Digging Deeper Reaching Further

Libraries Empowering Users to Mine the HathiTrust Digital Library Resources

Lesson Plans

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

Module 1 Getting Started: Text analysis with the HTRC

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 will 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.

Module 2.1 Gathering Textual Data: Finding Text

This lesson introduces the options available to researchers for finding and accessing textual data. In addition to discussing the variety of textual data providers, this lesson covers the process of building a text corpora in the HTDL interface and uploading it to HTRC Analytics for analysis.

Estimated time

35-50 minutes

Workshop audience

Librarians with little-to-no experience with text analysis who may be supporting research and teaching with text analysis at their institutions.

Prerequisites for participants

- Have some idea of text analysis concepts
- Have been introduced to the HTRC, or have completed Module 1

Learning objectives

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

- Differentiate the various ways textual data can be gathered in order to make recommendations for researchers.
- Evaluate textual data providers based on research needs in order to provide reference to researchers.
- Curate and select volumes to construct their own HTRC workset in order to gain experience building corpora.

Skills

- Build a collection in the HTDL and import it into HTRC Analytics as a workset

Getting ready

- Workshop participants will need:

- An account for HTRC Analytics (https://analytics.hathitrust.org). Instructors may guide participants in the registration process before officially starting the workshop session.

Session outline

- Introduction and outline
- Methods for accessing and downloading textual data
    - Challenges in finding text
    - Sources of textual data
- **Activity:** Strengths and weaknesses of different sources of textual data
- Evaluating sources of text data
- The process of building corpora
- Introduction to worksets
- **Activity:** Create an HT collection and upload a workset to HTRC Analytics
- Creativity Boom case study: How Sam built his corpora for analysis
- Discussion: What expertise do librarians already have to help with building a corpus for textual analysis?

Key concepts

- **Text corpus/corpora:** A “corpus” of text can refer to both a digital collection and an individual's research text dataset. Text corpora, the plural form, are bodies of textual data.
- **Workset**: In the HTRC environment, a workset is a sub-collection of HathiTrust content created by users.
- **Volume:** Generally, a digitized book, periodical, or government document.
- **Optical character recognition (OCR)**: Mechanical or electronic conversion of images of text into machine-readable text. The quality of the results of OCR can vary greatly, and raw, uncorrected OCR is referred to as "dirty" because it often contains mistakes, while corrected OCR is referred to as “clean”.

Key tools

- **HT Collection Builder**: An interface for creating collections via the HathiTrust Digital Library.

Key points

Kludging access: finding and gathering text	• Text can be approached as data and analyzed by corpus/corpora. • Before analyzing textual data, it is important to ensure the text is of sufficient quality (e.g., OCR-ed data is cleaned up) and fully prepared (certain unnecessary elements are discarded).
Methods for accessing and downloading textual data	• Finding text suitable for computational analysis is challenging, especially with issues of copyright and licensing restrictions, format limitations, and hard-to-navigate systems. • Three commonly used sources to find textual data are vendor databases, digital collections, and social media. Each source has its own strengths and challenges when it comes to downloading text.
Activity: Assess different textual data sources	• In small groups, discuss strengths and weaknesses of different sources of textual data. • *Goal:* Practice assessing benefits and drawbacks of various sources of textual data.
Evaluating textual data sources	• When assisting researchers in finding textual data, also consider how much flexibility is needed for working with the data, the technical skillset of the researcher, and any funding limitations.
Introduction to worksets	• HTRC Worksets are one way the HathiTrust allows users to create text corpora to analyze. • A workset is a user-created collection of HTDL text and can be cited and shared. Viewed on HTRC Analytics, you'll get metadata about the volumes in the workset but will not be able to read the text in this interface, so it suits non-consumptive research.

	• Users can import worksets from the HT Collection Builder, or compile volume IDs elsewhere.
Activity: Create and import a workset into HTRC Analytics	• Participants work alone or in pairs to create worksets • Encourage attendees to curate the volumes they select for their collection • Whole-group discussion of process when finished • *Goal:* Gain experience using a particular digital library interface to build a text analysis corpus.
Creativity Boom case study	• Introduce how Sam built his corpora for analysis
Discussion	• What expertise do librarians already have to help with building a corpus for textual analysis? • *Goals:* Encourage learners to tie their existing professional knowledge to skills that are useful for building textual datasets.

Additional Tips for Instructors

- **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.
- **Make sure to log in to HTDL before creating a collection in HT, and to log in to HTRC Analytics before uploading the collection.**
- 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.

