Digging Deeper Reaching Further

Libraries Empowering Users to Mine the HathiTrust Digital Library Resources

Module 1 Getting Started: Text analysis with the HTRC Instructor Guide

Further reading: go.illinois.edu/ddrf-resources

Narrative

The following is a suggested narrative to accompany the slides. Please feel free to modify content as needed when delivering workshops. Additional information and examples that are provided to aid the instructor's understanding of the materials are in *Italic text* – whether to present them to workshop audiences is entirely up to the instructor.

Slide M1-1

- This module covers the basics of text analysis and introduces the HathiTrust Research Center (HTRC) and the tools and services it provides to facilitate large-scale text analysis of the HathiTrust Digital Library.

Slide M1-2

- Here is a brief outline of what we will be covering in this module.
- We will first introduce the field of text analysis, and then introduce HathiTrust and the HathiTrust Research Center. Then we will look at the sample research question and case study examples that will frame the activities in this workshop.

Slide M1-3

- Our first topic is what exactly is text mining, or text analysis?
- Text analysis a subset of data analysis. Broadly speaking, it's the process by which computers are used to reveal information in and about text *(Marti Hearst definition).*
    - Computer algorithms can discern patterns in bodies of (often unstructured) text. "Unstructured" means that little is known about the semantic meaning of the text data and that it does not fit a defined data model or database. An algorithm is

simply a computational process that creates an output from an input. In text analysis, the input would be the unstructured text, and the output would be indicators to help you reveal different things about the text.
    - It should be noted that text analysis is more than just search, meaning it's not just about discovery and knowing that something is present in the text. It's also about exploring what it means for that thing to be there. For instance, knowing the word "creativity" appears in x number of volumes is only the first step; we also want to know what patterns in its appearance reveal something about literature, culture, or history.
- Text analysis can be used for a variety of purposes. It can be used for exploratory and analytical research, such as seeking out patterns in scientific literature to pick up trends in medical research otherwise difficult to see from the vantage point of an individual reader. It can also be used in developing tools that we use in our daily lives, for example creating spam filters to identify spam e-mail.

Slide M1-4

So how does text analysis work in general? Text analysis usually follows these steps:

- First, the text needs to be transformed from a form that human readers are familiar with to something that the computer can "read". This means we need to break the text into smaller pieces, and reduce (abstract) it into things that a computer can crunch.
- Counting is often what happens next. Some of the things that are often counted include words, phrases, and parts of speech. The number of these counts can be used to identify characteristics of texts.
- Then, researchers can apply computational statistics to the counts of textual features, and develop hypotheses based on these counts.

Slide M1-5

Understanding text via text analysis has a significant impact on how research is being conducted.

- In a general sense, the shift in the researcher's perspective leads to shifts in research questions. Text analysis techniques are sometimes called "distant reading". This is a term coined by Stanford professor Franco Moretti, meaning "reading" literature not by studying particular texts, but by aggregating and analyzing massive amounts of texts and "reading" them at a "distance". This scaling-up and "distancing" can bring out more insights from a very different vantage point.

- It is also worth mentioning that text analysis doesn't have to be the primary method in a research project. It may be just one step in the entire process, or it can be combined with close reading. This approach has been called "intermediate reading" or "distant-close reading".
- This shift in research perspective allows for new kinds of research questions to be asked, or for old questions to be "answered" in new ways. Here are some of the possibilities that text analysis can bring to researchers:
    - It can explore questions that cannot be answered by human reading alone
    - It allows larger corpora for analysis
    - It allows studies that cover longer time spans

As pointed out by researchers Laudin and Goodwin, text analysis techniques are often best when combined with qualitative assessment and theoretical context.

Slide M1-6

Let's pause for a brief discussion.

- What examples have you seen of text analysis?
- In what contexts do you see yourself using text analysis? What about the researchers you support?

Instructor facilitates whole-group discussion.

Slide M1-7

Text analysis research questions explore a wide range of topics, from biomedical discovery to literary history. Research questions that are conducive for text analysis methods may involve these characteristics:

- Change over time
- Pattern recognition
- Comparative analysis

Slide M1-8

Let's look at some real-world examples that show these characteristics, as well demonstrate the impacts on research we discussed earlier.

For this hands-on activity, please split into groups and review these summarized research projects http://go.illinois.edu/ddrf-research-examples:

- How do the projects involve change over time, pattern recognition, or comparative analysis?
- What kind of text data do they use (time period, source, etc.)?
- What are their findings?

Research project summaries are in the above URL for the instructor's use.

