Introduction to Text Mining and Analysis

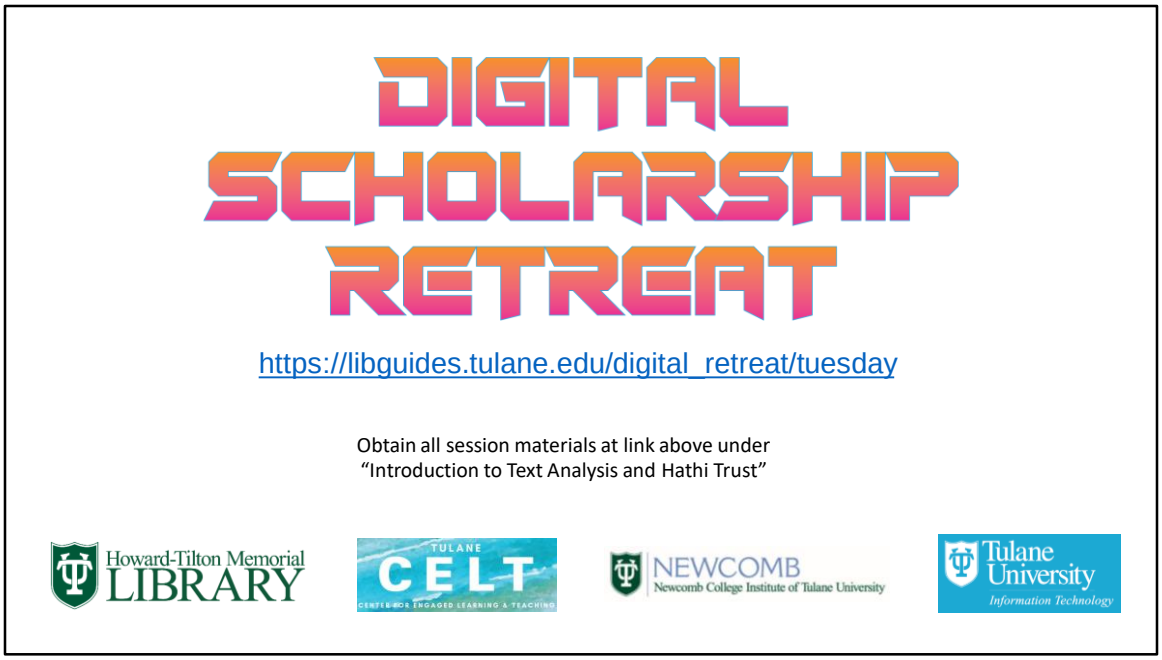
Jane Marie Pinzino
Tulane University

Artifact type: Workshop
Developed for: Teaching faculty and teaching assistants
Audience: Humanities faculty and graduate students
Time required: One 90-minute session
Method and tool: Text mining and analysis with JSTOR Text Analyzer, Hathi Trust Digital Library, and the Hathi Trust Research Center

Description:

- This workshop serves faculty and graduate students who teach undergraduate humanities courses, primarily English and History. By providing simple-to-use, out-of-the-box tools, the workshop empowers instructors to build up undergraduates' digital competencies and skills transferable to the workplace.
- The workshop begins with a discussion of key terms and ideas in text mining and analysis. Text mining is a subset of data mining, and provides an opportunity for processing large bodies of text to discern patterns in ways that human reading cannot. Text mining does not replace "close" or "human" reading, but supplements it.
- The workshop participants copy and paste a CNN article provided to all into JSTOR Text Analyzer. This tool is a real crowd-pleaser! The tool discerns keywords useful for searching, and connects the user with relevant articles for research.
- HathiTrust is an enormous digital library spearheaded by University of Michigan, together with the HathiTrust Research Center's text mining and analysis tools provided by University of Illinois and Indiana University. Workshop instruction first demonstrates how to login, search and create a collection in HathiTrust.
- HTRC Bookworm is a pleasant introduction to the analytic tools. Ask the workshop participants to search the history of terms and share with the group what they discovery.
- The final component of the workshop is to demonstrate how to create a workset and run the Named Entity Recognizer.
- Closing discussion goes around the room for participants to express what they would like additional workshops to offer, and a small, handwritten assessment is taken about today's workshop.

