HathiTrust Research Center User Requirements Study White Paper

Eleanor Dickson, Harriett Green, Leanne Nay, Angela Courtney, Robert McDonald

Submitted: March 9, 2018

Introduction

We present findings from an investigation into trends and practices in humanities and social sciences research that incorporates text data mining. As affiliates of the HathiTrust Research Center (HTRC), the purpose of our study was to illuminate researcher needs and expectations for text data, tools, and training for text mining in order to better understand our current and potential user community. Results of our study have and will continue to inform development of HTRC tools and services for computational text analysis.

The study sought to uncover and anticipate the needs of researchers who use text analysis as a method from both technical and behavioral perspectives. The study consisted of a series of interviews with researchers, librarians, and other academic staff whose work involves computational text analysis in order to explore why, when, and how scholars draw on this research methodology, broadly conceived. It investigated the tools and infrastructure required for this area of scholarship, as well as the research questions, methods, and skills-development that motivate and drive text analysis.

Additionally, this report presents a list of functional requirements and recommendations expressed during the interviews, as well as synthesized user personas. These requirements and personas offer a user-centered guide for HTRC development.

Study Design and Methods

We conducted 18 interviews with researchers, librarians, and academic staff who are involved in text analysis research. The interviews took place from 2015 through 2016, both by phone and at professional conferences, including HTRC UnCamp, the DLF Forum, and the Chicago Colloquium for Digital Humanities and Computer Science. Interviewees were recruited for the study via targeted recruitment emails. Several other participants were recruited via the HTRC user group email list, or because they were identified as scholars who use text analysis at either the University of Illinois or Indiana University.

Study participant demographics are described below:

- **Researchers:** 4 faculty members, 3 postdoctoral researchers, and 5 graduate students.
- **Academic staff:** 5 people who lead or work in digital humanities centers or other related initiatives and who participate in text analysis projects, such as directors of

campus research centers, research programmers, and developers. All had advanced degrees (master's or higher).

- **Librarians:** 8 people who have library-affiliated roles.
- **Disciplines represented:**
    - **Humanities:** English, early modern studies, comparative literature, history
    - **Social science:** Business, Law, Linguistics
- **Gender:** 7 people who self-identified as female and 11 who self-identified as male.

The HTRC Scholarly Commons team recorded and transcribed the interviews, and then performed an initial analysis through open coding using an approach based on grounded theory analysis.[1] Using the codebook developed during the initial analysis, through which themes and then codes were identified, the research team carried out further levels of coding using the qualitative data analysis software ATLAS.ti. Interview data was independently coded and then correlated by all authors to ensure intercoder reliability.

Findings

In this section we describe findings from the study in five key areas: data practices, research methods, tools for text analysis, training and skills development, and implications and impacts of text analysis research.

Data practices

Accessing data

Where a researcher gets data is highly dependent on both their research question and data availability. Many interviewees reported efforts to access data from multiple sources, describing situations where the data from one source was inaccessible, forcing them to look elsewhere, or where they needed multiple sources to create a "complete" dataset. For several respondents, the contents of a digital collection motivated their research question, though these interviewees tended to be affiliated with the digital collection they used most.

Data sources mentioned by interviewees

- Bodleian Ballad Archive
- California Digital Library
- Early English Books Online (EEBO)
- Eighteenth Century Collections Online (ECCO)

[1] *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*, Juliet Corbin and Anselm Strauss, 3rd ed., Los Angeles, CA: Sage Publications Inc., 2008

- English Broadside Ballad Archive (EBBA)
- English Short Title Catalog (ESTC)
- Gale
- Google Books
- Google Scholar
- HathiTrust
- JSTOR
- Project Gutenberg
- Proquest
- Twitter

Other researchers, typically those who study contemporary topics, made use of social media data or other web-available content, such as news articles, for their text data. These researchers were unlikely to see HathiTrust as a viable data source. When asked about HathiTrust, one business-school researcher who primarily uses Twitter and online forum data said, "the HathiTrust has all those un-copyrighted books in digital, right?"