Slide M1-9

Let's look at these examples in more detail together. Here is a well-known example of text analysis that used a method called Stylometry. When *The Cuckoo's Calling* by Robert Galbraith was published in 2013, people wondered if the book was actually written by the famous JK Rowling under a pen name.

- A research question such as "Did JK Rowling write *The Cuckoo's Calling* under the pen name Robert Galbraith?" is a good fit for using text analysis methods to explore, because this question is very difficult to answer and prove just by reading and comparing various texts.
- It involves comparative analysis (*The Cuckoo's Calling* vs. other books by Rowling) and recognizing patterns between her writing and *The Cuckoo's Calling*).

Slide M1-10

Therefore, the overall approach to and process of exploring this question was:

- At first, human reading led to hunch about authorship that raised the question.
- Then, Patrick Juola conducted a Stylometry analysis to complete a computational comparison of diction between this book and others written by Rowling. By comparing a set of linguistic variables between the text of *The Cuckoo's Calling* and four other texts (one by Rowling, three by other "distractor authors" for comparison), the results suggested that Rowling's writing style is the closest to the author of the *The Cuckoo's Calling*.
- This provided a kind of statistical "proof" of authorial fingerprint, and Rowling admitted to writing the novel under the pen name Galbraith shortly after.

Juola's approach was to compare The Cuckoo's Calling with Rowling's own writing, as well as with samples written by three other authors as "distractor authors" for comparison. He broke the text of The Cuckoo's Calling into chunks of 1000 lines and compared each chunk individually against the baseline model built from each of the four candidate novels. This comparison

consisted of conducting four separate types of analyses focusing on four different linguistic variables that indicate style, including the distribution of word lengths, the 100 most common words in the text, the distribution of character 4-grams (groups of four adjacent characters) and word bigrams (pairs of adjacent words). The results pointed strongly to Rowling as the author of The Cuckoo's Calling out of the four authors studied.

Slide M1-11

Here's another example. A researcher may want to explore a question such as "What themes are common in 19th century literature?"

- Matt Jockers and David Mimno asked this question in a paper in 2012.
- This is a "huge" research question that would require a very large corpus and an impossible amount of human reading to answer. Hence, it is a good fit for using text analysis, and more specifically topic modeling, to investigate.
- It involves the recognition of thematic patterns in the writing, and compares those patterns across many works from the period.

Slide M1-12

For Jockers's and Mimno's study:

- To explore what themes are common in 19th century literature, they ran large quantities of text through a statistical algorithm to discover "topics". Topic modeling is based on the idea that words that co-occur are likely to be about the same thing, so co-occurring words are represented as topics.

Note that topic modelling is unsupervised machine learning because human inputs are minimal (the only input is the text).

Slide M1-13

Here is a visualization of one of their results. It's a word cloud consisting of a theme they identified algorithmically, which they have labeled "female fashion". We see words such as gown, silk, dress, lace, and ribbon. These are words that tended to co-occur across their corpus of nineteenth century text. Through these results, Jockers and Mimno are able to argue that authors from this time period wrote about what women wore.

Note that "female fashion" is a term they applied to summarize the group of words that the computational analysis had identified as a topic; the algorithm itself cannot give you the name of the topic. This will be explained in more detail in one of our other modules.

Slide M1-14

Here's our last example. A question like "*What textual characteristics constitute 'literary language'?*" is a good one for text analysis.

- This question covers a very large time span, so it would also be almost impossible to answer with human reading.
- It explores change over time, as well as pattern deduction.

Slide M1-15

In this example, Ted Underwood and Jordan Sellers used classification algorithms to study what characteristics constitute "literary language".

- Their approach was to show the difference in words used in poetry, drama, and fiction with those used in nonfiction to demonstrate how "literary language" developed over time. First, they trained a computational model to identify literary genres. Then, they compared which words are most frequently used over time in non-fiction prose versus "literary" genres. Their results demonstrated tendency for poetry, drama, and fiction to use older English words.

More background on this example: In their research, the authors noted that defining the distinctive characteristics of "literary language" was historically a large part of literary criticism, but that it was largely abandoned because of the changing definition of literature over time. Literature referred to writing or learning generally, up until the middle of the 18th century. The concept of literature as imaginative writing emerged gradually between 1750 and 1850. The authors were interested in investigating the changes in literary language over time, as ideas of what constituted literature transformed. They used the relative use of newer words to older words to investigate differences. Their dataset consisted of a collection of 4,275 mostly book-length documents, and they included only the most common ten thousand words in the collection and excluded determiners, prepositions, conjunctions and pronouns.