Supporting materials:
Workshop slides with lecture notes, and workshop assessment tool


DIGITAL
SCHOLARSHIP
RETREAT
https://libguides.tulane.edu/digital_retreat/tuesday
Obtain all session materials at link above under
"Introduction to Text Analysis and Hathi Trust"
Howard-Tilton Memorial LIBRARY
TULANE CELT
NEWCOMB
Newcomb College Institute of Tulane University
Tulane University
Information Technology


Introduction to
Text Mining and Analysis
Jane Marie Pinzino
DIGITAL
SCHOLARSHIP
RETREAT

SLIDE 0

What is it? text mining? Digital text analysis?

Text mining is a subset of data mining, and text analysis that follows involves running algorithms on the text that is mined, and finally putting forth the interpretation on the patterns discerned in the results.
We **all** already do text analysis whenever we use Control F or Command F to find a word in an electronic document and consider all the instances of a word in a text. If, based on our comprehensive examination of the ways the word is used in the work, we were to advance a research hypothesis that would be a simple form of engaging digital text mining and analysis. But in digital humanities, the number of texts examined together tend to be large in number, several texts up to hundreds or thousands of texts, and even more.

In 2013, Robert Galbraith published a novel titled *The Cuckoo's Calling*. One scholar upon reading the work, had a hunch that that "Robert Galbraith" could be a pseudonym for J.K. Rowling. He peformed digital analysis on the entire corpus of Rowling's work, using a method called stylometry, which counts word frequencies, paragraph lengths, sentence lengths, the presence or absence of dialogue, etc. other

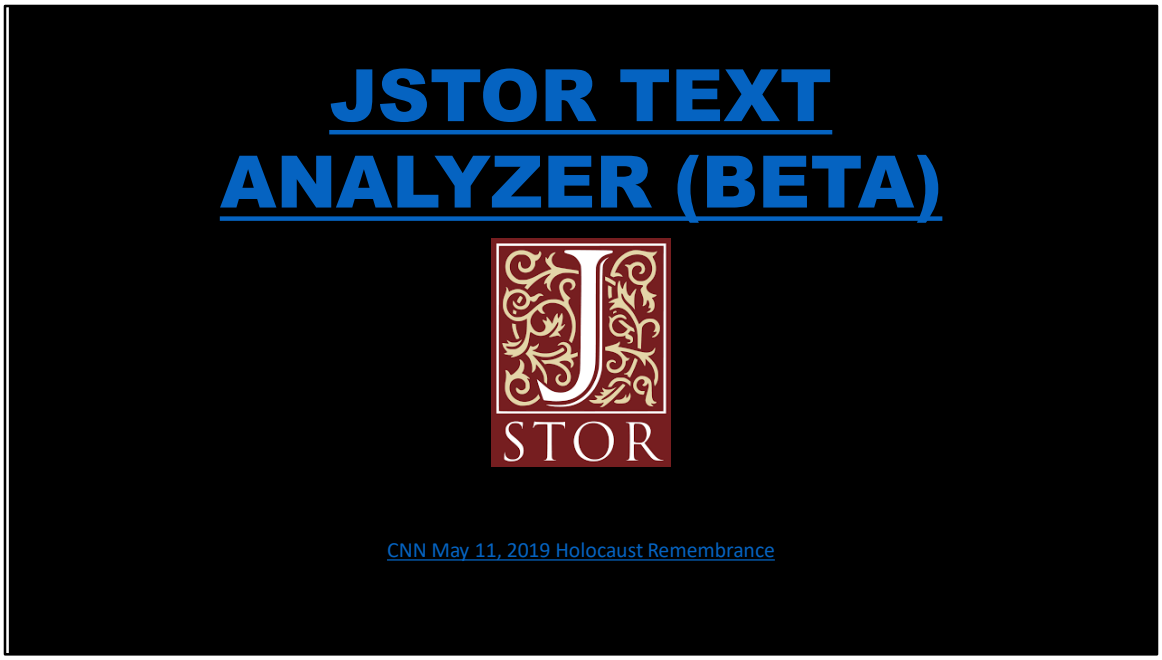
identifiers of style, and pronounced this novel a work of J.K. Rowling, which she "eventually" confirmed.