When data cannot be accessed via existing repositories, either because it does not exist or is not available due to rights issues, a number of researchers turned to generating their own text data. Respondents described purchasing or borrowing from libraries paper books, and then scanning and OCR-ing them. Those who work on twentieth-century, obscure, or popular culture were likely to employ this method.

We found that data availability continues to be a major sore-spot for researchers. An interviewee explained, "I think the biggest challenge is data, getting good data to work with. I think people underestimate the problems and difficulties in doing that." We heard from multiple respondents that they have not used HathiTrust for text data because they perceive that it has poor OCR quality, or because they faced difficulties when trying to get access because of licensing and rights restrictions.

Copyright is a particularly thorny challenge for researchers. Multiple interviewees described roadblocks in their research posed by copyright. One saying, "they're locked behind closed doors. And that's very frustrating. Because we know that they are digitized. We know that they are out there… To think that they're so close and yet so far away is very frustrating."

Building research corpora

Corpus size

Corpora size varied drastically amongst interview participants. They ranged from one novel in seven translations to 350,000 volumes. Additionally, respondents described corpora that dealt with text at a variety of scales, from the work or book level, to the book-chunk, paragraph, and citation level.

Several interviewees found comfort in working with smaller corpora and manageable datasets. The ability to curate a collection and dig into the details were cited as the key benefits of small datasets. Another respondent saw the value in working with small corpora as a way to "fool around at small scales and try to figure out how to scale up." While some interviewees were optimistic about the research possibilities opened by working with large-scale corpora, others felt overwhelmed. As one respondent explained, "datasets are getting too large to support traditional text analysis."

Data format and storage

A majority of respondents discussed working with PDFs as a step to accessing plain text, and also expressed the need for access to full text materials. Although some interviewees analyzed characters or parts of speech, they still wanted access to entire works. They were weary of too much intervention with the data by the provider on their behalf. As one interviewee remarked, "Even if you had somehow structured your texts, I would be saying, 'What was left out? How do I bring it back in?'"

A number of interviewees were engaged in projects that relied on metadata, either for corpus-building or as an object of analysis. One respondent commented that the field of digital humanities "makes metadata visible," explaining that DH work requires researchers to utilize previously obscured metadata, and to make their own metadata (and methods) publicly visible. The importance of strong metadata and the challenge of inadequate metadata was discussed in several interviews.

Study participants expressed plans to create their own databases, typically to store objects and to add robust metadata or improved search functionality. Several interviewees also described outgrowing simple spreadsheets to manage metadata as their projects grew. In general, the respondents wanted more control over their data and metadata, its storage, and its access.

Identifying corpus material

Even for materials ostensibly available to researchers, identifying items for analysis was described as challenging for some. Interviewees identified text through both known-item searches, for example by title, or through searching full-text and metadata. Researchers used metadata criteria, such as volumes cataloged under a certain subject heading or within a prescribed date range in order to create lists of volumes to analyze. Interviewees were less likely to mention full-text searches as an identification strategy.

Selecting the "best" representative of a text, managing duplicates, and discarding irrelevant parts of volumes were identified by study participants as part of their corpus-building processes. These issues become more acute at scale. According to another respondent, working from very large collections to create a research corpus means there is "a lot straw you have to get through to get to the needle."

Duplicates were of particular concern to interviewees. As one study participant described, "...our search has been clouded by all these duplicates... there are so many copies of the same thing and some of them are in different languages… it was just a messy dataset to begin with because we used subject headings to create what we thought would be [relevant text] based on subject headings." For this person, narrowing their corpus was an especially noisome problem.

Data cleaning

Data preparation represents a significant part of the research process, and nearly all respondents described data cleaning as a step in their text analysis workflow. One researcher said it had taken her research team over a year to clean their data satisfactorily.

For several highly-experienced researchers and those who consult on projects as academic staff, they noted that data cleaning can present a sticky issue for humanities researchers. One person who works in a digital humanities center said, "I think getting good data is the first challenge, or is one early challenge. We've done and continue to do a lot of scanning and OCR, which is very time-consuming, labor-intensive, and there's temptation on the part of scholar to want to turn it into an editing process, which can be endless, really. If you like to get the text right, it can be dangerous." He noted his group would, "talk about what's possible and what comes out and what matters and might disappear in the noise of analysis" to researchers who have questions about messy data.

