THE SERIALS LIBRARIAN
2017, VOL. 72, NOS. 1–4, 156–159
http://dx.doi.org/10.1080/0361526X.2017.1320876

Routledge
Taylor & Francis Group

Text Mining 101: What You Should Know

Patricia Cleary[a], Kristen Garlock[a], Denise Novak[a], Ethan Pullman[a], and Sanjeet Mann [b]

[a]Presenters; [b]Recorder

ABSTRACT

Scholars increasingly use text and data mining (TDM) methods to discover trends and relationships within complex digital data sets. In order to support this development in scholarly communication, librarians and publishers need to be knowledgeable about TDM methods, build partnerships with TDM researchers, and address challenges related to licensing and access to large document sets. The presenters of this NASIG session shared their experiences of supporting TDM as a library subject liaison, acquisitions librarian, and publisher representatives. Audience members discussed issues involving TDM of data from multiple publishers, local hosting of data sets and TDM activity by undergraduate students.

KEYWORDS

Data analysis; digital humanities; licensing; scholarly communications; text and data mining

Scholars are increasingly turning to text and data mining (TDM) methods to address innovative research questions through the analysis of complex data sets. The International Association of STM Publishers defines TDM as "explorative data analysis by way of automated, computational and linguistic processes and procedures."[1] These methods can be further distinguished by the format of the information being examined: text mining is the automated processing of structured digital texts, while data mining uses different computational tools and methods to analyze non-textual information. The reliance on textual documents makes text mining particularly appealing to scholars working in the digital humanities, but TDM methods can be used by researchers from any discipline and are well suited to studying relationships between concepts or people, or how an object of study has changed over time.

Supporting TDM projects both upholds and extends the traditional roles of publishers and librarians. Kristen Garlock explained that as a resource provider, JSTOR is interested in broadening access to information and supporting the development of new forms of scholarship, while Ethan Pullman noted that libraries have long worked to acquire information and provide access to their users. TDM challenges both communities to develop a deeper understanding of researchers' needs, including knowledge of how scholars use information after they gain access, and to cope with the complexities of collaboration and rights management on a larger scale.

TDM methods are continuously being refined as research projects produce new tools and advance existing analysis techniques. Given that TDM is characterized by rapid change, intensive use of technology, and lack of a single disciplinary "home," it can be difficult for newcomers to grasp the foundational understanding necessary for further learning. Ethan Pullman and Denise Novak of Carnegie Mellon University, Kristen Garlock of Ithaka and Patricia Cleary of Springer US structured their NASIG presentation to provide attendees with just such an introduction to TDM, by offering definitions and examples of TDM projects, sharing their stories of how they have supported TDM as librarians and publishers, and encouraging audience members to contribute their own experiences, questions, and concerns.

Ethan Pullman opened with a discussion of the library subject liaison's role in supporting text mining. Carnegie Mellon supports numerous programs, from undergraduate to doctorate, and maintains research-intensive collections to which it dedicates library liaisons such as Pullman (who works with

CONTACT Sanjeet Mann sanjeet_mann@redlands.edu

the departments of English, Modern Languages, and Philosophy). Pullman stays informed about trends in scholarship by looking over course syllabi and publications mentioned in faculty profiles, attending department meetings and frequenting research showcases and lectures. He encouraged audience members to attend talks across a wide range of disciplines, even if the subject matter is unfamiliar, in order to learn what sources are considered useful and reflect on the strategies needed to acquire and license them. Creating websites or research guides on text mining, partnering with data services and acquisitions librarians and joining online communities and conferences such as Text Analytics World are also effective ways for subject liaisons to learn about TDM projects.[2]

Pullman offered several examples of text mining projects. Tools such as Wordle or Voyant will produce word clouds to display the frequency of word occurrences in a single document or a large collection of texts (called a corpus).[3] A group of Carnegie Mellon researchers led by Christopher Warren created *Six Degrees of Francis Bacon*, a website that uses a concept map to display relationships between people mentioned in early modern texts.[4] Finally, Pullman introduced a student project that used text mining tools to analyze Authors Guild court documents to understand rhetorical strategies used in its argument in the Appeals Court Case *Authors Guild v. Google*. In response to an audience question, Pullman estimated that 80% of the TDM questions that he receives come from student researchers. A few large research projects, usually faculty projects such as *Six Degrees*, involve extensive collaboration with publishers.