The methods for text analysis grow more sophisticated, depending upon the complexity of the research question. It involves implementing algorithms that discern patterns in bodies of text, including counting words and determining frequencies, identifying parts of speech, bringing to the surface collocations of words which are used for stylometry and topic modeling. The DH community uses the Latin word *corpus* for a single body of text and *corpora* for multiple bodies examined together and in contrast to one another.

Text analysis can ferret out patterns of ideas in primary sources, or detect clues and identifiers in email that can automatically classify messages as spam to be routed to a separate folder. Machine reading can do these things for corpora, ie. large bodies of texts, in a way that human reading cannot. It is not physically possible to read and absorb that much. One does not ever replace the other, but they can be complementary research methods, distant reading and close reading.

The process for text analysis, generally speaking is to 1) Search for and Get access to your corpora; 2) clean the corpora (many corpora including HathiTrust are "dirty"; the OCR can be poor, or the corpora need to be normalized (for example if you are searching for uses of the word "color" you might have to go through and make changes for spelling-- selecting colour "OUR" or color "OR". Doing that manually can be daunting, so often coding is used to assist clean up.

A couple of more examples of text analysis projects: Ted Underwood, a prominent digital humanist out of Illinois, wanted to find out how the language for literary genres, including fiction, drama, and poetry, differed from that of non-fiction. So he carried out a sweeping linguistic comparison that covered a couple of centuries and thousands of English-language texts, and came to the insight that literary genres of text tend to make greater use of older English vocabulary, for example, adjectives like comely, diaphanous, and bucolic, while non-fiction reflected newer words and terms--for example, anthropocene, or microagression. Underwood, in a different project, wanted to know how women and men are portrayed in the literary genres from the 18th century to the present day. And he found that over time, the features that clearly distinguished a woman from a man leveled out over the centuries so that men and women became more like each other in literary portrayals.



Let's start with JSTOR Text-Analyzer, it is a good icebreaker. User friendly and useful for research.

Select all and copy the CNN article on "Holocaust Remembrance". This news article covers a recent white supremacist protest of a Holocaust Remembrance gathering at Arkansas State. The protesters were holding signs that read "The Holocaust did not happen but it should have". We are formulating the research question "what values are at stake in the white supremacist denial that the Holocause occurred?"
Adjust keywords to include "anti-Semitism", "Nazi", and see what comes up. Note the subject term "historical revision" appears, though that term never occurred in the CNN article!

This "out-of-the-box" tool offers useful keywords for searching, and connects users with scholarly articles on their topic. The user gets a sense of the scholarly conversation surrounding the topic, and locates specific articles, authors, and ideas useful to the research project.

Pronounciation is "Hah-Tee Trust" with a silent h. The word Hathi comes from the Hindi for elephant, referring to an elephant's capacity for memory in light of the mission of HathiTrust.

The **mission** of **HathiTrust** is to contribute to the common good by collecting, organizing, preserving, communicating, and sharing the record of human knowledge

There are more than 140 institutional members of the HathiTrust in the U.S. and beyond, and Tulane University is one of those members. Lucky for us, we have this unbelievable digital libary at out fingertips, which comes with out-of-the-box tools for analysis.

The collection contains over 17,000,000 volumes, including monographs, serials, and a complete collection of U.S. government documents. To get a sense of scale, Howard-Tilton has about 4,000,000 volumes so we can imagine the contribution to our collections that HathiTrust offers.