Another person who works on text mining projects said they would find helpful a tool that allows researchers, "ways to clean and use data that in a sense don't look like to the researchers like that's what they are doing." These imagined tools would mask, "the behind the scenes function of cleaning up the stuff."

Nevertheless, it seemed important to our respondents that they control the data cleaning process. Some interviewees were concerned about qualities of their data that may make out-of-the-box cleaning tools unusable, such as non-standard alphabets and foreign languages. And one researcher who works on machine learning projects said, "Then, from there, I want mess. I want a lot of noise. I want a lot of error." This researcher would have been dissatisfied with a pre-cleaned dataset.

Sharing data

Respondents were aware of the importance of data sharing for transparency and reproducibility, and 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 social scientist said that best practice for sharing data had been a "debate within [his] discipline." One respondent described the data sharing 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 were working with their library or institutional repository to preserve their data for the long term. Others were using third-party sources, such as Google Drive, Zotero, and GitHub. Still others planned to make their data available via their project's website. One interviewee said, "In some ways GitHub is an integral part of this. It's like, we can try to describe this code, or you can go look at our code, right...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, because you might be wondering, 'What the heck are they doing?' So it's a bit of a hybrid publication." This response characterizes the way in which respondents recognized data sharing as integral to the publishing process.

Research methods

Approaches and methods

Respondents sought to apply text analysis methods in order to answer research questions in new and exploratory ways.

One interviewee described the appeal of text analysis thusly, “But 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.” Respondents tended to describe the purpose of text analysis as challenging existing narratives, testing currently-held theories, or solving problems otherwise impossible without access to digital data.

Some observed that they have engaged in an extensive process of matching approaches and methodologies to their research questions. In describing their research collaboration, one respondent noted that for their project, "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."

Another interviewee described the relationship between data, research question, and results thusly: “I think if you are sort of in the, I don’t want to say traditional, but the more scientific method-type paradigm, you identify your research question, develop your hypothesis, go collect the data and test it. That’s one approach. But then the thing that strikes me about text analytics is that I don’t know if you know *a priori* what you’re necessarily measuring. I think it’s kind of a guess.”

From the respondents, we saw the way text analysis is being integrated as just one component of a research project. Interviewees described the way they combine close and distant reading, and emphasized the importance of returning to the source text during the analysis process. One described his work thusly, “we’re shifting over to what I’m starting to call distant close reading. The distant part is that large corpus you can look at, but I think it’s a misnomer that you can only make kind of high level [analysis] when [what] you’re actually doing is really looking at individual words and their relationships to other words, and putting those in context of what surrounds them. That’s the work of close reading. So the nature of that is shifting, I think actually fairly rapidly when you think about it.”

The following list synthesizes study participants’ description of their research:

- Building tools and developing methods for sentiment analysis
- Using topic modeling to find networks and commonalities between themes
- Analyzing Twitter streams to establish trends in contemporary life
- Named entity extraction on places, events, and figures in historical novels
- Analyzing online conversations
- Studying translation networks
- Detecting plagiarism or textual borrowing
- Mixing quantitative and qualitative methods to extract, code, and analyze historical text
- Sentiment analysis around particular historical figures
- Longitudinal tracking of word usage to study historical popular culture
- Studying how a writer's beliefs changed
- Citation analysis to study networks of scholars or documents
- Sentiment analysis of social media
- Corpus linguistics across translated text

Methods used

Respondents reported using a mix of quantitative, qualitative, and mixed methods, as described in the table below.

Methods used by interviewees:

- Content analysis
- Natural language processing
- Corpus linguistics
- Network analysis
- Crowdsourcing
- Regular expressions
- Machine learning, machine classification
- Sentiment analysis
- Metadata cleanup
- Topic modeling, Latent Dirochlet Allocation
- Mix of close and distant reading
- Translations

Tools for text analysis

Tool use

Respondents reported using a wide variety of tools in their text analysis workflows, and they demonstrated different understandings of what constitutes a “tool.” Some described software with a graphical user interface, such as Voyant, while others noted that their toolkit consisted primarily of various programming languages and their associated code libraries, such as SciKit Learn or the Natural Language Toolkit in Python.

