers. Technical capacity for text mining within the library is not always an option, or if it is, may not be at a scalable level, but there are additional ways to collaborate with and assist digital scholars beyond offering direct technical support. Liaison librarians can create guides to text mining, listing available corpora and tools for analysis, and be prepared to offer baseline consultation and referral services to faculty and students. As Okerson (2013) notes, "there is an important role for the data-mining-savvy non-technologist who can sit with the end user and interview him or her; then propose a strategy; then work with that user on a pilot mining exercise."

Kamada (2010) also observed that text mining research involves some technical knowledge, "but also often skills and knowledge in collecting and organizing data, in which librarians have unique training and background." Libraries can combine this knowledge with awareness of available texts and tools to provide consultation services to scholars initiating text mining research projects.

The level of support for text mining will vary across research libraries, based on available skillsets or capacity, but as Orcutt (2015) points out, "the crucial question for all institutions should not be whether to support [text mining] activities, but what that library support should look like."

3. Libraries and Text Mining

3.1. Corpora and Tools

As described in the literature review, libraries can facilitate text mining research in several ways, from providing access to the corpora that researchers at their institution need, to consulting with scholars on their research. There are several freely-available sets of texts to which libraries can point scholars on a guide or other listings on the library website. Librarians working with disciplines that tend to do text mining can also promote these data sets to researchers to make them aware of the possibilities. A few examples include JSTOR Data for Research (http://dfr.jstor.org/), which enables mining of the content in the JSTOR archive whether or not their library subscribes to the JSTOR database; *Chronicling America,* the open access database of U.S. newspapers digitized through the National Digital Newspaper Program (NDNP) of the Library of Congress and National Endowment for the Humanities (NEH); and the HathiTrust Digital Library.

There are also freely-available smaller data sets created by institutions or individual researchers. Some additional examples are available at the Data Collections and Datasets website, curated by Alan Liu (http://dhresourcesforprojectbuilding.pbworks.com/w/page/69244469/Data%20Collections%20and%20Datasets). There are links to "demo corpora," including inaugural speeches of U.S. presidents, "Feeding America: The Historic American Cookbook Dataset," several literature sets such as fiction collections collected and compiled from Project Gutenberg, and several more.

There are also sets of text that require purchase and/or licensing by libraries. As stated earlier, providing content such as journals, books, and databases for use by researchers is a central role for academic libraries, thus expanding that role to the acquisition of data sets and text corpora is logical extension.

In addition to providing texts for data mining, libraries can provide consultation and expertise in tools needed for this method of research. There are free, user-friendly tools available online, such as Voyant (http://voyant-tools.org/), an online tool for text analysis. Researchers can paste text into the Voyant window,

upload their own text files, or use one of the pre-loaded text sets, currently Shakespeare's plays and Jane Austen's novels. Voyant includes several tools for term frequencies, visualization, and more.

The Google Books Ngram Viewer (https://books.google.com/ngrams) is a free tool for viewing trends in the occurrence of words and phrases in the Google Books corpus. Since Voyant and Google Books Ngram Viewer are both freely-available web-based tools, scholars can access and use them on their own. Libraries' roles can be to understand how these tools work, what their limitations are, and in the case of tools like the Ngram Viewer, what the scope of the corpus is. Libraries can also make scholars, especially those new to text mining, aware of the availability of these tools.

Qualitative data analysis programs that are widely used in the social sciences, such as ATLAS.ti or NVivo, are additional options for libraries to support text mining and content analysis. ATLAS.ti, for example, is a tool for "qualitative analysis of large bodies of textual, graphical, audio and video data" (ATLAS.ti 2017). These programs are not free, but provision of these types of tools would be a natural extension for some libraries that already provide access to quantitative statistical analysis programs like SPSS in their computer labs. In addition, these robust tools have the potential to be used by scholars in a wide range of disciplines and broadly in digital scholarship. As one example, the University of Illinois at Urbana-Champaign lists information about both ATLAS.ti and NVivo on its Text Mining Tools guide (http://guides.library.illinois.edu/c.php?g=405110&p=2757865) and provides access to ATLAS.ti in their Scholarly Commons. For more advanced support, libraries could also provide expertise and consultation with using R, an open-source programming language, which can also be used for text mining.

