Digging Deeper Reaching Further

Libraries Empowering Users to Mine the HathiTrust Digital Library Resources

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

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

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.