Overall, the types of tools interviewees worked with ranged from the user-friendly to the more complex. One interviewee noted that non-technical faculty at their university had seen success with tools with a graphical user interface, saying, "Scholars that we work with who aren’t all that technical have become comfortable with Voyant and Juxta, and so for certain things they will just do that and they may not ever tell us that they ran a certain text through Voyant or Juxta."

Just over half of the respondents were engaged in tool building, most commonly because they reuse existing code, or because of the control it afforded them over their workflows. One interviewee noted, “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 the analytics workflow, at least, I like to know from beginning to end.” For the respondents, text analysis is a multi-step process carried out over a number of tools, systems, and technologies.

Tools used by interviewees:

- Bayesian classifier
- JQuery
- Selenium (web scraping)
- Beautiful Soup
- Juxta
- SPARQL
- Bookworm
- MALLET
- SQL, MySQL
- Brat rapid annotation tool
- MorphAdorner
- Tableau
- D3.js
- NLTK
- TEI
- Excel
- Python
- Voyant
- Gephi
- R
- Weka
- Ggplot
- RDF
- Wordle
- HTRC
- SciKit Learn
- Zotero

Tool needs

Our interview questions asked participants to articulate what features they would find useful in a text analysis tool. We summarize their responses in this section.

Data acquisition and management support

As we described above, data access is a major issue for researchers who engage in, or want to engage in, text analysis. Interviewees especially expressed a desire for improved ways to identify and extract the content they wanted for building a corpus, particularly navigating large-scale collections to find the volumes, passages, or phrases relevant to a research project.

Off-the-shelf tools

Respondents had mixed opinions of off-the-shelf tools, with some showing more enthusiasm than others. One respondent with a positive perspective said, "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." As one respondent pointed out, "The problem when I work with computer scientists is they're the type of people who like to rebuild the engine, but most of us just want to get in the car and never even change the oil if we don't have to." By creating more robust "off-the-shelf" tools, researchers with little experience in programming or statistics could have the opportunity to engage in digital humanities work with fewer obstacles.

Advanced researchers

Others, especially advanced researchers, were likely to see pre-built tools as something that would hem them in and expressed the value of doing the work oneself. For example, one person noted, "Prepackaged tools [and a] web interface? I don't actually know that I think that would be that helpful. I think it's really useful for people to have to wrestle with that a little bit. I like that people have to break it down and put the pieces together themselves, to a certain extent, and be aware of what's going on under the hood."

These researchers were also skeptical that plug-and-play tools could offer them high-end research functionality. About such tools, one interviewee said, "I don't really know what I would get above and beyond using R. I mean, I can get into R and then I can run multiple different types of analyses, and I also have the latest and greatest techniques available out there.” For these researchers, access to the data in a format conducive to large-scale analysis was of paramount concern.

Transparent, customizable tools

Many of the respondents emphasized the importance of tools that would allow researchers to engage with computational methods without obscuring the underlying processes. Several respondents highlighted the need for flexible tools that could be used at various stages of the research process and be accessible to users of different skill levels. One said, “I guess my thought about [text analysis tools], though, is that if you’re going to do that to make them very transparent. In terms of, this is the process that is going on under the covers, this is how we’re tokenizing, this is what a token means for this tool, these are the stop words lists, we’re segmenting by paragraph, we use this algorithm to determine the sentence’s structure.”

In addition to flexibility and transparency, many researchers emphasized the importance of being able to set their own parameters. Interviewees understand the dialogical process between themselves and computational methods as dynamic and evolutionary. One person observed, “I yearn for workflows where the scholar could actually set their own tokenization rules. It would be a way that we could create less language-specific [rules] or control the language specificity of the algorithm. I think that is the real need.”