3.2. Libraries as Partners in Text Mining

In addition to serving as a resource for scholars to support text mining research, libraries can be active partners in that research. There are many examples of text mining being used in scholarly research, and the few examples here focus on projects in which libraries initiated text mining projects or partnered with scholars.

One project is "Lincoln Logarithms: Finding Meaning in Sermons." Emory Libraries applied several digital humanities tools to a corpus of sermons delivered upon the death of U.S. President Abraham Lincoln. The four tools were used to mine the texts to view word frequencies, create topic models, create timelines and maps, and create visualizations (http://disc.library.emory.edu/lincoln/). One of the tools used in this project was Voyant (http://voyant-tools.org/), a free online tool for text analysis, which the researchers used to look at the frequency and distribution of the terms 'slavery' and 'peace' in two of the sermons. The texts used in this project had been digitized by the Beck Center for Electronic Collections at Emory Libraries.

"Robots Reading Vogue" is a project of the Yale University Library Digital Humanities Lab (http://dh.library.yale.edu/projects/vogue/). *Vogue* is a popular magazine published for over a hundred years, and its archive, consisting of hundreds of thousands of pages, has been digitized by ProQuest. One of the tools used to mine this archive included an n-gram search tool to analyse the use of words and phrases over time.

A project of the Digital Scholarship Lab at the University of Richmond utilized the digital archive of the *Daily Dispatch* for text mining and topic modeling. The newspaper had been digitized as part of an Institute of Museum and Library Services award, a joint project of the University of Richmond Libraries, Tufts University, and the Virginia Center for Digital History (http://dlxs.richmond.edu/d/ddr/index.html). The "Mining the *Dispatch*" project created topic models; topic modeling is a method that "uses statistical techniques to categorize individual texts and…to discover categories, topics, and patterns that we might not be aware of in those texts" (Nelson).

Researchers at Virginia Tech text mined newspapers in *Chronicling America* for their project "An Epidemiology of Information: Data Mining the 1918 Influenza Pandemic" (Hausman 2014). The project, which used multiple methods and was funded through the NEH Digging into Data Challenge, included a Social Sciences librarian and faculty members in History, English, Computer Science, and others, (http://www.flu1918.lib.vt.edu/?page_id=6).

4. Text Mining at the University of Arizona Libraries

Like many other libraries, the University of Arizona (UA) Libraries has been developing services for teaching and research in digital scholarship. One of the approaches is building awareness of what the library can provide. The authors presented twice about text mining at the UA Libraries' "Tech Talks," a series of presentations held at the library for faculty and students to share and learn about technology. One of these sessions was an overview of text mining that introduced concepts, tools, and examples of this method. The other presentation focused on the JSTOR Data for Research (DfR) tool that enables text mining of the journals and other content in the JSTOR digital archive.

Another goal is connecting with and integrating digital scholarship methods like text mining into the curriculum. One of the authors partnered with a faculty member in History to introduce digital humanities methods and concepts in a graduate student seminar course. The students were introduced to Voyant, among other tools, with a demonstration of visualization of word occurrences and a graph of word trends within newspaper articles on a given topic. The author also created a Digital History guide with readings, tools, and examples of digital humanities in history, including links to text mining tools (http://libguides.library.arizona.edu/digital-history).

The other author collaborated with faculty members in English on a grant-funded course exploring Shakespeare, diversity, and technology. With library colleagues, the author designed and taught four course sessions as part of an introduction to the digital humanities, with classes on text mining, digital mapping, and virtual reality content creation. At the conclusion of the course, students presented their final digital projects at a state-wide conference. Funding for the conference, as well as for software and equipment used in the course, was derived from a NEH grant in addition to a UA institutional grant supporting student engagement efforts. For English majors largely unfamiliar with applications of technology to literary analysis, text mining was one of most accessible methodologies, thanks to the ease of use of mining tools such as Voyant, and the availability of Shakespeare's corpus. The course represented an opportunity for the library to partner with faculty who had little or no experience in the digital humanities to help prepare students for future academic careers that will increasingly incorporate digital technologies.

The authors are continuing to identify opportunities to connect text mining and other digital scholarship methods and tools with the curriculum. For example, the UA Libraries provides ATLAS.ti in its Information Commons classroom in support of teaching and research. One of the authors conducted a workshop on using this program for analysing texts – from newspapers to tweets – for Journalism graduate students in a research methods course.