Module 2.2 Gathering Textual Data: Bulk retrieval

This lesson covers methods for gathering textual data from the web in bulk, including using APIs, file transfers, and web scraping, and also introduces the command line interface.

Estimated time

45-60 minutes

Audience

Librarians with some exposure to text analysis who may be supporting text analysis research at their institutions.

Prerequisites for participants

- Have some idea of text analysis concepts
- Have been introduced to the HTRC, or have completed Module 1
- Have been introduced to the concept of text as data in digital scholarship and are familiar with the options available to researchers for accessing textual data, or have competed Module 2.1

Learning objectives

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

- Execute basic commands from the command line interface in order to gain confidence with computationally-intensive research.
- Understand why automated access is valuable for building textual datasets in order to facilitate researcher needs around digital scholarship.

Skills

- Command line
- Execute a web scraping command

Getting ready

Workshop participants will need:

- Access to a computer, the Internet, and a web browser
- Access to PythonAnywhere and an account

Session outline

- Introduction to bulk retrieval and bulk HTRC data
- Introduction to methods of automating bulk retrieval
    - Web scraping
    - APIs
    - Transferring files
- **Activity:** Explore the basic HathiTrust Bibliographic API
- Introduction to the command line
- **Activity:** Run basic Bash commands
- **Activity:** Scrape a webpage
- Creativity Boom case study: How Sam did bulk HTRC data retrieval
- Discussion: Does your library provide access to digitized materials in a way that is conducive to text analysis?

Key concepts

- **Command line**: A text-based interface that takes in commands and passes them to the computer's operating system. Commands can be used to accomplish (and script) a wide range of tasks. The interface is often called a **shell**, such as the **Bash shell**.
- **API (Application Programming Interface)**: A set of clearly-defined communication methods (may include commands, functions, protocols, objects, etc.) that can be used to interact with an external system. They are basically instructions (written in code) for accessing systems or collections.
- **Script**: A file containing a set of programing statements that can be run using the command line.
- **Web scraping:** The process of extracting data from webpages.

Key tools

- **File Transfer Protocol (FTP):** A protocol that computers on a network use to transfer files to and from each other. A protocol is a set of rules that networked computers use to talk to one another, like a language.

- **Secure/SSH File Transfer Protocol (SFTP):** Works in a way similar to FTP, but is a separate protocol that encrypts the connection to enable a secure file transfer.
- **rsync:** A fast file-copying tool widely used for backups. It's well-known for its efficiency, because it reduces the amount of data sent over the network by sending only the differences between the files at the source location and the files at the destination location.
- **PythonAnywhere:** A browser-based programming environment that's also a code editor and file hosting service. It comes with a built-in Bash shell and does not interact with your local file system.
- **wget:** A command line tool for retrieving files from a server. It can scrape the contents of a website, with options that can be modified to tailor more specifically to how you want the contents to be retrieved.
- **Beautiful Soup:** A Python-based web scraping tool that pulls data out of HTML and XML files. It has several options for specifying what you want to scrape (within the HTML) and is good for getting clean, well-structured text.

Key points

Introduction to bulk retrieval and bulk HTRC data	• Gathering large amounts of textual data is a time-consuming process – it's necessary to automate retrieval when possible. • Some HT and HTRC datasets can be retrieved using APIs and rsync.
Introduction to methods of automating bulk retrieval	• Some methods for automating retrieval are: web scraping using tools or via running commands/scripts; using APIs; transferring files with FTP, SFTP, or rsync.
Activity: Use an API	• Retrieve metadata using the HathiTrust's Bibliograpic API. • *Goal:* Demystify data APIs to show how they facilitate data transfer.
Introduction to the command line	• The command line is a text-based interface that takes in commands and passes them on to the computer's operating system to accomplish tasks.

	• You can use a web-based tool called PythonAnywhere with a built-in Bash shell to run commands and scripts.
Activity: Run basic Bash commands	• Use video to introduce some basic Bash commands, such as “pwd” and “cd”, and guide participants in practicing them in PythonAnywhere. Participants will also unzip and move the activity files that will be used in later activities. • *Goal:* Gain hands-on experience with the command line in preparation for the following activity.
Activity: Run wget to scrape a webpage	• Guide participants in running a command on PythonAnywhere that scrapes the text from a webpage version of George Washington’s *Fourth State of the Union Address*. • Review the scraped text, summarize the process, and discuss next steps. • On their own, participants revise the command to scrape George Washington’s *Second State of the Union Address.* • *Goal:* Build confidence on the command line and show how automated data retrieval makes it easier to grab data than manual copying.
Creativity Boom case study	• Sam used rsync to bulk retrieve HTRC Extracted Features files.
Discussion	• Question: Does your library provide access to digitized materials in a way that is conducive to text analysis? • *Goal:* Prompt librarians to consider how library collections are data sources for text analysis.