Training & skills development

Skills gap

A majority of respondents described significant technical challenges for researchers entering the field of digital humanities. A lack of experience in computer programming and statistics were the most common obstacles. While many respondents shared a similar frustration over technical barriers, one respondent explained, “I find it much easier to bring a humanist along and teach them enough computer science to be dangerous than try to get a computer scientist to understand the humanities.” Experienced researchers were more likely to note that humanists using computational methods need to understand the underlying math in order to make qualified claims about the results.

Statistics and programming were most commonly cited when discussing areas for training and development. Several respondents mentioned that students lacked skills in mathematics disciplines and also noticed a common fear of math among humanities students. Respondents dealt with the lack of statistics proficiency by seeking out consultations, forging collaborations, or hiring research help. For example, one social scientist said, “If there’s something that I feel like is a well-defined task, I’ll try to get someone to do it and I’ll just hire someone to do it.” However, several respondents noted that statisticians and computer scientists are interested in research of their respective fields, for example in developing methods, whereas humanists or social scientists tend to be consumers of that research and the methods it develops because they are concerned with the application of statistical and computational concepts.

Some experienced researchers were critical of those who they believe have over-fit their statistical analyses. Speaking about the pitfalls of text analysis research, one interviewee said, “The reason is that there’s lies, damn lies in statistics. [Text analysis] is statistics. It’s all statistics. Right? And so how you decide to optimize your dataset, and how you decide which features you’re going to use and how you’re going to parameterize the algorithm, you get different kinds of results... And so I think parameters are probably, at this point, not really understood.”

Training

In order to overcome these shortcomings, respondents identified different approaches to learning new skills including self-education, integrating digital humanities into college curricula, providing training opportunities such as workshops, and creating better readymade tools.

Self-education

Several respondents described teaching themselves new technical skills, in addition to seeking out collaborators to fill knowledge gaps. One respondent noted that while education is important, the burden of learning to perform research and write a dissertation has traditionally fallen on the individual scholar and digital humanities work is no different; the challenge lies in self-education.

Library training

Respondents who work in libraries were more likely than others to cite the library as a resource for learning new skills, whereas researchers outside the library tended to recommend that humanities students take courses in statistics or computer science. One respondent said of library-based text analysis consultations, “...optimally we would like to be able to say, ‘this is what we recommend doing, let’s work with you, and teach you how to do it.’”

When asked if he turned to the library for support, one researcher said, “To be frank, not really. To lay all the cards on the table, I think [my university] recently built a center for digital scholarship, but it really needs an NLP expert if it’s wants to be of assistance to those who are trying to do serious work. I mean, experts to whom grad students in CS would go for advice, you know, something like that, that’s what they need, I think.” This researcher instead turned to other scholars or experts on his campus for assistance.

Interviewees who are affiliated with the library spoke of workshops and events offered by their library that are geared towards beginners. These opportunities were described as an entry point for digital humanities work and a way to introduce things like the command line. Some of the training they described took place through an apprenticeship or assistantship model. One respondent said, "We have a summer workshop where undergraduate and graduate students are deeply involved, full-time, for eight weeks or so. So during that summer period it’s often graduate students who are producing the digital visualizations or who are running Mallet."

Classroom training

Relatively few respondents taught undergraduate or graduate students in semester-long courses, and of those who did, they were unlikely to teach text analysis. Several pointed to departmental culture as the issue preventing them from incorporating computational text analysis into their curricula. One researcher, who had been skeptical overall of pre-built tools did show interest in drag-and-drop tools for the classroom. He explained, “I once imagined teaching a class in which students learned a script and actually run analysis against data, but I was told that basically that class isn’t a humanities class anymore, that belongs in computer science. So at least at [my institution], that wouldn’t work within my home department. But I could bring in GUI-fronted tools, I think.”

Implications and impacts of text mining

Collaboration

Most of the respondents worked collaboratively or with the help of others. Project teams ranged from two to more than 25 members, and mostly consisted of several persons. Interviewees described working with students and faculty members, programmers, and technical staff at their university and at other universities. Collaboration presented a barrier to study participants. They noted that costs, such as money, time, and energy;

blockages in cross-institutional collaboration; and logistics and coordination were the most common barriers they face.

