# Digging Deeper Reaching Further

Libraries Empowering Users to Mine the HathiTrust Digital Library Resources

## Overview of Activities

## **Module 1** Introduction

### Activity 1: Read and explain text analysis examples

- **Description:** Participants read and review three summarized text analysis research examples and discuss the key points, kinds of data, and findings of the projects.
- **Goal:** Gain exposure to text analysis research and how it is being used by scholars.
- **Slides:** M1-8 to M1-16
- **Handout:** Master Handout p. 3 (Module 1 Handout p. 1)
- **Instructor Guide:** Full Guide pp. 8-11 (Module 1 Instructor Guide pp. 3-5)
- **Requirements:**
    - Files: None
    - Other: Access to a computer, the Internet, and a Web browser for reading project summaries online: http://go.illinois.edu/ddrf-research-examples

---

## **Module 2.1** Gathering Textual Data: Finding Text

### Activity 2.1-1: Assess different textual data sources

- **Description:** Participants discuss the strengths and weaknesses of three kinds of textual data sources for building a corpus for political history.

- **Goal:** Practice assessing benefits and drawbacks of various sources of textual data.
- **Slides:** M2.1-9

  **Note: Also see slides M2.1-6 to M2.1-8 for an overview of three sources for textual data.*
- **Handout:** Master Handout p. 4 (Module 2.1 Handout p. 1)
- **Instructor Guide:** Full Guide p. 8 (Module 2.1 Instructor Guide p. 4)
- **Requirements:**
  - Files: None
  - Other: None

### Activity 2.1-2: Create and import a workset into HTRC Analytics

- **Description:** Participants create a textual dataset of volumes related to political speech in America with the HT Collection Builder, and upload it to HTRC Analytics as a workset for analysis.
- **Goal:** Gain experience using a particular digital library interface to build a text analysis corpus.
- **Slides:** M2.1-17 to M2.1-38
- **Handout:** Master Handout pp. 5-7 (Module 2.1 Handout pp. 2-4)
- **Instructor Guide**: Full Guide pp. 11-15 (Module 2.1 Instructor Guide pp. 7-11)
- **Requirements:**
  - Files: None
  - Other: Access to a computer, the Internet, and a Web browser (Internet Explorer is *not recommended* for this activity) for using the following websites: HathiTrust Digital Library (HTDL) and HathiTrust Research Center (HTRC) Analytics ; HTDL account and HTRC account

---

# **Module 2.2** Gathering Textual Data: Bulk Retrieval

## Activity 2.2-1: Explore the HathiTrust Bibliographic API

- **Description:** Participants retrieve metadata of a HathiTrust Digital Library volume of their choice using the HathiTrust Bibliographic API.
- **Goal:** Demystify data APIs to show how they facilitate data transfer.
- **Slides:** M2.2-10 to M2.2-11

  **Note: Also see slides M2.2-8 to M2.2-9 for background information about APIs and the HathiTrust Bibliographic API.*
- **Handout:** Master Handout p. 8 (Module 2.2 Handout p. 1)
- **Instructor Guide:** Full Guide pp. 8-9 (Module 2.2 Instructor Guide pp. 4-5)
- **Requirements:**
  - Files: None
  - Other: Access to a computer, the Internet, and a Web browser

## Activity 2.2-2: Run basic Bash commands

- **Description:** Participants watch a video that introduces some basic Bash commands, such as “pwd” and “cd”, and practice using them in a Bash console. Participants also unzip and move the activity files that will be used in activities in later modules (this part can be skipped when using this activity independently in other instruction contexts).
- **Goal:** Gain hands-on experience with the command line.
- **Slides:** M2.2-16 to M2.2-20

  **Note: Also see slides M2.2-12 to M2.2-15 for background information about the command line, Bash, and PythonAnywhere.*
- **Handout:** Master Handout p. 9 (Module 2.2 Handout p. 2)
- **Instructor Guide:** Full Guide pp. 11-12 (Module 2.2 Instructor Guide pp. 6-8)
- **Requirements:**
  - Files: activity_files.zip

**Note: No files are needed if using this activity independently in other instruction contexts.*

- Other:
  - Access to a computer, the Internet, and a Web browser for accessing video at http://go.illinois.edu/ddrf-bash-video
  - Web browser and PythonAnywhere account for using web-based Bash console on PythonAnywhere. You can also use Terminal (Mac systems) or other configured Bash shells (Windows systems) instead.