Approximately 60% of the collection is in copyright or of undetermined copyright status, and that impacts the type of availability for research a particular volume has. HathiTrust may archive the full-text of an item that is in copyright but the user cannot download it. In some cases they may view it, but in all cases the user may run text analytics on it, that generate a list of results. HathiTrust won a lawsuit that allows its members to do this. Thus, they get a list of results of a text-mining search useful for their research, and it is up to the researcher to obtain a print copy of the work to examine the generated list in light of a hard copy. I will demonstrate this shortly.

In the digitization process, there is a plain-text OCR file created, useful for text-mining, but it is not hand-corrected, so there are errors. In a volume where the plain-text file is just too rough to be of use, HathiTrust works on obtaining a cleaner plain-text file and a user can request this.

The HathiTrust Research Center, is a joint project between the University of Illinois and Indiana University that provide computational tools and support for the HathiTrust Digital Library members to perform text analysis on the corpus or the collections created by the user.

Non-consumptive reading is machine reading not human reading. You are getting at the ideas in the text, but not the expression of them.

What is in the collection? Whatever the member libraries scan and contribute. About 50% of the titles are in English, followed by numbers in German, French, and Spanish, so the collection is and has been heavily U.S. and European, though increasingly other languages including tribal languages both American and global have been scanned and incorporated.

There are number of duplicates in the collection, both title duplicates and edition duplicates. That requires decisions on the user's part of what and what not to include.

The titles are mostly **not** born-digital, but have been digitized by the member institutions,
with materials from the 15th century, including manuscripts, to the present day, with a heavy concentration on the print materials from the 20th century. The scope of the collection has entirely to do with the choices of the contributing libraries. The largest single contributor is University of Michigan, and the second is the university system of California, followed by Harvard 3rd.

HathiTrust bookworm allows a user to visualize the usage of a term(s) over time. Google ngram viewer is a precursor to Hathi Trust bookworm, and has enhanced features over Google, including faceting options as well as a collection with texts up to the current year. Google last updated its ngram collection in 2009.
HathiTrust bookworm allows a user to visualize the usage of a term(s) over time. Google ngram viewer is a precursor to Hathi Trust bookworm, and has enhanced features over Google, including faceting options as well as a collection with texts up to the current year. Google last updated its ngram collection in 2009.

Exercise: Do sample searches in Bookworm and go around the room to hear how this might be used in the classroom.

HathiTrust Research Center

The Hathi Trust Research Center, for which you created a fancy password,

Three approaches to using HTRC:

Web-based tools that allow you to extract and analyze a dataset.
The user can run a tool having no access to the underlying data for human reading (because of copyright restrictions). Text-mining without reading provides lists of identifiers with no customized tweaks, therefore suitable for classroom research, or for the researcher who does not want to learn to code.

2) "Derived datasets"—provide transformation of extracted features
Crunches the data in advance; pre-processing; one step of the way there; processes tokens (words) and how many times they appear on the page: parts of speech tags, page level metadata; make fuzzy the expression of the text; techniques that work with bags of words, but do not account for the context the words

3) Secure "data capsules"--access to the full text allowed but the outputs are highly controlled because of copyright protections

Exercise: Create Workset (metadata from a list of sources) and run Named Entity Recognizer

Highly Recommended On YouTube

Text Mining with Hathi Trust (Library of Congress)

Be in touch!

jpinzino@Tulane.edu
Coordinator for Scholarly Resources for the Humanities

Workshop Assessment Tool

Rate the facilitator's ability to present the material in a comprehensible way.

very good | good | fair | poor | very poor

I learned something new from this session that will help me with my research or teaching.

strongly agree | agree | neither agree or disagree | disagree | strongly disagree

How satisfied are you with the session?

very satisfied | satisfied | neutral | dissatisfied | very dissatisfied

Please share any comments about this session: