HTRC Analytics Algorithms

Read about the different algorithms HTRC offers, and the kinds of data each algorithm provides.

HTRC Algorithms are click-to-run tools for text analysis. They require no programming, and researchers can set the parameters for their analysis. You can run HTRC algorithms against collections of HathiTrust data, called worksets, in order to analyze them or download their Extracted Features. Worksets can be cited, and researchers can choose to make their worksets public or private.

Run an algorithm Follow a tutorial

The basics

The data

HTRC Algorithms are designed to be run on sub-collections of volumes from HathiTrust, called worksets. You cannot run HTRC Algorithms on non-HathiTrust data.

HTRC Algorithms can analyze volumes in a workset so long as they have been synched with HTRC from HathiTrust. While syncing happens regularly, there may be occasional discrepancies.

HTRC Algorithms can analyze in-copyright ("limited view") as well as public domain ("full view") volumes from HathiTrust.

HathiTrust data is not exposed or viewable within HTRC Algorithms or worksets. A researcher applies an algorithm to their workset (collection) and the data is called and crunched behind the scenes. **Only the results are viewable.**

Create and browse worksets Follow a tutorial

Algorithm results

Every time you run an algorithm against a workset, that's called a *job*. You are able to view the status of the jobs that you have submitted. You can also delete jobs, for example if you have made an error in your set-up and want start again.

The results of your jobs are stored in HTRC Analytics and you can also download certain results files for each algorithm.

New: Starting in October 2019, job results will be available for 18 months from the run date, after which they will expire. An expiration date will appear next to each job on your jobs page. Download your results prior to the expiration date to make sure you don't lose them.

Run an algorithm   Follow a tutorial

The algorithms

Extracted Features Download Helper

Helps you download HTRC Extracted Features files.

Generates a shell script that, when run locally from the command line on your computer, will download the Extracted Features files for the volumes in a workset. Extracted Features files are accessed via a command line utility called rsync. The script produced by this algorithm contains the commands to rsync (download) each volume in your workset.

Result of job

A shell script to download Extracted Features data files

Volume limit

None

Specifications and parameters

- If there is a volume in your workset for which there is not a corresponding Extracted Features file, that volume will be skipped in the shell script. Extracted Features files are periodically generated and may be out-of-sync with the full HathiTrust corpus.
- For more information on the Extracted Features data, see https://analytics.hathitrust.org/features.

How to utilize results

Steps with screenshots

- Download the shell script file, called EF_Rsync.sh
- If you like, rename and/or move the file from your download folder

- From the Bash shell (Terminal on a Mac, or Cygwin/GitBash on Windows), navigate to the directory where your results script is saved. Then run the shell script, modifying the file name if you changed it earlier:

```
sh EF_Rsync.sh
```

- The HTRC Extracted Features json file for each volume in your workset will be downloaded to your machine.

Author

Colleen Fallaw

Current Version

3.0.2

InPhO Topic Model Explorer

Uses a machine learning process called topic modeling to group the words in your workset into "topics" most prevalent in the text.

Trains multiple LDA topic models using a tool called the InPho Topic Explorer, which has been integrated into HTRC Analytics, but can also run locally or in a Data Capsule. The algorithm allows you to export files containing the word-topic and topic-document distributions, along with an interactive visualization. For full detailed description of the InPho Topic Explorer, please review the documentation.

Result of job

Four files are generated. Three are for displaying a visualization of topic clusters and top terms: topics.html, cluster.csv, topics.json. The final file (workset.tez) can be used with a local install of the Topic Explorer to access the complete word-topic and topic-document matrices, along with other advanced analytics.

