Scholarly Needs for Text Analysis Resources: A User Assessment Study for the HathiTrust Research Center

Harriett E. Green
University of Illinois at Urbana-Champaign, green19@illinois.edu

Eleanor F. Dickson
University of Illinois at Urbana-Champaign, dicksone@illinois.edu

Leanne R. Nay
Indiana University Bloomington, lnay@indiana.edu

Ewa Zegler-Poleska
Indiana University Bloomington, ezeglerp@indiana.edu

Harriett E. Green, Eleanor F. Dickson, Leanne R. Nay, and Ewa Zegler-Poleska, "Scholarly Needs for Text Analysis Resources: A User Assessment Study for the HathiTrust Research Center" (2016). *Proceedings of the Charleston Library Conference.*
http://dx.doi.org/10.5703/1288284316464

Scholarly Needs for Text Analysis Resources: A User Assessment Study for the HathiTrust Research Center

Harriett E. Green, English and Digital Humanities Librarian, The University of Illinois at Urbana Champaign

Eleanor F. Dickson, Visiting HathiTrust Research Center Digital Humanities Specialist, The University of Illinois at Urbana Champaign

Leanne R. Nay, Scholarly Technologies Librarian, Indiana University Bloomington

Ewa Zegler-Poleska, IDEASc PhD Fellow, Indiana University Bloomington

Abstract

The HathiTrust Research Center (HTRC) is undertaking a study to better understand the needs of current and potential users of the center's tools and services for computational text analysis. In this paper, we report on the results of the first phase of the study, which consisted of interviews with scholars, administrators, and librarians whose work involves text data mining. Our study reveals that text analysis workflows are specific to the individual research project and are often nonlinear. In spite of, and in some cases because of, the wealth of textual data available, scholars find it most difficult to locate, access, and curate textual data for their research. While the goals of the study directly relate to research and development for the HTRC, our results are useful for other large-scale data providers developing solutions for allowing computational access to their content.

Introduction

Libraries and textual data providers, including digital libraries, digital repositories, and subscription databases, are called to update their service and access models to meet the increasingly data-driven research needs of humanists and social scientists. The HathiTrust is one such textual data provider also developing means by which researchers can perform computational analysis on material in its repository. As of fall 2016, the HathiTrust Digital Library (HTDL) contains over 14 million digitized volumes. The HathiTrust Research Center (HTRC) aims to facilitate large-scale computational text analysis of the contents of the HTDL through data services and analytical tools.

This paper shares preliminary findings of a study we conducted that seeks to better understand current and potential users of the HTRC's needs for text data mining tools in order to understand and anticipate how scholars integrate text analysis into their research. We first describe current practices in computational text analysis as reported by our interviewees. We then focus on two areas of importance to text analysis researchers: data acquisition and use and tools for text analysis.

The study has consisted of a series of interviews with fifteen researchers, librarians, technologists, and administrators whose work involves computational text analysis. Many of our interviewees had interacted with the HathiTrust's Digital Library and Research Center before. Their experience levels ranged from longtime digital humanities practitioners to those just starting out in the field. The interviews were transcribed and then coded and analyzed in Atlas.ti.

Our findings reveal that text analysis workflows are complicated and individualized, and researchers have the most challenges building and curating textual datasets and understanding how text analysis tools work. The results of the study will help us assess the effectiveness of the HTRC, as well as make suggestions for development of future iterations of the HTRC's data services and toolkit. The results of the study will help us assess the effectiveness of the HTRC, as well as make suggestions for development of future iterations of the HTRC's data services and toolkit.

This paper builds on existing research into humanities scholars' use of digital tools (Frischer et al., 2006; Toms and O'Brien, 2008; Gibbs and Owen, 2012;

https://doi.org/10.5703/1288284316464

Green and Courtney, 2015). While the interviews were conducted with specific focus on the HTRC, the results of our study point to opportunities for other large-scale data providers to develop solutions for allowing computational access to their content.

Text Analysis Research Practices and Culture

Motivation for Analysis

Respondents sought to apply text analytic methods to answer research questions in new and exploratory ways. Many of their research questions involved testing previous claims about literary and cultural history using data-driven methods. One historian observed that " . . . when I say people have been studying this time period for 300 years, people who are much smarter than me, better writers, have better access to the archives, who can read more than I can, the only way we can say something new is if we get new perspective on old data."

