Extracted Features Dataset [v.0.2]

Page-level features from 4.8 million volumes

*Note that this is a **beta** data release. It has been superseded by the HTRC Extract Features Dataset v.2.0 , which contains data for 17+ million volumes.*

A great deal of useful research can be performed non-consumptively with pre-extracted features. For this reason, we've prepared a data export of features for the public domain volumes of the HathiTrust Digital Library.

Features are notable or informative characteristics of the text. Also, we have processed a number of useful features, including part-of-speech tagged token counts, header and footer identification, and various line-level information. These are provided *per-page*. Providing token information at the page-level makes it possible to separate text from paratext; for instance, a researcher may use the information to identify publishers' ads at the back of a book. For cleaner text, headers and footers are also identified distinctly from page content. The specific features that we extract for each page are described in more detail below.

The most useful extracted feature that we are provide is the token (unigram) count, on a per-page basis. Term counts are specific to the part-of-speech usage for that term, so that a term used as both a noun and a verb, for example, will have separate counts provided for both these modalities of its use. We also include line 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. This information can illuminate genre and volume structure: for instance, it helps distinguish poetry from prose, or body text from an index.

The following release is a beta release.

Data Stats

# of volumes	4,801,237
# of pages	1,825,317,899
Median pages /volume	330

More information about HathiTrust datasets.

Downloading the files

See directions for Ecxtracted Features v.0.2

Attribution

Boris Capitanu, Ted Underwood, Peter Organisciak, Sayan Bhattacharyya, Loretta Auvil, Colleen Fallaw, J. Stephen Downie (2015). *Extracted Feature Dataset from 4.8 Million HathiTrust Digital Library Public Domain Volumes* (0.2)[Dataset]. HathiTrust Research Center, doi:10.13012/j8td9v7m.

Feature File Documentation

The HTRC Extracted Features provides a small amount of metadata in addition to the quantitative features.

Features

Volume-Level Features

schemaVersion: A version identifier for the format and structure of the feature data (HTRC generated).

dateCreated: The time the batch of metadata was processed and recorded (HTRC generated).

pageCount: The number of pages in the volume.

pages: An array of JSON objects, each representing a page of the volume.

Page-Level Features

Pages are contained within volumes, they have a sequence number, and information about their header, body, and footer.

Page-level information

seq: The sequence number. More details on this value.

tokenCount: The total number of tokens on the page.

lineCount: The total number of non-empty lines on the page.

Metadata

A small amount of bibliographic metadata for identifying the volume is included in this dataset. See also: "Where can I find detailed bibliographic metadata?".

id: A unique identifier for the current volume. This is the same identifier used in the HathiTrust and HathiTrust Research Center corpora.

schemaVersion: A version identifier for the format and structure of this metadata object. *metadata.schemaVersion* is separate from *features.schemaVersion* below.

dateCreated: The time this metadata object was processed. *metadata.dateCreated* is not necessarily the same as the *features.dataCreated* below.

title: Title of the volume.

pubDate: The publication year.

language: The primary language of the volume.

emptyLineCount: The total number of empty lines on the page.

sentenceCount: Total number of sentences identified on the page using OpenNLP. Details on parsing.

languages: Automatically inferred language likelihood for the page, Shuyo Nakatani's Language Detection library. Language code reference.

Header, Body, and Footer information

The fields for *header*, *body*, and *footer* are the same, but pertain to different parts of the page. Read about the differences between the sections.

tokenCount: The total number of tokens in this page section.

lineCount: The number of lines containing characters of any kind in this page section. This pertains to the layout of the page; for sentence counts, see the *sentenceCount* field.

emptyLineCount: The number of lines without text in this page section.

sentenceCount: The number of sentences found in the text in this page section, parsed using OpenNLP.

tokenPosCount: An unordered list of all tokens (characterized by part of speech using OpenNLP), and their corresponding frequency counts, in this page section. Tokens are case-sensitive, so a capitalized "Rose" is shown as a separate token. There will be separate counts, for instance, for "rose" (noun) and "rose" (verb). Words separated by a hyphen across a line break are rejoined. No other data cleaning or OCR correction was performed. Details on POS parsing and types of tags used.

beginLineChars: Aggregated frequency counts of the first non-whitespace character on each line.

endLineChars: Count of the last character on each line in this page section (ignoring whitespace).

capAlphaSeq: The longest length of the alphabetical sequence of capital characters starting a line. (Body only).

htBibUrl: The HathiTrust Bibliographic API call for the volume.

handleUrl: The persistent identifier for the given volume.

oclc: The array of OCLC number(s).

imprint: The place of publication, publisher, and publication date of the given volume.

Acknowledgements

Computation was supported by the Cline Center for Democracy through its allocation on the Blue Waters sustained-petascale computing project, which is funded by the National Science Foundation (awards OCI-0725070 and ACI-1238993) and the state of Illinois. Blue Waters is a joint effort of the University of Illinois at Urbana-Champaign and its National Center for Supercomputing Applications.

This release has been made possible, in part, by the National Endowment for the Humanities: Celebrating 50 Years of Excellence. Any views, findings, conclusions, or recommendations expressed in this release do not necessarily represent those of the National Endowment for the Humanities.

Questions

How are tokens parsed?

Hyphenation of tokens at end of line was corrected using custom code. Apache OpenNLP was used for sentence segmentation, tokenization, and part of speech (POS) tagging. No additional data cleaning or OCR correction was performed.

OpenNLP uses the Penn Treebank POS tags.

Can I use the page sequence as a unique identifier?

The *seq* value is always sequential from the start. Each scanned page of a volume has a unique sequence number, but it is specific to the *current* version of the full text. In theory, updates to the OCR that add or remove pages will change the sequence. The practical likelihood of changes in the sequence is low, but uses of the page as an id should be cautious.

A future release of this data will include persistent page identifiers that remain unchanged even when page sequence changes.

Where is the bibliographic metadata? Who wrote the book?; When was it published, etc.?

This dataset is foremost an extracted features dataset, with minimal metadata included as a convenience. For additional metadata information, i.e. subject classifications, etc., HT offers Hathifiles, which can be paired to our feature dataset through the volume *id* field.

The metadata that *is* included in this data includes MARC metadata from HathiTrust and additional information from Hathifiles:

- imprint: 260a from HathiTrust MARC record, 260b and 260c from Hathifiles.
- language: MARC control field 008 from Hathifiles.
- pubDate: extracted from Hathifiles. See also: details on HathiTrust's rights-determination.

- oclc: extracted from Hathifiles.

Additionally, schemaVersion and dateCreated are specific to this feature dataset.

What do I do with beginning- or end-of-line characters?

The characters at the start and end of a line can be used to differentiate text from paratext at a page level. For instance, index lines tend to begin with capitalized letters and end with numbers. Likewise, lines in a table of contents can be identified through arabic or roman numerals at the start of a line.

What is the difference between the header, body, and footer sections?

Because repeated headers and footers can distort word counts in a document, but also help identify document parts, we attempt to identify repeated lines at the top or bottom of a page and provide separate token counts for those forms of paratext. The "header" and "footer" sections will also include tokens that are page numbers, catchwords, or other short lines at the very top or bottom of a page. Users can of course ignore these divisions by aggregating the token counts for header, body, and footer sections.

Contact Us

htrc-help@hathitrust.org

Tools

If you've built tools or scripts for processing our data, let us know and we'll feature them here!

Projects

Let us know about your projects and we'll link to them here.