Introduction

July 10, 2016

The HathiTrust holds nearly 15 million digitized volumes from libraries around the world. Individually and in aggregate, these works are valuable for humanities scholars. Spanning centuries and genres, a scholar can make inferences about cultural, linguistic, historic, and structural trends in the published word. To simplify access to this collection, the HathiTrust Research Center (HTRC) has released the Extracted Features dataset (https://analytics.hathitrust.org/features).

In this workshop, we introduce a library, the HTRC Feature Reader, for working with the Extracted Features dataset using the Python programming language. However, the skills introduced here are useful beyond this dataset, though. The HTRC Feature Reader is structured to support work using popular data science libraries, particularly Pandas. In teaching analysis of HathiTrust materials through the HTRC Feature Reader, this workshop teaches skills that will benefit general data analysis in Python.

Today, you'll learn:

- How to work with "notebooks", a useful, interactive way of data science in Python;
- Methods to read and visualize text data for millions of books with the HTRC Feature Reader; and
- Data malleability: selecting, slicing, and summarizing Extracted Feature data using the flexible "DataFrame" structure.

0.1 Online Workshop Materials

The newest version of materials related to this workshop is at https://github.com/htrc/ef-workshop. There is also a download page for lesson-files.zip (https://github.com/htrc/ef-workshop/releases/).

0.2 Additional Resources

Beyond the workshop, more information can be found online. Keep this links for continuing your learning beyond today.

- Dataset: The landing page for the HTRC Extracted Features Dataset (https://sharc.hathitrust.org/features) describes the dataset in more detail, and the Documentation Wiki (https://wiki.htrc.illinois.edu/display/COM/HTRC+Extracted+Features+Dataset) provides more detail on other people's uses of the dataset and how to download highly-specific subsets.
- Python Library: The Github repository for the HTRC Feature Reader (https://github.com/htrc/htrc-feature-reader) has another tutorial, and a folder of examples. The Feature Reader code documentation (http://htrc.github.io/htrc-feature-reader/htrc_features/feature_reader.m.html) provides detailed information on all the functionality in the library.