Additional Tips for Instructors

- **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.

- When demonstrating the commands in PythonAnywhere, instructors may use “Ctrl” and “+” (“Command” and “+” on Macs) to enlarge the content on the screen. It can be very difficult to see the command line 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.
- It could be helpful to have at least two instructors teaching this module, with one demonstrating commands and running scripts in the front, and the another moving around the room to help participants troubleshoot any issues.

Module 3: Working with Textual Data

In order to do text analysis, a researcher needs some proficiency in wrangling and cleaning textual data. This module addresses the skills needed to prepare text for analysis after it has been acquired.

Estimated time

50-65 minutes

Audience

Librarians who want to learn more about data preparation for working with text in particular.

Prerequisites for participants

Ideally, participants:

- Are familiar with concept of text as data
- Have been introduced to the HTRC, or have completed Module 1
- Have used the command line, or have completed Module 2 Lesson 2

Learning goals

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

- Distinguish cleaning and preparing as one step in the text analysis workflow in order to understand best practice in the field.
- Recognize key strategies for preparing data in order to make recommendations to researchers.
- Run a Python script from the command line in order to gain experience with the utility of Python for working with data.

Skills

Upon completion of the module, participants should be able to obtain the following skills:

Execute data cleaning methods, such as:

- Running a Python script to remove HTML tags from text scraped from the web.
- Experimenting with stop word lists.

Getting ready

Workshop participants will need:

- Access to a computer, the Internet, and a web browser
- Access to PythonAnywhere
- The following data downloaded and uploaded to PythonAnywhere:
    - The scraped text file of *Fourth State of the Union Address* from Module 2.2
    - remove_tag.py
    - remove_stopwords.py

Session outline

- Introduction to humanities data
- Approaching text as data
- Overview of preparing data before conducting text analysis
    - Common steps
    - Key concepts: chunking and grouping text, tokenization
    - Preparation impacts results and takes time and effort
- **Activity:** Read and review data cleaning steps
- Introduction to Python
- **Activity:** Run Python scripts to strip HTML tags from a text file and to remove stop words
- Creativity Boom case study: how Sam refined his corpus in preparation for analysis
- Discussion: read and reflect on passage from “Against Cleaning” by Katie Rawson and Trevor Munoz

Key concepts

- **Humanities data:** In a humanities research setting, “data” can be defined as material generated or collected while conducting research. Humanities data may include databases, citations, software code, algorithms, documents, etc. (Adapted from definition provided in *Data Management Plans for NEH Office of Digital Humanities Proposals and Awards*)
- **Script:** A file containing a set of programing statements that can be run using the command line. Python scripts are saved as files ending with the extension “.py”.
- **Chunking text:** The process of splitting text into smaller pieces before analysis. May be divided by paragraph, chapter, or a chosen number of words.
- **Grouping text:** The process of combining text into larger pieces before analysis.
- **Stop words:** Frequently used words (such as “the”, “and”, “if”) that are often removed from text before performing analysis.

- **Tokenization:** Breaking text into pieces called tokens. Often certain characters, such as punctuation marks, are discarded in the process.
- **N-grams**: A contiguous chain of n items from a sequence of text where n is the number of items. Unigrams refer to one item chains, bigrams to two item chains, and so on.

Key tools/platforms

- **Python:** A programming language that is commonly used when working with data. Python has high-level data structures, is interpretive in nature, and has a relatively simply syntax.
- **OpenRefine:** A tool like Excel that is powerful for exploring, cleaning, and manipulating tabular data. Originally known as Freebase Gridworks and later as Google Refine, OpenRefine became an open community resource in 2012.

Key points

Introduction to humanities data	• In a humanities research setting, data is material generated or collected while conducting research. • Humanities data may include citations, software code, algorithms, etc.
Approaching text as data	• Text can be approached as data and analyzed by corpus/corpora. • Before analyzing textual data, it is important to ensure the text is of sufficient quality (e.g., OCR-ed data is cleaned up) and fully prepared (certain unnecessary elements are discarded).
Overview of preparing data before conducting text analysis	• Common steps include: correcting OCR; removing titles or headers; removing html or xml tags; splitting (chunking) or combining (grouping) files; removing certain words, punctuation marks; making text lowercase, tokenization, stemming • Preparation impacts research results and takes time and effort. When possible, these tasks should be automated, and scripting is a helpful way to do this clean up.
Activity: read and review data cleaning steps	• In small groups, read and explain to one another concepts in data cleaning.