### Activity 2.2-3: Run the wget command to scrape a webpage

- **Description:** Participants run the wget command on PythonAnywhere to scrape text from a webpage version of George Washington's *Fourth State of the Union Address*, and revise the command to scrape George Washington's *Second State of the Union Address.*
- **Goal:** Gain experience with the command line and show how automated data retrieval makes it easier to grab data than manual copying.
- **Slides:** M2.2-23 to M2.2-31

  **Note: Also see slide M2.2-21 for background on wget and slide M2.2-32 for other options for scraping online text.*
- **Handout:** Master Handout pp. 10-11 (Module 2.2 Handout pp. 3-4)
- **Instructor Guide:** Full Guide pp. 13-15 (Module 2.2 Instructor Guide pp. 9-11)
- **Requirements:**
  - Files: None.
  - Other: Access to a computer, the Internet, and a Web browser (Internet Explorer is *not recommended* for this activity) for using web-based Bash console on PythonAnywhere; PythonAnywhere account

### Activity 2.2-4: Read and reflect: *Santa Barbara Statement on Collections as Data*

- **Description:** Participants read some parts of the *Santa Barbara Statement on Collections as Data* and discuss the development of digital collections in libraries.

- **Goal:** Gain a basic understanding of the idea of “collections as data” and how it relates to library digital collections development.
- **Slides:** M2.2-36 to M2.2-38
- **Handout:** Master Handout p. 11 (Module 2.2 Handout p. 4)
- **Instructor Guide:** Full Guide p. 17  (Module 2.2 Instructor Guide p. 13)
- **Requirements:**
    - Files: None.
    - Other: Access to a computer, the Internet, and a Web browser to read full statement online: https://collectionsasdata.github.io/statement/

---

## **Module 3** Working with Textual Data

### Activity 3-1: Read and review data preparation techniques

- **Description:** Participants read and explain to one another concepts and techniques in data preparation in pairs or small groups.
- **Goal:** Reinforce the variety of data preparation strategies a researcher may use to clean text data.
- **Slides:** M3-11

    **Note: Also see slides M3-3 to M3-10 for more information on preparing data for analysis and key concepts.*
- **Handout:** Master Handout pp. 12-13 (Module 3 Handout pp. 1-2)
- **Instructor Guide:** Full Guide pp. 8-10 (Module 3 Instructor Guide pp. 4-6)
- **Requirements:**
    - Files: None.
    - Other: The handout is required for this activity (reading material is included only in handout and not on slides).

### Activity 3-2: Run Python scripts to clean text data

- **Description:** Participants run a Python script under guidance to remove HTML tags from a scraped text and review results, and 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.
- **Slides:** M3-17 to M3-25

  **Note: Also see slides M3-12 to M3-15 for background on Python and running scripts.*
- **Handout:** Master Handout p. 14 (Module 3 Handout p. 3)
- **Instructor Guide:** Full Guide pp. 11-13 (Module 3 Instructor Guide pp. 7-9)
- **Requirements:**
  - Files:
    - Washington_4.txt uploaded to PythonAnywhere

      **Note: If Activity 2.2-3 was completed before this activity, this file is already in PythonAnywhere and no additional action is required. If instructors intend to use this activity independently, they need to create this file according to instructions in Activity 2.2-3 and directly provide the file to participants to upload.*
    - Remove_tag.py, remove_stopwords.py, stopwords.txt uploaded to PythonAnywhere

      **Note: If the unzip activity files part of Activity 2.2-2 has been completed, these files are already in PythonAnywhere and no additional action is required. If the unzip activity files part of Activity 2.2-2 has NOT been completed, the three files need to be uploaded to PythonAnywhere.*
  - Other: Access to a computer, the Internet, and a Web browser (Internet Explorer is *not recommended* for this activity) for using web-based Bash console on PythonAnywhere; PythonAnywhere account.