Types of Methods

We asked respondents about the proportion to which they used quantitative and qualitative methods, as well as mixed methods. These broke down fairly evenly. Other specific approaches were influenced by the nature of their research, such as one respondent's use of "a set of network diagrams, or data that can be played with independently. So sometimes we produce things that are visualizations, digital visualizations, as web pages for local use, sometimes we produce things . . . ."

Research Culture

According to interviewees, text analysis creates opportunities to explore change the scale, scope, and pace of their research. Some respondents noted skepticism they had received from their colleagues, as well as difficulty they perceived in building an academic career in digital humanities. Respondents were variously critical of "neophytism" in digital humanities, expectations to produce innovative results, and the lack of collegiality in the field.

Collaborations

Respondents worked both alone and with research teams. Solo work primarily consisted of data preparation and analysis, searching for materials, finding methods, and learning new skills. Collaborative work was done in teams ranging from two to 20 people and primarily consisted of building databases, assembling corpora, and finding solutions to research questions. Some respondents faced challenges in getting assistance for their work. One respondent noted, "I had tried really hard to find a more experienced linked open data programmer once I got the grant, but had a really awful time, because everyone who I talked with seemed to have noncompete agreements that banned them from working with me even though I'm really not going to be taking over the world or interfering with anyone's DH project."

Publishing

The respondents' publications take a variety of forms, and they disseminated both interim and final-phase research (see Table 1). One interviewee said of publishing text analysis research, "In some ways GitHub is an integral part of this. We can try to describe this code, or you can go look at our code, so it's interesting in that if you read the paper without actually looking at the code, you've gotten sort of a broad overview of the method, but you couldn't replicate it. And if you just tried to read our code, you might not be able to replicate it either . . . So, it's a bit of a hybrid publication."

Table 1. Publishing formats.

Dissertations
White papers
Conference papers
Journal articles
Blogs
Books
Software
Code documentation
Raw data
Derived data

Getting Funding

From the perspective of the interviewees, funding was crucial to conducting text analysis work. Respondents received funding from local sources, such as their department or library, as well as national and international grant-giving agencies. They noted that funding was crucial for collaborative projects, and their institutions lacked a business model to make collaborations work. On the difficulty of obtaining funding, one respondent described how "there are so many good ideas and the ones that are going to have traction, they have to have viability to a funder . . . where I'm stuck right now is just developing enough knowledge that I can put together a viable grant to NSF or Mellon or someone else. It's that first step, and it's been really difficult."

Textual Data Acquisition and Use

Object of Analysis

The size of the corpora used by interviewees varied greatly, ranging from studying one novel in seven translations to mining several hundred thousand texts. While some interviewees were optimistic about working with large scale corpora, many felt overwhelmed. As one respondent explained, "datasets are getting too large to support traditional text analysis." Many interviewees found comfort in working with smaller corpora, one noting that they prefer to "fool around at small scales and try to figure out how to scale up." The unit of analysis for respondents was likewise variable. Respondents noted that they often worked with subsets of entire items, such as individual speeches, end paragraphs, encyclopedia articles, diary entries, or citations pulled from published volumes. Respondents described their research as operating at the work level, the page level, and the character level. Additionally, a high number of interviewees reported text analysis research in non-English languages, including German, Greek, Chinese, French, and Hebrew.

Working With Textual Data

Building a dataset. The most frequently mentioned data sources are listed in Table 2. Several interviewees described working with multiple data providers either to find one who could fulfill their request or because their desired data was siloed across systems. Respondents reported being unable to use existing corpora for their research and were scanning books and newspapers themselves to build a dataset. Several interviewees desired a tool that would assist in identifying and navigating documents; for example, one said, " . . . document navigation would be extremely helpful, and that's the kind of thing that people have to do a lot of: searching, bookmarking, grouping things, and looking at several segments together. Key-word-in-a-larger-context-type displays and that sort of thing would be very helpful."

Table 2. Data sources.