	• *Goal:* Reinforce the variety of data preparation strategies a researcher may use to clean text data.
Introduction to Python	• Python is a programming language that's very useful for working with data. It has high-level data structures, is interpretive in nature, and has a relatively simply syntax. • Python can be used to write and run scripts, or it can be used for doing interactive programming.
Activity: Run Python scripts to clean text data	• Using PythonAnywhere, instructors will guide participants in running a Python script to remove HTML tags from a scraped text, and reviewing results. • Using PythonAnywhere, participants will execute a Python script on their own to remove stop words. • *Goal:* Practice basic data cleaning techniques to understand how data is readied for text analysis.
Creativity Boom case study	• Sam removed all of the pages in his workset that did not contain a form of a creativ* • He also discarded all words that belonged to a certain category of part of speech called "closed" parts of speech, which are pronouns, conjunctions, and other words.
Discussion	• "Against Cleaning" is a piece by Katie Rawson and Trevor Munoz that proposes as humanities-centric approach to data standardization. • Participants will read the passage on the slide and reflect on it via discussion. • *Goal:* Encourage participants to consider what is lost (or gained) when data is standardized.

Additional Tips for Instructors

- **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.
- 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.

Module 4.1 Performing Text Analysis: Using Off-the-shelf Tools

In this lesson, we will be focusing on supporting beginner researchers in performing text analysis by using off-the-shelf, pre-built tools. It will discuss the advantages and constraints of web-based text analysis tools and programming solutions, introduce basic text analysis algorithms available in the HTRC algorithms, and demonstrate how to select, run, and view the results of the topic modeling algorithm.

Estimated time

45-60 minutes

Audience

Librarians with minimal experience with digital humanities, or who will be working with others with limited experience.

Prerequisites for participants

Ideally, participants:

- Are familiar with concepts of how to acquire and manage text data, or have completed Module 2
- Have been introduced to the HTRC, or have completed Module 1

Learning goals

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

- Recognize the advantages and constraints of web-based text analysis tools and programming solutions in order to evaluate researcher questions and requests.
- Match appropriate tools to research problems, and distinguish different approaches to text analysis in order to suggest options for researchers.
- Demonstrate text analysis using web-based tools in order to gain experience with off-the-shelf solutions text mining.
- Evaluate the results of running a text analysis algorithm in order to build confidence with the outcomes of data-intensive research.

Skills

Upon completion of the lesson, participants should be able to obtain the following skills:

- Run a text analysis algorithm

Getting ready

Workshop participants will need:

- An account for HTRC Analytics (https://analytics.hathitrust.org)

Session outline

- Introduction to tools for performing text analysis
    - Benefits and drawbacks of pre-built tools and do-it-yourself tools
    - Choosing a pre-built tool
- Introduction to the HTRC algorithms: features of the off-the-shelf algorithms, how to choose an algorithm
- Introduction to topic modeling: bag-of words model, how does topic modeling work, tips for topic modeling
- **Activity:** Think about what kinds of research questions certain HTRC algorithms can help answer.
- **Activity:** Run topic modeling algorithm in HTRC Analytics
- Creativity Boom case study: how Sam experimented with HTRC Algorithms to explore his corpus
- Discussion: How are librarians teaching digital scholarship tools to students and researchers?

Key concepts

- **Algorithm:** A process a computer follows to solve a problem, creating an output from a provided input.
- **Topic modeling**: A method of using statistical models for discovering the abstract "topics" that occur in a collection of documents.
- **Bag-of-words model:** A concept for working with text where all grammar and word order has been taken out and all the words are like being mixed up in a bag.
- **Job (in HTRC context)***:* An algorithm run against a workset in HTRC Analytics.
- **Results (in HTRC context)***:* The results of your job(s) outputted by the algorithm. You can view or download them.