### Activity 3-3: Read and reflect: Passage from *Against Cleaning* by Katie Rawson and Trevor Muñoz

- **Description:** Participants read an excerpt from *Against Cleaning* by Katie Rawson and Trevor Munoz and discuss in pairs or small groups what it means to "clean" data in the humanities as well as how to broach the topic with researchers.
- **Goal:** Encourage participants to consider what is lost (or gained) when data is standardized.
- **Slides:** M3-27 to M3-29
- **Handout:** Master Handout p. 15 (Module 3 Handout p. 4)
- **Instructor Guide:** Full Guide pp. 14-15 (Module 3 Instructor Guide p. 10)
- **Requirements:**
    - Files: None.
    - Other: None.

---

## **Module 4.1** Performing Text Analysis: Using Off-the-Shelf Tools

### Activity 4.1-1: Discuss research applications for HTRC algorithms

- **Description:** Participants read descriptions of HTRC algorithms and discuss in pairs or small groups what the algorithms can do and what research questions they might help answer.
- **Goal:** Gain confidence in pairing research questions to tools.
- **Slides:** M4.1-13

  **Note: Also see slides M4.1-7 to M4.1-8 for background on HTRC algorithms.*
- **Handout:** Master Handout p. 16 (Module 4.1 Handout p. 1)
- **Instructor Guide:** Full Guide p. 10 (Module 4.1 Instructor Guide pp. 5-6)
- **Requirements:**

- Files: None.
- Other: Access to a computer, the Internet, and a Web browser to read full algorithm descriptions online: https://wiki.htrc.illinois.edu/x/HoJnAQ

### Activity 4.1-2: Run topic modeling algorithm in HTRC Analytics

- **Description:** Participants run the topic modeling algorithm in HTRC Analytics to explore the most prevalent topics in a president public papers workset.
- **Goal:** Develop hands-on experience with text analysis algorithms.
- **Slides:** M4.1-15 to M4.1-29

  **Note: Also see slides M4.1-9 to M4.1-12 for key concepts in topic modeling and information about the HTRC InPhO Topic Model Explorer.*
- **Handout:** Master Handout pp. 17-18 (Module 4.1 Handout pp. 2-3)
- **Instructor Guide:** Full Guide pp. 11-15 (Module 4.1 Instructor Guide pp. 6-10)
- **Requirements:**
  - Files:
    - poli_science_DDRF@eleanordickson workset
      **Note: This is a public workset provided in HTRC Analytics so no additional action is required.*
  - Other: Access to a computer, the Internet, and a Web browser (Internet Explorer is *not recommended* for this activity) for running algorithms on HTRC Analytics; HTRC Analytics account.

---

## **Module 4.2** Performing Text Analysis: Basic Approaches with Python

### Activity 4.2-1: Match research projects to text analysis areas and methods

- **Description:** Participants match three research examples 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.

- **Slides:** M4.2-4 to M4.2-6
  **Note: The first two slides introduce the areas and specific methods, the last slide shows activity instructions.*
- **Handout:** Master Handout p. 19 (Module 4.2 Handout p. 1)
- **Instructor Guide:** Full Guide pp. 8-9 (Module 4.2 Instructor Guide p. 2)
- **Requirements:**
  - Files: None
  - Other: Access to a computer, the Internet, and a Web browser for reading project summaries: http://go.illinois.edu/ddrf-research-examples

### Activity 4.2-2: Install a Python library and run a script to view most-used adjectives in a set of volumes

- **Description:** Participants install the HTRC Feature Reader library and run a Python script to create a list of the most-used adjectives and the number of times they occur in different sets of volumes of presidential papers (using the Extracted Features files of these volumes). Participants can also modify the script to search verbs or another part of speech as an additional challenge.
- **Goal:** Gain exposure to programming concepts, understand how counts of features can reveal information about text, practice basic text analysis.
- **Slides:** M4.2-26 to M4.2-32
  **Note: Also see slides M4.2-7 to M4.2-15 for background on HTRC Extracted Features, and slides M4.2-19 to M4.2-23 for information on Python libraries/packages and the HTRC Feature Reader library.*
- **Handout:** Master Handout pp. 19-20 (Module 4.2 Handout pp. 1-2)
- **Instructor Guide:** Full Guide pp. 15-17 (Module 4.2 Instructor Guide pp. 8-10)
- **Requirements:**
  - Files:
    - All necessary files are in PythonAnywhere and no additional action is required if the unzip activity files part of Activity 2.2-2 has been completed.
      **Note: If the unzip activity files part of Activity 2.2-2 has NOT been completed, follow the directions below:*