Funding

Many of our respondents spoke about sponsored projects, or otherwise noted the relationship between grant funding and digital humanities. They tended to see funding as important for making collaboration work, even though it was time-consuming to get. They felt that the shift to sponsored and data-driven projects had increased pressure for "successful" deliverables or positive results, which sometimes led to researchers releasing premature or low-quality results. Participants wished to see funding for exploratory projects, for meetings where they could share information, or for coordinated work. Interviewees were frustrated by the lack of a business model to make collaboration and interdisciplinary research happen within the university.

Funding sources mentioned:

- Institutional support, such as a local department or the library
- Andrew W. Mellon Foundation
- Institute for Museum and Library Services
- National Endowment for the Humanities
- European Association for Digital Humanities

Reception from colleagues

Respondents reported mixed reception from colleagues to their text analysis work. Several interviewees noticed humanities scholars' skepticism and even resistance to quantification and analytic methods. They also spoke to challenges building a career in digital humanities, finding appropriate venues for publication, the expectancy of innovation in their work, and lack of collegiality. Others found that their departments were receptive to digital methods, for example one graduate student who was training faculty in his department about digital humanities.

HTRC Functional Requirements and Recommendations

This list synthesizes and summarizes recommendations from interviewees.

Interfaces

- Use simple, jargon-free language for people who are not technical
- Minimize “platform fatigue” and moving between sites

Documentation

- Write and share improved descriptions of research examples
- Write tips on what makes a good workset and guidelines for creating one

Services:

- HTRC experts with programming skills and knowledge of HathiTrust data structures provide support for the “proof of concept” step of projects

Bookworm:

- Expand search options
    - Example: bigrams or greater
- Visualize worksets

Datasets, worksets, and data access:

- Release as much data as possible, as freely as possible
    - Example: derived datasets, public domain data, sample or curated datasets
    - Make it easier to download public domain text
    - Include METS record, H-OCR, OCR, and the images themselves
- Include less-technical options for access
- Ability to work with large corpora
    - Example: Compare two 10,000 item datasets
- Provide curated worksets
    - Example: n-text de-duplicated
    - Example: workset around certain themes, such as gender, genre, time period, etc.
- Improve search
    - Search and discovery via regular expressions
    - Iterative search and selection workflows
- Facilitate item selection

    - Provide statistical information about a workset to help with corpus-building
        - Example: identify outliers in a selection of volumes
    - Allow researchers to see keyword-in-context during corpus-building
    - Ability to drill into specific parts of text
- Improve data cleaning opportunities
    - Example: build OCR clean-up tool
    - Example: automatically remove headers and footers, or front matter and back matter
- Make it possible to integrate non-HathiTrust sources into analysis
- Allow researchers to enrich metadata for their worksets

HTRC algorithms:

- Build new functionality, such as:
    - Sentence boundary detection
    - Influence detection
    - Machine classification
    - Thesaurus or text normalization tool
    - Foreign language analysis (including foreign language tokenization)
- Improve the outputs and results
    - Generate better visualizations and provide tools for making them
        - Example: maps and timelines
    - Results from HTRC algorithms should not be significantly less than what is offered by external tools, such as MALLET
- Allow for robust parameter setting to give the researcher control over their work
- Rename the algorithms to make them less opaque
- Prioritize reproducibility by allowing people to run the same algorithm on the same data

Data Capsule

- Expand disk allocations
- Include standard text analysis packages in the capsule
- Explore possibility of allowing researchers to run their analysis on Karst or Blue Waters

User Personas

1) **LAURA**, DIGITAL PROJECTS LIBRARIAN

Advises and collaborates with researchers on digital humanities projects

"The algorithms don't need to be dumbed, the output doesn't need to be dumbed down even if [the researcher] doesn't know what to do with it."

Uses Mallet, R, NLTK
Wants Flexible, transparent tools
Is support staff

Laura's primary job is to advise faculty on tools, methodologies, and resources for their long-term digital projects. She tries to balance what the researcher wants with their skills.