Key tools

- **HTRC algorithms:** A set of off-the-shelf text analysis algorithms provided via HTRC Analytics for users to analyze their worksets, such as algorithms for extracting named entities and doing topic modeling.
- **Voyant:** A tool that can create many types of visualizations such as word clouds, bubble charts, networks, and word trees. It has a user-friendly interface that works great as a learning tool. See more at: http://voyant-tools.org/
- **Lexos:** A web-based tool that can be used for pre-processing, analysis, and visualization of digitized texts. Lexos can also be downloaded and installed locally. See more at: http://lexos.wheatoncollege.edu/upload
- **AntConc:** A freeware corpus analysis toolkit for text analysis, especially for analyzing concordances. See more at: http://www.laurenceanthony.net/software/antconc/
- **Weka:** A collection of machine learning algorithms for data mining tasks. It contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. See more at: http://www.cs.waikato.ac.nz/ml/weka/

Key points

Introduction to tools for performing text analysis	• There are pre-built tools and do-it-yourself tools for performing text analysis. Pre-built tools are easy to use but have limited capacities. Do-it-yourself tools allow for more customization and control but requires more technical knowledge. • How to choose a pre-built tool depends on the goal of the analysis. Some tools are better than others at conducting certain types of analysis.
Introduction to the HTRC algorithms	• HTRC algorithms are pre-built tools that can extract, refine, analyze, and visualize worksets. They are limited in parametrization but good for learning. • Different HTRC algorithms accomplish different types of tasks. Some are task oriented, while others are more analytic.

Introduction to topic modeling	• Topic modeling is a method of using statistical models for discovering the abstract "topics" that occur in a collection of documents. • In topic modeling, the text is chunked, stop words are removed, and the computer treats texts as bags of words, and guesses which words make up a “topic” based on their proximity to one another. • “Topics” aren’t necessary true reflections of aboutness – tweaking your input affects the output. • When doing topic modeling, treat it as one part of a larger analysis, be familiar with your input text and check your results, and be aware of how changing stop word lists and tweaking parameters can affect results. Additionally, gain some basic knowledge about your tool.
Activity: Discuss research applications for web-based text analysis tools	• Think about what kinds of research questions certain HTRC algorithms can help answer. • *Goal:* Gain confidence pairing research question to tool.
Activity: Run topic modeling algorithm in HTRC Analytics	• Instructors will guide participants in running the HTRC topic modeling algorithm to see what topics are present in a sample workset of political speech texts. • *Goal:* Develop hands-on experience with text analysis algorithms.

Creativity Boom case study	• Think about how Sam could have used HTRC Algorithms to explore his corpus
Discussion	• How are librarians teaching digital scholarship tools to students and researchers? • *Goal:* Encourage attendees to map concepts they learn in the workshop to teaching and learning in their library.

Additional Tips for Instructors

- **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.
- It is helpful to run the topic modeling job exactly as described in the activity in advance to make sure you have a completed job to show to the participants during the workshop, just in case your live demonstration of the job gets stuck in the queue and cannot be completed in time.
- 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.

Module 4.2 Performing Text Analysis: Basic Approaches with Python

More advanced researchers will prefer to conduct text analysis outside of pre-built, off-the-shelf tools, opting instead for a toolkit of command line programs and custom code. This module introduces the concept of programming packages and provides hands-on experience with running Python code to analyze an Extracted Features file from the HTRC Extracted Features dataset.

Estimated time

50-65 minutes

Workshop audience

Librarians who want to develop their skillset for supporting researchers who want to engage in computational text analysis.

Learning goals

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

- Identify the needs of advanced text mining researchers in order to make skill-appropriate recommendations.
- Recognize text analysis methods in order to understand the kinds of research available in the field.
- Successfully interact with a pre-defined textual dataset in order to gain experience with programming skills for data-driven research.

Skills

Upon completion of the module, participants should be able to obtain the following skills:

- Install a Python library using Pip
- Run a Python script to work with an HTRC Extracted Features file

Prerequisites for participants

Ideally, participants:

- Have been introduced to the HTRC, or have completed Module 1
- Have used the command line, or have completed Module 2 Lesson 2

- Are acquainted with the programming language Python or, or have completed Module 3

Session outline

- Introduction to toolkit for do-it-yourself text analysis
- Overview of package managers and installing libraries/packages
- Introduction to HTRC Extracted Features
- **Activity:** Install a Python library and run a script to view most-used adjectives in a set of volumes
- Introduction to exploratory data analysis
- **Activity:** Install the HTRC Feature Reader and run Python script to view the word count in a volume based on its Extracted Features file
- Advanced text analysis with the HTRC Extracted Features example
- Discussion of the librarian's role in supporting text analysis research

Getting ready

Workshop participants will need:

- Access to a computer, the Internet, and a web browser
- Access to PythonAnywhere
- The following files in PythonAnywhere:
    - top_adjectives.py
    - word_count.py
    - mdp.49015002221860.json.bz2
    - mdp.49015002221878.json.bz2
    - mdp.49015002221886.json.bz2
    - miua.4925052,1928,001.json.bz2
    - miua.4925383,1934,001.json.bz2
    - mdp.49015002203033.json.bz2
    - mdp.49015002203140.json.bz2
    - mdp.49015002203157.json.bz2
    - mdp.49015002203215.json.bz2
    - mdp.49015002203223.json.bz2
    - mdp.49015002203231.json.bz2
    - mdp.49015002203249.json.bz2

    - mdp.49015002203272.json.bz2
    - mdp.49015002203405.json.bz2
    - mdp.49015002221761.json.bz2
    - mdp.49015002221779.json.bz2
    - mdp.49015002221787.json.bz2
    - mdp.49015002221811.json.bz2
    - mdp.49015002221829.json.bz2
    - mdp.49015002221837.json.bz2
    - mdp.49015002221845.json.bz2
- HTRC Feature Reader Python library installed to PythonAnywhere

Key concepts

- **Natural Language Processing (NLP):** Using computers to understand the meaning, relationships, and semantics within human-language text.
- **Named entity extraction:** Using computers to locate and classify named entities (such as the names of persons, organizations, and locations) in text.
- **Stylometry:** The application of the study of linguistic style. It is often used to determine authorship to anonymous or disputed texts.
- **Sentiment analysis:** Using computers to systematically identify attitudes or emotions present in text.
- **Machine learning:** A process that gives computers the ability to learn without being explicitly programmed. Machine learning is based on researchers constructing and using algorithms that can learn from and make predictions on data. It can either be unsupervised (with minimal human intervention) or supervised (with more human intervention).
- **Topic modeling**: A method of using statistical models for discovering the abstract "topics" that occur in a collection of documents.
- **Naïve Bayes classification:** A method based on Bayes' Theorem from statistics that uses machine learning to classify texts based on information present in the texts of each class.
- **Functions:** Reusable code blocks that perform an action.
- **Libraries/packages:** Collections of functions that can be implemented in a script or program.
- **Package Manager:** A tool that facilitates the download and installation of programming packages.

- **Exploratory data analysis:** An approach for familiarizing oneself with a dataset before analyzing it that often involves visualizations, including visualizations of raw counts and simple statistics, or comparative visualizations.

Key tools/platforms

- **Python:** A programming language that is good for working with data. Python has high-level data structures, is interpretive in nature, and has a relatively simply syntax.
- **pip:** Package manager for Python (alternatives: Homebrew, Conda).
- **R:** A programming language optimized for (statistical) data analysis.
- **HTRC Extracted Features:** A downloadable dataset of text data and metadata extracted and abstracted from volumes in the HathiTrust Digital Library.
- **HTRC Feature Reader:** Python library for working with HTRC Extracted Features.
- **pyplot:** Visualization function in the Python data science package, Pandas.

Key points

Key approaches to text analysis	• Among others, there are 2 key approaches to text analysis: natural language processing and machine learning • Natural language processing is the use of computers to understand the meaning, relationships, and semantics within human-language text. It includes named entity extraction, sentiment analysis, and stylometry. In many, but not all, cases, the researcher will require full text. • Machine learning is training computers to recognize patterns in text, and it can be supervised or unsupervised. It includes topic modeling and Naïve Bayes classification.
Activity: match project to method	• Participants match each of the research examples from Module 1 with a broad text analysis area and specific method. • *Goal:* Reinforce understanding the kinds of research questions that particular text analysis methods are suited to answer.
HTRC Extracted Features dataset	• A dataset of JSON files, one for each volume in the HTDL

	• The files contain metadata, including bibliographic metadata and computationally-derived metadata, such as word and line counts • They also include part-of-speech tagged token counts at the page-level
Do-it-yourself text analysis	• Some researchers will not be satisfied with pre-built, off-the-shelf tools. • They will want more control over the process via do-it-yourself tools
The text analysis toolkit	• The toolkit more advanced researchers will use depends on individual preferences • The researcher will likely need an understanding of statistics, and they may collaborate with other experts • The toolkit will consist of command line tools and programming languages • MALLET and Stanford NLP are common command line tools for text analysis • R and Python are common programming languages for text analysis
Programming concepts of modules, packages, and libraries	• Programming packages and libraries are collections of reusable code blocks; Packages are made up of modules • Packages for text analysis may facilitate tasks such preparing, reading or loading, and analyzing text with preset routines. • Packages are installed using a “package manager” which are command line tools that help make sure the packages are installed correctly
Activity: Install a Python library and run a script to view most-used adjectives in a set of volumes	• Using PythonAnywhere, instructors will guide participants through the process of installing the HTRC Feature Reader Python library and run a Python script to create a list of the most-used adjectives and the number of times they occur in a set of volumes in a workset.