In addition, the UA Libraries are making available sets of texts for researchers at our institution and beyond. Through a project initiated several years ago, the Historic Mexican and Mexican American Press (MMAP), the library digitized and provides open access to twenty newspapers and magazines, and uncorrected OCR text files are available along with the page images.

Like other libraries, we are beginning to license text files from vendors, such as the corpus of *The New York Times*, to which we already provide database access, but also want to provide the raw text files for researchers to text mine. This has been initiated by the authors in anticipation of supporting scholars, and we have also received requests from faculty for access to these types of data sets.

The authors are also developing a guide to text mining and textual analysis that will include examples of corpora the library provides and those freely-available, information about how to request additional texts, links to text mining tools, and contact information for consultations.

The UA Libraries also recently established a Digital Scholarship Working Group, composed of liaison librarians in the social sciences and humanities, librarians who work as specialists in areas such as data management, and technical experts. The working group serves to define and establish infrastructure, including library services and resources, for supporting digital

scholarship efforts on campus and to share information about developments, both locally and within the broader scholarly realm, in the field.

5. Conclusions

Academic libraries are uniquely well suited to offering support to digital scholars interested in applying text mining to their research. Providing access to corpora and tools for analysis and collaborating with faculty to embed emerging technologies into the curriculum are some of the ways libraries can build out support for these initiatives. As more researchers become engaged with digital scholarship methods, libraries should be prepared to both support them and partner with them in their research.

References

Anderson, C. and Craiglow H. (2017). Text mining in business libraries, *Journal of Business & Finance Librarianship*, Vol. 22, No. 2, 1 – 17.

ATLAS.ti. (2017). What is ATLAS.ti? Accessed April 19, 2017. http://atlasti.com/product/what-is-atlas-ti/.

Cheney, D. (2013). Text mining newspapers and news content: new trends and research methodologies, *Proc. IFLA World Library and Information Congress*. Accessed April 20, 2017. http://library.ifla.org/233/1/153-cheney-en.pdf.

Hausman, B. L., Pencek, B., Ramakrishnan, N., Eysenbach, G., Ewing, E. T., Kerr, K., & Gad, S., (2014). *An Epidemiology of Information: Data Mining the 1918 Influenza Epidemic Project Report*. Accessed April 20, 2017. https://vtechworks.lib.vt.edu/handle/10919/46991.

Jockers, M. and Underwood, T. (2015). Text-mining the humanities. In Schreibman, S., Siemens, R., and Unsworth, J. (Eds.), *A New Companion to Digital Humanities* (pp. 291-306). Wiley, Hoboken, New Jersey.

Kamada, H. (2010). Digital humanities: Roles for libraries? *College & Research Libraries News*, Vol. 71, No. 9, 484 – 485.

King, L. (2015). Data Mining on Vendor-Digitized Collections, *Library Connect*. Accessed April 19, 2017. https://libraryconnect.elsevier.com/articles/data-mining-vendor-digitized-collections.

Nelson, R.K. Introduction. *Mining the* Dispatch. Accessed April 30, 2017. http://dsl.richmond.edu/dispatch/pages/intro.

Okerson, A. (2013). Text & data mining-a librarian overview, *Proc. IFLA World Library and Information Congress*. Accessed April 20, 2017.

Orcutt, D. (2015). Library support for text and data mining, *Online Searcher*, Vol. 39, No. 3, 27 – 30.

Reilly, B. (2012). CRL reports, *The Charleston Advisor*, Vol. 14, No. 2, 75 – 76.

Stewart, N., Secker, J., Morrison, C., and Horton, L. (2016). Liberating data: how libraries and librarians can help researchers with text and data mining, *LSE Impact Blog*. Accessed April 30, 2017. http://blogs.lse.ac.uk/impactofsocialsciences/2016/07/12/how-libraries-and-librarians-can-help-with-text-and-data-mining/

Underwood, T. (2015). Seven Ways Humanists are Using Computers to Understand Text. *The Stone and the Shell*. Accessed April 19, 2017. https://tedunderwood.com/2015/06/04/seven-ways-humanists-are-using-computers-to-understand-text/.

Williams, L., Fox, L., Roeder, C., & Hunter, L. (2014). Negotiating a text mining license for faculty researchers, *Information Technology and Libraries*, Vol.33, No. 3, 5-21