HathiTrust Digital Library
Early English Books Online (EEBO)
JSTOR
ARTFL
Google Scholar
English Short Title Catalog
Project Gutenberg
ProQuest historical newspapers

Normalizing and preparing data. All respondents engaged in text data normalization and cleaning. Such procedures include spelling regularization, part-of-speech tagging, translation, and tokenization. Deduplication was likewise an important part of preparing data and was done via several methods, including hand-selecting documents, using algorithms to match text, and comparing metadata. Data preparation also included structuring previously unstructured text data and storing it in databases that allowed the respondent to create visualizations or interact with it using mechanisms for linked open data.

Data sharing. Many of the respondents had plans in place for sharing their data. They valued keeping track of data, particularly derived data, as well as the underlying code used to carry out text analysis. Several of the interviewees noted that humanists were not accustomed to data sharing, but most acknowledged the importance of allowing others to reproduce their work. One respondent described this process as especially important with growing collections, such as the HathiTrust Digital Library, because it is "shifting ground" as the collection changes and develops. Some work with their library or institutional repository to preserve data for the long term. Others turned to third-party sources, such

as Google Drive, Zotero, and GitHub, to store their data, or they planned to make their data available via their project's website.

Data Challenges

Respondents noted that gathering data is often a difficult, involved process. As one respondent observed, building a dataset was "Very time-consuming, labor-intensive, and there's temptation on the part of scholar to want to turn it into an editing process." Copyright was a frequent obstacle in accessing desired data in the first place, and several respondents cited how their research required in-copyright text, or they needed institutional or publishers' permissions to use the data. As one respondent stated, "I did work with ProQuest and *The New York Times*. I had an article that came out in an academic journal, and I worked with them to get permission to use an image before, so I know that it is possible. But the process was so long, and it was for three images. I'm going to have thousands of files."

After acquiring the data, there were a multiplicity of challenges involved in cleaning and filtering it. One respondent noted that messy OCR was a frequent headache, and that "getting good data is the first challenge." Part of 'good data' was having accurate and clean metadata, which was highlighted as a key difficulty by more than one respondent. As one explained, "There's so much that we want to do text analytics-wise on the collection, but using [and] cleaning up the metadata and getting us to a collection that we feel is clean enough to give us back interesting results has been what we have been spending the last one and a half years on at least."

Verification and authoritative review of the textual data was another key issue: One respondent expressed that "We need that corpus, and we need basically the data exposable. I mean, we have to be able to view, not only manipulate, but also view the data." Another respondent suggested a potential peer review process for data, observing that "I would want more mechanisms for having the data sort of checked and rechecked, potentially using something like double blind methods and things like that."

Another major issue in working with text data was interoperability of data sources and melding together multiple content sources into a dataset. One respondent observed, "the newspaper archives that exist . . . all have their own siloed, siphoned search system and metadata collection mechanism, and these things do not necessarily always talk to each other, try as I will to get everything into Zotero in some kind of unified form." Another similar issue was being able to analyze data from the HathiTrust alongside data gathered from elsewhere, as one respondent described, "we're going to have to perhaps digitize some texts ourselves that we can't find through Hathi . . . . How do we turn that photo of the text into something that can then be comparable and run alongside with the stuff that we're getting from HathiTrust?"

Tools for Text Analysis

Types of Tools

Respondents described using many different tools, as seen in Table 3. They demonstrated multiple understandings of what constituted a tool. Some described software with a graphical user interface, and one interviewee noted that nontechnical faculty at their university had experienced success with these off-the-shelf tools. Others described toolkits that consisted primarily of various programming languages and their associated code libraries. When asked to describe the kinds of tools for text analysis they would find useful, respondents most frequently mentioned tools for visualization and document discovery and often suggested tools that allowed them to do iterative, incremental work. For example, one interviewee described a tool that would allow a researcher to, "get the documents, do an analysis, dump out locations or something, and then feed that back up. Then look at . . . the output of the topic model, upload the output, and then you could use that to navigate the documents. Without actually requiring them to do the topic model, have some kind of interchange format."

Building Tools

Just over half of the respondents were engaged in tool building. For some, tool building is the results of reusing existing code or of matching method to research question. As one respondent described their research collaboration, they noted that "the sentiment analysis has involved making up tools to fit the question, too, making up approaches and methods to say how do we do that, are we

interested in the whole thing." Others preferred maximum control over their workflows, one respondent noting, for example, "I end up doing a lot of things myself, because I want to know how things work, the complete pipeline. We stop at some point, no one is building their own operating system or anything like that, but I like to know from beginning to end."