The authors explain that there are social implications to word age, because there were 200 years during which English was almost exclusively spoken, while French was used for writing. When English began to be written again, around 1250, literate vocabulary was borrowed from French and Latin. So, the older words tend to be more informal and the newer words were more learned or literate language. This is their basis in using the changing ratio of older and newer words over time to reveal patterns in the development of literary language.

They used a similar process to differentiate literary language in the different genres, and to see how this changed over time.

Slide M1-16

Here is another look at one of the graphs from their paper. It shows one way that literary diction differs from that of nonfiction writing between 1700 and 1900. The Y axis shows the ratio of old words to new words, and the X axis shows years. So, the graph illustrates word age over time. The words used in Poetry, Drama, and Fiction are shown in purple, and nonfiction prose are shown in dark gray. The graph shows a gradual increase in the use of new words in both categories until about 1775, when older words began to be more prevalent in Poetry, Drama, and Fiction.

Therefore, this analysis reveals some patterns in the development of literary language.

The authors point out that by the end of the 19th century there was a "sharply marked distinction between literary and nonliterary diction: novels were using the older part of the lexicon at a rate almost double that of nonfiction prose."

Slide M1-17

In the next section, we will introduce HathiTrust, the HathiTrust Digital Library (HTDL), and the HathiTrust Research Center (HTRC), and how they work to support large-scale text analysis. Here is a diagram illustrating where the HTDL and HTRC are in a basic text analysis workflow. Please note that this is not a naturally occurring workflow, but an optional approach that the researcher can initiate. We will talk more about how a text analysis workflow can often be non-linear and messy later in this module.

- As we briefly introduced in the previous section, in a basic text analysis workflow, a researcher:
    - Gathers digitized text (text that has been scanned and OCR-ed) *Note: OCR (Optical character recognition) refers to the mechanical or electronic conversion of images of typed, handwritten or printed text into machine-encoded text.*
    - Applies computational methods to that text, such as word counts, classification techniques, and topic modeling
    - And then analyzes the results generated by the algorithm or technique

- The HTRC enters the workflow at the points of providing digitized text at scale from HTDL and providing tools and services that enable computational research.
- The researcher, of course, still brings her own analysis to bear on the results.

Slide M1-18

- HathiTrust is a library consortium founded in 2008. It's a community of research libraries committed to the long-term curation and availability of the cultural record.
- It is an offshoot of the Google Books project, and most digitization has happened at academic research libraries.
- Currently, it has grown to over 120 partner institutions, mostly in the US, but also abroad.

Slide M1-19

- The HathiTrust Digital Library is concerned with collecting, preserving, and providing access to the content digitized at the partner institutions. It contains over 16 million volumes made up of more than 5 billion total pages.
- Over half of the volumes are in English.
- It has material going back to the 15th century, though the majority of the content comes from the 20th century.
- As a result of the 20th century concentration, about 63% of the material is in copyright or restricted because its rights status is unknown, and about 37% is in the public domain. The HT has a review project for determining rights statuses for that chunk of "unknown" material.
- While the HTRC is getting set-up to provide access on certain terms to in-copyright content, for now it provides access to the 37% of the HTDL that's in the public domain.

The HTDL has its own interface for searching, reading, and building collections of volumes in the public domain.

Slide M1-20

The HTRC is concerned with allowing users to gather, analyze and produce new knowledge primarily via computational text analysis.

- It facilitates large-scale computational research of HTDL content.

- The “center”, which is located at Indiana University and the University of Illinois, does various R&D for non-consumptive text analysis, a concept that we will be explaining in later slides.
- It conducts user studies, finds cutting-edge technical solutions, and builds tools and services.

Slide M1-21

HTRC tools and services operate within a “non-consumptive” research paradigm.

- This is a term for text analysis research that lets a person run tools or algorithms against data without letting them read the text. The precise definition is shown here on the slide *(Research in which computational analysis is performed on text, but not research in which a researcher reads or displays substantial portions of the text to understand the expressive content presented within it).*
- Non-consumptive research complies with copyright law because of the distinction in law between “ideas” and “expressions”. It is sometimes called non-expressive use (because it works with “ideas” instead of specific “expressions”, hence the term “non-expressive”).
- The non-consumptive research paradigm is the foundational underlying structure of HTRC work.
- *This term relates to how one interacts with the text, regardless of whether the data is in-copyright or in the public domain, though it comes out thinking around how to provide analytic access to rights-restricted text.*

Slide M1-22

Before we move on, let’s do a short discussion related to what we just covered. Are you (or your colleagues) currently offering research support for text analysis? How so? Why or why not? What kinds of questions and/or projects does your library handle?