	• *Goal:* Gain exposure to programming concepts, understand how counts of features can reveal information about text, practice basic text analysis.
Exploratory data analysis	• It is often difficult to grasp the contents of a dataset—its scope, range, and potential errors—from reading files alone. • Exploratory data analysis is the process by which one familiarizes themselves with a dataset before analysis • Often exploration involves visualization to make it easier to understand the data.
Activity: Visualize word count in an HTRC Extracted Features file	• Using a Python script, plot raw counts in an HTRC Extracted Features file • Visualize word count over a single volume • *Goal:* Develop comfortability with how basic text analysis can be aided by graphing data.
Advanced text analysis example	• Ted Underwood completed a text analysis project that used the HTRC Extracted Features dataset to classify volumes in the HTRC by genre • This work is an example of what can be done using the data fields in the Extracted Features and also of supervised machine learning • Ted released his derived dataset at the end of the project and it's available for others to use in their own analysis projects
Creativity Boom case study	• On his limited corpus of only pages containing the forms of "creativ*", Sam performed topic modeling • That way he ended up with the themes around the concept of creativity in the literature. • He then mapped the topics over time to see how their usage changed through the twentieth century.

Discussion	• In what ways can librarians support advanced text analysis research? • What additional skills would you need to learn in order to do so? • *Goal:* Encourage librarians to consider how they might apply what they have learned in the workshop.

Additional Tips for Instructors

- **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.
- 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.

Module 5 Visualizing Textual Data: An Introduction

This lesson is an introduction to data visualization in general, with a focus on textual data analysis. It also introduces the HathiTrust+Bookworm interface that allows the user to visualize word usage over time.

Estimated time

30-45 minutes

Workshop audience

- Beginners with an interest in text analytics and/or the HTRC more generally
- Anyone interested in data visualization, especially the visualization of textual data
- Anyone interested in learning about basic tools for interacting with the HTDL corpus

Learning goals

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

- Recognize common types of data visualizations in order to communicate with researchers about their options.
- Explore results in HathiTrust+Bookworm and begin making connections using available data and data points in order to develop experience reading data visualizations.

Skills

- Using library metadata to impact how a visualization is displayed
- Reading and interpreting graphs
- Perform a keyword search
- Fine-tune search results through faceting

Prerequisites for participants

None! While Module 1, Getting Started, provides useful background about the HTRC and its mission, learners can dive into HathiTrust+Bookworm without much introduction.

Session outline

- What is data visualization and when is it used in the research process?
- Common types of textual data visualizations
- **Activity:** Match type of use to the type of visualization

- Examples of web-based tools and programming libraries for visualizing textual data
- Introduction to HathiTrust+Bookworm:
    - What is HathiTrust+Bookworm?
    - Examples of HathiTrust+Bookworm visualizations
    - Overview of HathiTrust+Bookworm interface
- **Activity:** Hands-on exploration of HathiTrust+Bookworm
- Case study: How Sam visualized his data
- Discussion: Visual literacy and data literacy

Getting ready

Workshop participants will need:

- Access to a computer, the Internet, and a web browser.

Key concepts

- **Data visualization:** The process of converting data sources into a visual representation. It often also refers to the product of this process.
- **Word tree**: A type of visualization that displays the different contexts in which a word or phrase appears in a text, with the contexts arranged in a tree-like structure to reveal recurrent themes and phrases.
- **Node-link diagram:** A type of visualization for displaying networks. It captures entities (such as people, places, and topics) as nodes (also called “vertices”) and relationships as links (also called “edges”), with a circle or dot representing a node, and a line representing a link.
- **Word cloud/tag:** A graphical representation of word frequency, usually presenting words that appear more frequently in the source text larger than those that appear less frequently.
- **N-grams**: A contiguous chain of n items from a sequence of text where n is the number of items. Unigrams refer to one item chains, bigrams to two item chains, and so on.
- **Timeline:** A graphic design displaying events in chronological order.