Table 3. Tools used for text analysis.

Voyant
Juxta
Gephi
Tableau
MorphAdorner
D3.js
MALLET
Zotero
Weka
Python: NLTK, SciKit-Learn

Text Analysis Tools Challenges

The challenges that respondents encountered with tools primarily focuses on understanding and trusting tools. The phrase *black boxes*, in particular, was invoked by interviewees. Several respondents acknowledged that some in the humanities are intimidated by the statistics necessary for conducting text analysis. One person said, "When I talk about computational methods, my sense is that it's a little black-boxy to [students] . . . Things go in, stuff comes out, we visualize it. Therein lies a huge danger: How to create understanding around or even excitement for something like statistics." One technologist stated, "I think what we have to do is be able to offer humanists tools that are powerful, can work with the data, but not require them to do any kind of complex thinking about the computational aspect. They don't want to do the command line."

Nevertheless, several the respondents cautioned against tools that obscured the technical and mathematical processes in a black box. One said, "if you're going to [build tools] make them very transparent . . . 'this is how we're tokenizing, this is what a token means for this tool, these are the stop words lists, we're segmenting these by paragraph, we use this algorithm to determine the sentence structure.'" A researcher just getting started in text analysis said, "stumbling upon the [HTRC] portal and seeing the algorithms, it's a little daunting to know how to get the information in and what comes out. And how to format it."

Discussion and Conclusion

Our work thus far reveals that the scholarly practices involving text analysis are collaborative and complex. The wealth of text data available has facilitated computational text analysis, but it is still difficult and time consuming for researchers to access content in desired formats. In fact, larger collections mean there is more to weed through: One respondent observed that "the million books paradigm is fascinating, but there's a lot of straw you have to get through to get to the needle." Our findings reveal that textual analytics is a multistep research process carried out over numerous systems and technologies that researchers wish to be able to easily move between. Additionally, we found that researchers are concerned about data sharing and reproducibility, and as such, they want to be able to reference their datasets as well as understand, and place trust in, the tools they use to do their work. Our findings indicate a preference for tools and services that privilege both the inward and outward flow of data.

This paper is a preliminary analysis of our study thus far. In a future phase of this study, we will incorporate a broader representation of disciplines, including the addition of interviews with social scientists. We also hope to explore the integration of HTRC tools into research workflows. While the results of the study are informing current technical development of the HTRC to best meet researchers' needs, our findings also begin to reveal the requirements of researchers as digital humanities tools and resources evolve, and sheds light on how libraries can begin to meet researchers' resource needs. These preliminary findings provide insights into how librarians, technologists, and publishers of textual content can best support digital scholarship.

Acknowledgments

Nicholae Cline, Sayan Bhattacharyya, Angela Courtney, and Alex Kinnaman also contributed to this research. The HathiTrust Research Center is made possible by generous funding from the Andrew W. Mellon Foundation and the National Endowment for the Humanities. Thanks to an Institute for Museum and Library Services grant (RE-02-14-0023), which afforded some staff contributions to this project.

References

Frischer, B., Unsworth, J., Dwyer, A., Jones, A., Lancaster, L., Rockwell, G., & Rosenzweig, R. (2006). Summit on digital tools for the humanities: Report on summit accomplishments. Institute for Advanced Technology in the Humanities, University of Virginia. Retrieved from http://www.iath.virginia.edu/dtsummit/SummitText.pdf

Gibbs, F., & Owens, T. (2012). Building better digital humanities tools: Toward broader audiences and user-centered designs. *Digital Humanities Quarterly, 6*(2). Retrieved from http://www.digitalhumanities.org/dhq/vol/6/2/000136/000136.html

Green, H., & Courtney, A. (2015). Beyond the scanned image: A needs assessment of scholarly users of digital collections. *College & Research Libraries, 76*(5), 690–707. https://doi.org/10.5860/crl.76.5.690

Toms, E. G., & O'Brien, H. (2008). Understanding the information and communication technology needs of the e-humanist. *Journal of Documentation*, *64*(1), 102–130. https://doi.org/10.1108/00220410810844178