(Optional: break participants into pairs or groups for discussion)

(If time runs short, you can skip Discussion 2 and combine Discussion 2 and 3 into one group discussion at the end. See Discussion 3 for how to adjust questions.)

Slide M1-23

The main topics of this module end here. In the final section, we will introduce the outline and structure of our workshop, and talk about a sample research question and case study that we will keep returning to throughout the workshop.

Here is the outline of the workshop.

The workshop content is divided into modules, and the modules generally follow the research process of conducting text analysis.

- We will cover two modules about gathering textual data
- One module about working with textual data
- Two modules about analyzing textual data
- One module about visualizing textual data

In each module, we will complete hands-on activities around a central research question and discuss a case study example at each step of the research process.

We will use both HTRC and non-HTRC tools as our goal today is to develop transferable skills.

Slide M1-24

This workshop will cover: Process, approaches, and key topics in text analysis.

You will leave today having a better sense of text analysis and how it might be used by researchers on your campus. We cover both programming concepts and computational methods, but don't intend that you would be able to go from this workshop straight to a high-end text analysis research.

Slide M1-25

In the workshop, we will keep returning to this sample text analysis reference question.

- Imagine that a student comes to you with this research topic: "I'm a student in history who would like to incorporate digital methods into my research. I study American politics, and in particular I'd like to examine how concepts such as liberty change over time."
- As we work from one module to the next following the research process, we'll practice different approaches for answering this question throughout the workshop.
- The question will also help us frame the role of the librarian in supporting or participating in text analysis research.

Slide M1-26

Throughout the workshop, we will also be introducing a real case study of text analysis step by step.

- The case study is called *Inside the Creativity Boom* and was conducted by researcher Samuel Franklin. He worked with HTRC to explore how the use and meaning of the words *creative* and *creativity* changed over the 20th century.
- We'll discuss how this researcher approached his question throughout the workshop to give you a more concrete idea of what needs to be considered and done in each step of a text analysis research project.

Modified from project report abstract:

The project set out to make use of the HathiTrust corpus to map the career of the words "creative," "creativity," and their less common variants over the last several hundred years, with an emphasis on the twentieth century. It was already known that the word "creative" emerged gradually over the course of the modern era, increasing in use rapidly in the twentieth century, and that "creativity" only barely appeared around the turn of the twentieth century and exploded into the regular English lexicon in the post-WWII era. In order to discern the relationship between these two patterns, and to figure out if the recent rise of "creativity" signifies simply the popularization of a pre-existing concept or a new conceptual formation, a more granular analysis of these trends was necessary. Have the increases in the use of the word types "creative" and "creativity" been distributed evenly throughout the printed corpus, or have they clustered around certain fields, genres, or communities of discourse? To what topics, activities, and types of people have those words pertained? Is it possible to discern variation and change in the meanings of those words across and between genres, fields, and eras? To answer these questions, [he] proposed utilizing a number of different analytical tools such as collocates, topic modelling, and faceted queries, using the HathiTrust corpus and subsets therein.

Link to final report:
https://wiki.htrc.illinois.edu/download/attachments/31588360/Franklin_ACS_End-Report.pdf?version=1&modificationDate=1506961462000&api=v2

Slide M1-27

Before we move on, there is one point that needs to be emphasized. Our workshop outline and modules may appear to suggest the research workflow of a text analysis project is linear and follows a predetermined sequence one step after another, like this diagram on the slide. One finds text, prepares the text, analyzes it with algorithms and visualizes the results.

Slide M1-28

However, an actual text analysis workflow is usually much more complicated and is rarely a linear, sequential process. It's often more like the diagram on this slide. Depending on the project, a researcher may repeat certain steps in small cycles, or return to previous steps, or do some exploratory steps to determine next steps. Therefore, it is important to note that while we are using the modules to represent the general, progressive process of text analysis research, things are a lot messier in real life.

Slide M1-29

Let's end this module with a group discussion. In small groups or pairs, please discuss what do you think are some of the characteristics of a good candidate research question/project for using text analysis methods?

(Instructor organizes discussion)

(If Discussion 2 on Slide 22 was skipped, you may ask participants to discuss the questions below:

Are you (or your colleagues) currently offering research support for text analysis? What kinds of questions and/or projects does your library handle? If not, what do you think are the characteristics of a good candidate research question/project for using text analysis methods?

This can be a group discussion of about 5-8 minutes.)

Slide M1-30

That's all we have for this module, and we will be happy to take any questions from you.

Slide M1-31

(*Display references so attendees know they can find them later.)*