Digging Deeper Reaching Further

Libraries Empowering Users to Mine the HathiTrust Digital Library Resources

Module 3: Working with Textual Data
Lesson Plan

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

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.