Digging Deeper Reaching Further

Libraries Empowering Users to Mine the HathiTrust Digital Library Resources

Module 1 Getting Started: Text analysis with the HTRC Lesson Plan

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

This lesson is a basic introduction to text analysis and the research methods it encompasses. It also introduces the HathiTrust Research Center (HTRC) and the tools and services it provides to facilitate large-scale text analysis of the HathiTrust Digital Library.

Estimated time

20-30 minutes for set-up, 30-45 minutes for module

Audience

Librarians with little-to-no experience with text analysis and/or the capabilities of the HathiTrust Research Center.

Prerequisites for participants

None! This lesson is for the true beginner.

Learning objectives

At the end of the module, the participants will be able to:

- Recognize research questions for which text analysis can be used in order to better support text analysis research on their campus.
- Relate the HTRC to text analysis research in order to understand the context for one digital scholarship tool provider.
- Understand broad text analysis workflows in order to make sense of digital scholarly research practices.

Getting ready

There's nothing for workshop participants to do in advance for this module.

Session outline

- Introduction to text analysis research in the humanities and social sciences
    - Impact on research
    - Text analysis research questions
- **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?
- **Activity:** Read and explain text analysis examples
- Introduction to HT, the HTDL, and the HTRC
- Overview of key concepts for working with the HTRC: HTRC's access model and services, and the non-consumptive research paradigm
- **Discussion:** How are librarians currently offering research support for text analysis?
- Introduction to workshop outline
    - Modules generally follow research process
    - Sample reference question for hands-on activities
    - Case study
- Discussion: What are some of the characteristics of a good candidate research question/project for using text analysis methods?

Key concepts

- **Text analysis**: A form of data mining, using computer-aided methods to study textual data.
- **Distant reading**: As compared to close reading, which finds meaning in word-by-word careful reading and analysis of a single work (or a group of works), distant reading takes large amounts of literature and understands them quantitatively via features of the text. (Conceptualized by Franco Moretti)
- **Non-consumptive research**: 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.
- **Algorithm**: A process a computer follows to solve a problem, creating an output from a provided input.
- **Optical character recognition (OCR)**: Mechanical or electronic conversion of images of typed, handwritten or printed text into machine-encoded text. The quality of the results of

OCR can vary greatly, and raw, uncorrected OCR is often described as “dirty”, while corrected OCR is referred to as “clean”.

Key tools/platforms

- **HathiTrust:** A library consortium founded in 2008. HathiTrust is a community of research libraries committed to the long-term curation and availability of the cultural record.
- **The HathiTrust Digital Library (HTDL):** A digital preservation repository and highly functional access platform under HathiTrust. It provides long-term preservation and access services for public domain and in copyright content from a variety of sources, including Google, the Internet Archive, Microsoft, and in-house partner institution initiatives. Overall, the content mostly consists of digitized books from libraries.
- **The HathiTrust Research Center (HTRC):** A research center under HathiTrust that facilitates computational, scholarly research using the 16+ million volumes in the HathiTrust Digital Library. The HTRC provides mechanisms for non-consumptive access to content in the HathiTrust corpus, as well as tools for computational text analysis.

Key points

Introduction to text analysis research in the humanities and social sciences: key approaches and examples	• Text analysis: the process by which computers are used to reveal information in and about text. • Text analysis usually involves breaking text into smaller pieces; reducing (abstracting) text into things that a computer can crunch; counting words, phrases, parts of speech, etc.; using computational statistics to develop hypotheses. • Text analysis impacts research by shifting the researcher’s perspective of the text, and makes it possible to ask questions that cannot be answered by human reading alone, larger corpora for analysis, and longer periods of study. • Text analysis research questions often involve change over time, pattern recognition, and comparative analysis.
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?

	• *Goal:* Encourage learners to make personal connections to the content of the workshop.
Activity: Text analysis research questions	• In pairs or small groups, read the research examples and discuss the key points and methods. • *Goal:* Gain exposure to text analysis research and how it is being used by scholars.
Introduction to HT, the HTDL, and the HTRC	• The HathiTrust organization is divided into roughly two parts: the HathiTrust Digital Library (HTDL) and the HathiTrust Research Center (HTRC). • The HTRC is concerned with allowing users to gather, analyze and produce new knowledge primarily via computational text analysis, based on the digitized content collected, preserved, and provided to users by the HTDL.
Overview of key concepts for working with the HTRC	• The foundational underlying structure of HTRC work is the "non-consumptive" research paradigm, which is text analysis research that lets a person run tools or algorithms against data without letting them read the text.
Discussion	• How are librarians currently offering research support for text analysis? • *Goal:* Encourage learners to make personal connections to the content of the workshop.
Introduction to workshop outline and structure	• Workshop has seven modules, modules generally follow text analysis research process • One sample reference question • One case study • Note: actual text analysis research workflows can be quite messy and are rarely linear
Discussion	• What makes our sample reference question and the case study good candidates for using text analysis methods?

	• *Goal:* Build confidence assessing whether a research question is suitable for text analysis methods.

Additional Tips for Instructors

- Leave plenty of time for participants to complete the set-up part on the handout.
- **Recommend participants NOT to use Internet Explorer for the web-based activities and choose an alternative browser such as Chrome or Firefox.** Participants using IE may encounter some issues with some of the activities. Additionally, for participants using Safari on a Mac, note that the activity_files zip may be automatically unzipped into a folder when downloaded. They will need to manually compress the folder into a zip file again by right clicking on the folder and selecting the "Compress 'activity_files'" option. Then they can upload the compressed file to PythonAnywhere for our activities.
- **Remind participants to create accounts for BOTH HTRC Analytics and HTDL for the hands-on activities.**
- When demonstrating activities in web browsers, instructors may use "Ctrl" and "+" ("Command" and "+" on Macs) to enlarge the content on the screen. It can be quite difficult to see things from the back of the room! Use "Ctrl" and "-" ("Command" and "-" on Macs) to zoom back out when you need to demonstrate other things in regular size.