Denise Novak recalled that she became involved with text mining when Pullman asked her whether a particular database permitted library users to mine its content. Novak contacted the database's representative, who sent a license agreement in order to obtain the data. Novak now asks for text mining rights to be added to the license agreement for each new textual database, using language derived from the LIBLICENSE model license and working in partnership with university legal counsel. Novak also helped to develop workflows for subject librarians interested in obtaining TDM rights and created a list of sites and publishers that offer complimentary text mining for Carnegie Mellon's research guides.[5] She commented that many publishers charge a "reasonable" fee for access to mine their datasets. The costs vary greatly, sometimes exceeding a thousand dollars and sometimes just covering the cost of the hard drive on which the data are delivered. Carnegie Mellon pays the fees from the library acquisitions budget.

Kristen Garlock explained that JSTOR has supported text analysis since the 1990s and has written text mining permissions into their standard terms and conditions of use for journal and pamphlet collections, as referenced from library licensing agreements. The volume of requests has grown over the years, prompting JSTOR to create standard processes for researchers to ask for data sets and for JSTOR staff to fulfill these requests. The self-service site Data for Research launched in 2008, allowing researchers with small projects to create a user account, define their desired data set, and obtain metadata, keywords, word counts, and engrams.[6] Investigators who need full text or customized data sets can sign an agreement form, or the library may sign on their behalf. Garlock said over 500 of these requests have been fulfilled so far and offered to send interested audience members a bibliography of research projects utilizing JSTOR data sets. Some of these research projects lead to functionality beneficial to the entire JSTOR user community. For example, researchers at the University of Washington ran a citation analysis across JSTOR's entire data set and produced an algorithm for identifying related research, upon which JSTOR is now basing new recommendation functionality.

Garlock observed that JSTOR faces the challenge of scaling up support for its increasingly popular data service. Researchers are posing complex questions about how to use the self-service portal, how to analyze their data, and whether their intended uses and reuses are permissible under license terms. One increasingly frequent request is whether the primary source data set could be made available to other researchers after the end of the project for verification of results, something that is currently difficult for JSTOR to support. Garlock observed that there are no standards to ensure that data sets from different publishers will be compatible with each other, which can confound scholars whose project involves licensing and analyzing multiple sources of data.

Patricia Cleary said that Springer's TDM policy, first introduced in 2014, automatically grants mining rights to subscribed content for non-commercial research purposes.[7] Requests for data outside the institutional subscription or permission to reprint images will also be considered on a case by case basis. Springer limits the amount of text from a data set that researchers can quote in a publication to 200 characters, 20 words, or 1 sentence, and asks that the original content be cited using a Digital Object Identifier (DOI) link. Cleary emphasized that libraries are not required to monitor researcher activities for compliance or reporting purposes as a condition of signing a TDM agreement. She added that Springer allows unrestricted mining of SpringerLink Open Access content governed by the Creative Commons Attribution (CC-BY) license. These license terms are included in all new contracts for academic or government institutions, and customers who do not already have the TDM clause can add it at any time. Cleary commented that Springer is revising these terms to make the process easier for scholars, but expected that no significant policy changes would result.

Individual researchers can download data directly from the SpringerLink platform without entering an Application Programming Interface (API) key, but a metadata API is provided for text mining applications that need it to search through Springer content. The API currently covers journal articles, book chapters, and protocols, with the full text of book archives coming soon. Content is delivered free of Digital Rights Management (DRM) and researchers can choose to either have Springer host the data set for them on SpringerLink or deliver the data to the library for offline hosting on a hard drive.

Having heard individual accounts of how Carnegie Mellon, JSTOR, and Springer support text mining, audience members took a broader view during the question and answer session, refocusing the conversation on unresolved problems related to licensing permissions and the need for greater collaboration between publishers. Audience members remarked that cross-publisher partnerships would make it easier for smaller publishers to support text mining, but that barriers to collaboration arise from fears over loss of control, or potential rights-holder lawsuits if data were to be mishandled. Another audience member reported problems using the CrossRef Metadata API to obtain data sets from multiple publishers: license negotiations dragged on for months due to a conflict with campus licensing rules, and only a small amount of data was delivered. Cleary had described the potential of the CrossRef TDM initiative to allow automated downloads of these cross-publisher data sets, but acknowledged the initiative was still in progress with no estimated completion date.[8]

Librarians asked about TDM licenses that require researchers to use the publisher's API or interface to analyze the data set, and questioned whether batch downloads of textual data might violate other licensing language prohibiting the use of spidering on the publisher's website. Many students are knowledgeable enough to script their own analysis tools and create databases to hold textual data, but other TDM scholars are less knowledgeable and would benefit from a librarian's advice. Pullman opined that researchers should be able to choose to use their own analysis tools whenever appropriate, and explained that there is a difference between analysis tools and "text mining" software. The former does not violate anti-spidering license terms designed to protect publishers' servers, while the latter should be carefully considered, since publishers provide many of these tools as part of their TDM licensing agreements. Locally developed tools should be vetted for compliance with publisher agreements. Garlock added that JSTOR often reaches out to researchers who have downloaded unusually large quantities of articles to find out if they need assistance with data services.

Another question addressed the use of text mining by undergraduates. Novak stated that none of Carnegie Mellon's TDM licenses prohibit this practice, while an audience member reported having to negotiate a prohibition on undergraduate access out of a license agreement. Pullman commented that faculty interests are also at stake, since their recommendations drive many undergraduates to try TDM. These research methods present librarians with a valuable opportunity to teach source evaluation skills by explaining how the presence or absence of sources in the underlying data set will affect data analysis.

In response to a final question, Garlock identified a need for more transparency and consistency in tracking the outcome of text and data requests from publishers. Her comment that publishers are still experimenting with models for supporting large-scale research using TDM holds true for libraries as well. Few librarians or publishers consider themselves intimately familiar with text mining; when Pullman surveyed the audience at the beginning of the session, nearly all rated themselves as text mining novices. TDM techniques have a bright future in the academy, but are still very much in flux. Wider appreciation of TDM's potential to answer innovative questions and understanding of current challenges in supporting TDM is a necessary first step toward lasting solutions. Attendees left this session better prepared to build the partnerships needed to realize this goal.

Notes

1. International Association of STM Publishers, "STM Statement on Text and Data Mining (TDM)," March 15, 2012, http://www.stm-assoc.org/2012_03_15_STM_Summary_Statement_Text_Data_Mining_final.pdf (accessed July 11, 2016).
2. Text Analytics World, "Text Analytics World," http://www.textanalyticsworld.com (accessed July 9, 2016).
3. Jonathan Feinberg, Wordle, http://www.wordle.net (accessed July 12, 2016); Stefan Sinclair and Geoffrey Rockwell, Voyant, http://voyant-tools.org (accessed July 12, 2016).
4. Christopher Warren et al., "Six Degrees of Francis Bacon," http://www.sixdegreesoffrancisbacon.com (accessed July 9, 2016).
5. Carnegie Mellon University Libraries 2015, "Text & Data Mining," http://www.library.cmu.edu/research/tdm/overview (accessed July 11, 2016).
6. JSTOR, "Data for Research," http://about.jstor.org/service/data-for-research (accessed July 11, 2016).
7. Springer US, "Springer's Text- and Data-Mining Policy," http://www.springer.com/gp/rights-permissions/springer-s-text-and-data-mining-policy/29056 (accessed July 11, 2016).
8. CrossRef, "CrossRef Text and Data Mining Services," http://tdmsupport.crossref.org (accessed July 11, 2016).

Notes on Contributors

Patricia Cleary is Global eProduct Development Manager at Springer US.

Kristen Garlock is Associate Director of Education and Outreach—JSTOR at Ithaka.

Denise Novak is Acquisitions Librarian at Carnegie Mellon University, Pittsburgh, Pennsylvania.

Ethan Pullman is Humanities Liaison & Library Instruction Coordinator at Carnegie Mellon University, Pittsburgh, Pennsylvania.

Sanjeet Mann is Arts and Systems Librarian at University of Redlands, Redlands, California.

ORCID

Sanjeet Mann http://orcid.org/0000-0003-4442-1053