The researchers she advises have a diverse range of skills and experience. She needs a tool to recommend to researchers that can accommodate that range. She envisions an interface easily navigated by less technical users that provides tips on building worksets, as well as statistics and visualizations about a workset's content. She also imagines a tool that would visualize results and the significance of tweaking parameters. She also thinks the ideal tool for text analysis wouldn't limit her advanced researchers, who will want to be able to set parameters for algorithms.

She is skeptical that humanities scholars need to know exactly how text analytic algorithms work or how to build one, and favors flexible, transparent out-of-the-box solutions.

Background

- Early- to mid-career librarian
- PhD in Religious Studies
- Advises faculty on what tools and methodologies best fit their digital project goals

Use case

- Worked with faculty member using topic modeling to study 150 texts
- Advised graduate student to use classifier to study trait in small corpus of text

Challenges

- Inaccessible textual data
- Data cleaning and management
- Matching researcher with tool

2) **Liz**, PROFESSOR OF COMPARATIVE LITERATURE

Publishes text analysis research in scholarly journals

"The analytic workflow, at least, I like to know from beginning to end… If there were a black box, I could peek into it. That's what I like."

Uses Sci-Kit learn
Wants compute resources, flexibility
Is an experienced researcher

Liz has been involved in large, interdisciplinary, collaborative digital projects both through her DH center and her own research.

She is always looking for new tools and methodologies that are more efficient for her projects, preferably ones that are open-source. Liz constantly encounters problems in accessing material both in and out of copyright, and is still looking for a reliable source to draw text from.

Liz needs access to clean text data with detailed metadata. She wants a system that does keyword-in-context and handles foreign language text analysis. She once tried to use the HTRC Data Capsule, but didn't think it had enough compute power. In an ideal tool, she would have access to algorithms that she can manipulate and adjust how she wants. Liz thinks the researchers she interacts with at her DH center would benefit from an improved tool for visualization.

Background

- Tenured faculty member
- 15 years experience
- Runs campus DH center

Use case

- Primarily uses Sci-Kit Learn, a machine-learning package in Python, to try to identify similarities in texts.
- Works in collaboration with a local computer scientist and a colleague at a peer institution.

Challenges

- Copyright limitations
- Finding good text
- Collaborating with others

3) **Stefan**, GRADUATE STUDENT

Uses text analysis in his dissertation research

"How do you need to clean your data? What do you need to remove? ...That kind of thing. More data management to build the workset."

Uses regular expressions, SQL
Wants more examples
Is a new researcher

Stefan is involved in several projects, both in his campus DH center and his own research.

His personal project involves a dataset of over 3000 items that He wants to analyze with text analytic methods. He is still learning, so he usually consults with others at his university for help.

He has experimented a little with pre-built tools, but isn't sure they will work for his. He knows what he wants to do, even if he doesn't know how to do it yet. He is looking for a tool that can do named entity extraction, sentiment analysis, and topic modeling. He thinks he would benefit by having access to how-to guides on how to build corpora and what to expect in an algorithm's output.

He is interested in learning by example, and is hoping he could find a website or toolkit of results and corpora that have already been used that he can browse through.

Background

- Graduate student in English
- 1-2 years experience with DH
- Plans to integrates text analysis into his dissertation research

Use case

- Conducts detailed searches of the HTDL and JSTOR
- Assembles text and other information in a database
- Creates visualizations

Challenges

- Finding text and drilling to the parts of interest
- Understanding statistics
- Working across silos of data

Appendices

Interview protocol

Workset Creation for Scholarly Analysis Study (User Requirements for Textual Analytics study) Interview Guide

Estimated length: 40 minutes

Goals of the interview: (1) To ensure that the services that the HathiTrust Research Center develops presently and in the future are adequate to the needs of the users; (2) Develop a suite of general services for HTRC users through the Scholarly Commons; (3) To obtain ideas for illustrative use cases for use in workshops and tutorials.

NOTE: "Prompts" are potential follow-up questions that are designed to draw out more in-depth explanation and detail from the interviewees when needed. Not all prompts will be asked during each interview session.

