HTRC Derived Datasets

HTRC Extracted Features

HTRC Extracted Features datasets consist of metadata and derived data elements that have been extracted from volumes in the HathiTrust Digital Library. The dataset is periodically updated, including adding new volumes and adjusting the file schema. When we update the dataset, we create a new version. The current version is v.2.0.

Download the data Follow a tutorial

The basics

A great deal of useful research can be performed with features extracted from the full text volumes. For this reason, we generate and share a dataset called the HTRC Extracted Features. The current version of the dataset is Extracted Features 2.0. Each Extracted Features file that is generated corresponds to a volume from the HathiTrust Digital Library. The files are in JSON-LD format.

An Extracted Features file has two main parts:

Metadata

Each file begins with bibliographic and other metadata describing the volume represented by the Extracted Features file.

Features

Features are notable or informative characteristics of the text. The features include:

- Token (word) counts that have been tagged with part-of-speech in order to disambiguate homophones and enable a greater variety of analyses
- Various line-level information, such as the number of lines with text on each page, and a count of characters that start and end lines on each page
- Header and footer identification for cleaner data.

Within each Extracted Features file, features are provided *per-page* to make it possible to separate text from paratext. For instance, feature information could aid in identifying publishers' ads at the back of a book.

Examples and tutorials

- Use Cases and Examples
- Extracted Features in the Wild
- Programming Historian lesson (for EF v.1.0)

Tools for working with HTRC Extracted Features

- HTRC Feature Reader (Python)
- Hathidy (R) – this tool is community-built and not supported by HTRC
- Topic Explorer Extracted Features integration

The versions

Version 2.0 (current)

Documentation

Basic walk-through of an Extracted Features 2.0 file

Get the data

Version 1.5

NOTE: this dataset has been superseded by Extracted Features versions above.

Documentation

Get the data

Version 0.2

NOTE: this dataset has been superseded by Extracted Features versions above.

Documentation

Get the data

Partner-created derived datasets

HTRC has partnered with researchers to create other derived datasets from the HathiTrust corpus. Follow the links below to learn more and access the data.

NovelTM Datasets for English-Language Fiction, 1700-2009 (Ted Underwood, Patrick Kimutis, Jessica Witte)

Description

This dataset is descriptive metadata for 210,305 volumes of English-language fiction in HathiTrust Digital Library. Nineteenth- and twentieth-century fiction are also divided into seven subsets with different emphases (for instance, one where men and women are represented equally, and one composed of only the most prominent and widely-held books). Fiction was identified using a mixed approach of metadata and predictive modeling based on human-assigned ground truth. A full description of the dataset and its creation is available in the dataset report linked below.

Read the report

Get the data from GitHub

Get the data from Zenodo

Word Frequencies in English-Language Literature, 1700-1922 (Ted Underwood)

Description

This dataset contains the word frequencies for all English-language volumes of fiction, drama, and poetry in the HathiTrust Digital Library from 1700 to 1922. Word counts are aggregated at the volume level, but include only pages tagged as belonging to the relevant literary genre. Fiction was identified using a mixed approach of metadata and predictive modeling based on human-assigned ground truth. A full explanation of the dataset's features, motivation, and creation is available on the dataset documentation page below.

Documentation

Get the data

Geographic Locations in English-Language Literature, 1701-2011 (Matthew Wilkens)

Description

The dataset contains volume metadata as well as geographical locations and the number of times the location is mentioned in the text of works of fiction written in English from 1701 - 2011 that are found in the HathiTrust Digital Library. This dataset relied on Ted Underwood's novelTM dataset to determine which volumes to include, and it is part of Matthew Wilkens' larger Textual Geographies Project. Information about the Textual Geographies Project can be found at the Textual Geographies Project link below. A full explanation of the Textual Geographies in English Literature dataset is available at the documentation link below.

Textual Geographies Project

Documentation

Get the data