- *Upload top_adjectives.py to your user home directory (folder) in PythonAnywhere*
- *In your home directory, create a new sub-directory (folder) called "1930" and upload these files to the directory:*
  *mdp.49015002221860.json.bz2,*
  *mdp.49015002221878.json.bz2,*
  *mdp.49015002221886.json.bz2,*
  *miua.4925052,1928,001.json.bz2,*
  *miua.4925383,1934,001.json.bz2*

- *In your home directory, create another new sub-directory (folder) called "1970" and upload these files to the directory:*
  *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*

- Other:
  - Access to a computer, the Internet, and a Web browser (Internet Explorer is *not recommended* for this activity) for using web-based Bash console on PythonAnywhere; PythonAnywhere account
  - Link to Penn Tree Bank if doing challenge: https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

### Activity 4.2-3: Visualize word count in an HTRC Extracted Features file

- **Description:** Participants create a visualization using the the native plotting functionality in Pandas, called pyplot, to see the word count over a volume based on its Extracted Features file.
- **Goal:** Develop comfortability with how basic text analysis can be aided by graphing data.
- **Slides:** M4.2-36 to M4.2-40

  **Note: Also see slides M4.2-33 to M4.2-34 for information about visualization as an exploration strategy and visualization libraries.*
- **Handout:** Master Handout p. 21 (Module 4.2 Handout p. 3)
- **Instructor Guide:** Full Guide pp. 18-19 (Module 4.2 Instructor Guide pp. 11-13)
- **Requirements:**
  - Files:
    - All necessary files are already in PythonAnywhere and no additional action is required If the unzip activity files part of Activity 2.2-2 as well as Activity 4.2-2 has been completed.

      **Note: If either Activity 2.2-2 or Activity 4.2-2 has NOT been completed, follow the directions below:*
      - *Upload word_count.py to PythonAnywhere if it is not there.*
      - *Create a new sub-directory (folder) called "1970" if you do not see one and upload file mdp.49015002203033.json.bz2 to the directory.*
      - *If the instructor is planning to have participants modify the script to look at other volumes, please also upload some additional Extracted Features files to the directory.*
  - Other: Access to a computer, the Internet, and a Web browser (Internet Explorer is *not recommended* for this activity) for using web-based Bash console on PythonAnywhere; PythonAnywhere account.

---

# **Module 5** Visualizing Textual Data: An Introduction

## Activity 5-1: Match types of use to types of visualization

- **Description:** Participants match types of visualizations to the kinds of information they are suited to convey and 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.
- **Slides:** M5-14

  **Note: Also see slides M5-7 to M5-13 for background on common types of visualizations.*
- **Handout:** Master Handout p. 22 (Module 5 Handout p. 1)
- **Instructor Guide:** Full Guide pp. 10-11 (Module 5 Instructor Guide p. 5)
- **Requirements:**
  - Files: None
  - Other: None

## Activity 5-2: Explore HathiTrust+Bookworm

- **Description:** Participants use HathiTrust+Bookworm to explore lexical trends related to their own interests.
- **Goal:** Gain experience using web-based visualization tools, the parameters that can be adjusted, and the information they convey.
- **Slides:** M5-26 to M5-29

  **Note: Also see slides M5-18 to M5-24 for how to use HathiTrust+Bookworm.*
- **Handout:** Master Handout p. 22 (Module 5 Handout p. 1)
- **Instructor Guide**: Full Guide p. 14 (Module 5 Instructor Guide p. 9)
- **Requirements:**
  - Files: None

- Other: Access to a computer, the Internet, and a Web browser (Internet Explorer is *not recommended* for this activity) for using HathiTrust+Bookworm.

---