Key tools

- **HathiTrust + Bookworm:** A tool that visualizes word frequencies over time in the HathiTrust Digital Library. It can be accessed at: https://bookworm.htrc.illinois.edu/develop .
- **Google Books Ngram Viewer:** Similar to HathiTrust+Bookworm, a tool that enables users to search for words in corpora of texts and visualize their usage over time. Link: https://books.google.com/ngrams

- **Voyant:** A tool that can create many types of visualizations including word clouds, bubble charts, networks, word trees, etc. It has a user-friendly interface that works great as a learning tool. Link: http://voyant-tools.org/
- **Wordle:** A tool for creating word clouds, mostly for exploration and decorative purposes because not much fine-tuning can be done. Link: http://www.wordle.net
- **ArcGIS Online/StoryMaps**: A visualization tool that can be used to incorporate GIS information and maps into interactive timelines and stories. Link: https://storymaps.arcgis.com/en/
- **Tableau:** A set of software that can be used for data preparation, visualization, and analysis. Among the different versions of Tableau Desktop (geared towards individual usage), Tableau Public is available for free. See more at: https://public.tableau.com/s/ and https://www.tableau.com
- **Gephi:** A free visualization and exploration software that can be used to create graphs and networks. It works especially well for exploratory data analysis. See more at: https://gephi.org
- **NodeXL:** An add-in for Microsoft Excel that supports social network and content analysis. Available in Basic and Pro versions. See more at: http://www.smrfoundation.org/nodexl/
- **DH Press:** A digital humanities toolkit that enables users to mashup and visualize a variety of digitized humanities-related material, including historical maps, images, manuscripts, and multimedia content. It can be used to create a range of digital projects and is designed for non-technical users. See more at: http://dhpress.org
- **ggplot:** Python library for data visualization.
- **pyplot:** Visualization function in the Python data science package, Pandas.
- **ggplot2:** R library for data visualization.
- **D3.js:** JavaScript library for web-publishable visualizations.

Key points

What is data visualization?	• Data visualization is the process of converting data sources into a visual representation. • Visualization is a way of interpreting and presenting data.

Common textual data visualizations	• Some common visualizations include: word clouds, trees/hierarchies, networks, temporal/spatial-based visualizations, and other "multi-dimensional" visualizations.
Activity: Match type of use to the type of visualization	• Participants match types of visualizations to the kinds of information they are suited to convey. If time allows, consider the kind of data each visualization might require. • *Goal:* Practice thinking about applications for data visualization, and when and with what data they might be employed by researchers.
Examples of web-based tools and programming libraries for visualizing textual data	• Examples of web-based tools include: Voyant, Wordle, ArcGIS Online/StoryMaps, Google Books Ngram Viewer, HathiTrust+Bookworm, Tableau, Gephi, NodeXL, DH Press • Programming libraries for visualizations: matplotlib, pyplot, and ggplot library in Python; ggplot2 in R; D3.js.
What is HathiTrust+Bookworm?	• Bookworm is a tool that visualizes language usage trends in repositories of digitized texts. It is good at finding and understanding categories in a library. • Bookworm can visualize and quantify the dynamics of language evolution. • HathiTrust + Bookworm is a visualization of word frequencies over time in the HathiTrust Digital Library.
Examples of HathiTrust+Bookworm visualizations	• Using HT+BW to track social change: "lady" vs. "woman" • Using HT+BW to Bookworm to track words in translation across time and place: "liberté" and "liberty"

Overview of HathiTrust+Bookworm interface	• Type in search words and click on the funnel icon to facet the search by genre, language, and more. • Use the tabs “Dates”, “Metric”, and “Case” to fine-tune results. • After the visualization is generated, click on a specific spot on the curve to be directed to corresponding volumes in the HathiTrust Digital Library.
Activity: Hands-on exploration of HathiTrust+Bookworm	• Guide participants in using HT+BW to visualize lexical trends. • *Goal:* Gain experience using web-based visualization tools, the parameters that can be adjusted, and the information they convey.
Case Study	• Sam used HT+Bookworm to visualize the use of “creative” in the HTDL over time • Sam also used an experimental HT+BW interface to create different kinds of visualizations
Discussion	• Where does visual literacy fit into data literacy overall? • What would it mean to be visually literate, particularly with regard to text analysis? • *Goal:* Encourage librarians to consider pedagogical applications for concepts they have learned.

Additional Tips for Instructors

- **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.
- 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.

- For the HT+BW hands-on activity, instructors may encourage workshop participants to discuss their search results with each other. This can make the activity more interactive and keep the participants more fully engaged.
- Data visualization is a huge topic, and the information provided in this lesson can only scratch the surface. For instructors who have little previous experience in this area, it may be helpful to do some additional background reading (the materials provided in the further reading section of our website is a good place to start) to familiarize themselves with other types and formats of data visualization and more visualization tools.