Digging Deeper Reaching Further

Libraries Empowering Users to Mine the HathiTrust Digital Library Resources

Curriculum Re-use Guide

Introduction to the HathiTrust Research Center

Duration:

3 hour workshop

Modules:

- 1: Getting Started
- 2.1: Gathering Textual Data: Finding and Curating Textual Data
- 4.1: Analyzing Textual Data: Using Off-the-Shelf Tools

Description:

In this workshop, participants will learn about the HathiTrust Research Center and get some experience with text analysis using off-the-shelf tools. Participants will create an HTRC workset and analyze their workset using a topic modeling algorithm.

Notes and set-up requirements:

This version of the curriculum requires participants to create an account with the HathiTrust Digital Library and the HathiTrust Research Center Analytics site. There are no programming activities in this version, so there is no need to create a PythonAnywhere account or download activity files.

Text as Data

Duration:

2 hour workshop

Modules:

- 1: Getting Started (Slides 1-15)
- 2.2: Gathering Textual Data: Bulk Retrieval
- 3: Working with Textual Data

Description:

In this workshop, participants will gain experience working with text as data. Activities include a basic introduction to the command line, scraping text from a webpage, and removing HTML tags.

Notes and set-up requirements:

This version of the materials focuses on text analysis more broadly, and will not teach participants about the HathiTrust Research Center. Participants will only need to create a PythonAnywhere account and download the activity files in order to participate.

Data Visualization

Duration:

2 hour workshop

Modules:

- 1: Getting Started
    - View “Introduction to the Command Line” video from Module 2.2
- 4.2: Performing Text Analysis: Basic Approaches with Python
- 5: Visualizing Text Data: An Introduction

Description:

In this workshop, participants will learn about the HathiTrust Research Center and create simple data visualizations via the command line and off the shelf tools. Participants will use Python to visualize the most frequently used adjectives in a workset and the number of words per page in a volume, followed by an introduction to data visualization and the HathiTrust+Bookworm visualization tool.

Notes and set-up requirements:

In order to participate in this version of the curriculum, attendees will need to create a PythonAnywhere account and download the activity files. It is not necessary to create an HTRC account in order to use the HathiTrust+Bookworm tool.