Introduction

Hello, my name is […] from the University of Illinois / Indiana University. Thank you for agreeing to be interviewed as part of the User Requirements for Textual Analytics study. First let me tell you about the study. The research team is headed by J. Stephen Downie at the Graduate School of Library and Information Science at the University of Illinois at Urbana-Champaign. The full study consists of a set of interviews designed to discover the needs of users who use or intend to use textual analytics. We hope to apply the resulting findings to developing and refining services and tools for the HathiTrust Research Center that will address the needs of our users. Your responses to the interview questions are confidential. Only summary data will be reported and no individual or institution names will be used. Before we begin, let me review the consent form and ask for your verbal consent.

Interview Questions

(NOTE FOR INTERVIEWER: The bolded questions are the main questions you should ask, and the 'Prompts" are optional. Only ask one or more of the Prompts if the respondent's initial answer does not contain as many details we would like.)

Part 1. General information

1. Describe your research area and interest in text analytics.

Prompts
How did you first become aware of the potential for text analytics in your research?
At which stage(s) of your research do you employ textual analytics?

2. Describe the relationship between your research question and the text analytic methods and approaches that you use.

Prompts
How do you determine what methods and approaches to use for your research?
How does textual analytics fit into your methods?
At what point in your research do you use analytics? (e.g., to frame research questions, for exploratory work, for confirmation/disconfirmation of your interpretive work?)
When might you use text analysis in conjunction with other methods and approaches in your discipline?

3. Where / to whom do you go for assistance when applying the textual analysis algorithms?

Prompts
Could you tell me about any experiences you've had with carrying out text analysis research with collaborators as part of a team?
Have you approached the library, digital humanities center, or other institutional resource outside of your department /college unit for assistance with your text analysis research?
What is your view of a collaborative/team process for textual analytics research?

Part 2. "Research project" questions:

4. Please walk me through a recent research project that used text analytical methods from beginning to end:

- Purpose of research / Research question(s)
- Duration of project
- Project collaborators worked with
- Desired / Actual set of texts
- Size/scope of texts
- Algorithms/Methods (if you have settled on them yet)
- Format/presentation of results (Numerical results? Plots/Graphs? Other kinds of visualizations?)

5.What challenges have you encountered when conducting textual analysis?

- Prompts
- Scale
- Access and Copyright
- Communication/collaboration
- Validation/reproducibility
- Dissemination
- Disciplinary culture

6. What do you do with your resulting data after you complete your analysis?

Prompts

How have your textual analytics research been received by colleagues in your disciplinary field?

Do you perceive any barriers to disseminating your textual analytics research in the primary journals/publication outlets for your field?

Reproducibility/reuse

Part 3: Teaching with Text Analysis Tools and Broader Needs

7. How have you used text analysis in your teaching?

Prompts

Why did you choose to use text analysis in your course(s)?

Have you used any of the text analysis tools or services provided by the HathiTrust Research Center in your classes?

When has it been effective to integrate text analysis into your course curriculum?

How does text analysis approaches fit with the learning outcomes for your discipline?

8.What tools would you be most interested in using for text analysis?

Prompts

Have you used any of the text-analytic tools or services provided the HathiTrust Research Center? If so, which ones?

Can you suggest a few additional text analytic tools / services / algorithms/resources that you are likely to find useful in the context of HTRC?

Do you use probabilistic algorithms?

Do you assess statistical significance?

Do the tools you'd be interested in working with allow you to set parameters? If so, what criteria do you use to select parameters?

Part 4. Demographic information

[For phone interviews only, in lieu of written Demographic Questionnaire]

This information will help us to characterize responses, minimize bias, ensure representative series of focus groups, inform recruitment. These questions are completely optional.

1. What is your position/title?
2. What are your academic degrees?
3. What is your organization, school or department?
4. What are your research areas?
5. What is your year of birth?
6. What is your gender?
7. What is your nationality?
8. How many years have you been doing this type of research?

Closing

Thank you for your time. Your responses will be combined with those of others to provide information about uses of large collections of digitized books and materials.

[Note that, in accordance with common interview practice, we expect to make minor adjustments to the instrument for individual participants during the course of each interview, based on their responses, the relevance of questions to their research, and how the conversation evolves.]