Volume limit

3000 volumes OR 3GB

Specifications and Parameters

- The text for the volumes in the workset is tokenized using the Topic Explorer's *init* functionality, which:
    - Normalizes the text.
    - Performs well with Indo-European languages, including English, Polish, Russian, Turkish, Greek, Italian, Latin, French, German, Spanish. Tokenizer does not perform well with East Asian languages or any other languages without spaces in the orthography.
- Stoplisting is performed based on the frequency of terms in the corpus. The most frequent words (accounting for 50% of the workset) and the least frequent words (accounting for 10% of the workset) are removed.
- Topic models are created based on the parameters set for the job.
    - *Iterations:* A lower number of iterations (i.e. 200 iterations) will process faster and is good for experimentation. A higher number will give publication-ready results (i.e. 1000 iterations).
    - *Topics:* You set the number of topics to be created, and multiple numbers can be added for one job. Entering "20 40 60 80" in the number of topics, the algorithm will train separate models with 20 topics, 40 topics, 60 topics and 80 topics.
- Generates a bubble visualization that shows how topics across models cluster together.
- More documentation of the Topic Explorer is available at https://inpho.github.io/topic-explorer/.

How to utilize results

1) Bubble visualization

- Enables you to see the granularity of the different models and how terms may be grouped together into "larger" topics.
- If you trained multiple models, you can toggle the views of the corresponding bubbles for each model. The size of the bubbles/nodes relates to the number of topics generated (larger for fewer topics, smaller for more topics).
    - You have the option to turn collision detection on and off.
- The bubbles cluster based on the similarity of the topics.
- The clusters and colors are determined automatically by an algorithm, and provide only a rough guide to groups of topics that have similar themes. The different axes also do not have any intrinsic meaning, but may be interpretable as representing historical or thematic dimensions in the underlying corpus.

2) topics.json

- See the words in each topic and the corresponding probability for each word appearing in the topic.
- Download the file and do further analysis and visualization locally.

3) workset.tez

- Download the file to access a more robust visualization of your topics.
- Follow the instructions to install the InPho Topic Explorer locally, and then import your workset.tez file to view a more interactive visualization.
    - The InPho Topic Explorer tool may not work correctly on a Windows computer. HTRC is working to fix this.
    - We recommend renaming the workset.tez file after downloading it in order to disambiguate results from multiple jobs.
    - Navigate to the directory where you've put your workset.tez file and run the following commands:

        - topicexplorer import workset.tez
          topicexplorer launch workset.ini
    - The models were already created in HTRC Analytics. These commands simply load the results into the tool and display them locally.

Watch a video of a locally-running Topic Explorer on YouTube.

Author

Jaimie Murdoch

Current Version

1

Token Count and Tag Cloud Creator

Identify the tokens (words) that occur most often in a workset and the number of times they occur. Create a tag cloud visualization of those most frequently occurring words, where the size of the word is displayed in proportion to the number of times it occurred.

Result of job

Tag cloud showing the most frequently occurring words, and a file (token_counts.csv) with a list of those words and the number of times they occur.

Volume limit

3000 volumes or 3GB

Specifications and Parameters

- Prepares the data by identifying page header, body, and footer and extracting the page body only for analysis.
- Combines of end-of-line hyphenated words in order to de-hyphenate the text.
- Removes stop words using either:
    - The default stop word list available in the algorithm,
    - Or a custom list saved as a text file that exists somewhere with a web-accessible URL. Add the URL to the appropriate field when setting parameters for your job.
    - If left blank, no stopwords will be removed.
- Applies replacement rules (i.e. corrections), maintaining the original case of the replaced words, using either:
    - The default list of corrections available in the algorithm,
    - Or a custom list saved as a CSV file that exists somewhere with a web-accessible URL. The file should be formatted so that the first column is the word to be replaced and the second is the replacement. The file must have a header row. Add the URL to the appropriate field when setting parameters for your job.
    - If left blank, no replacements will be made.
- Tokenizes the text using the Stanford NLP model for the language specified by the user, or does white-space tokenization
    - Enter the 2-letter code for the most prominent language in the workset.
    - If the language is not English (en), French (fr), Arabic (ar), Chinese (zh), German (de), or Spanish (es), then the text will be tokenized using white space.
- Regular expression pattern matching is used to control what appears in the tagcloud.
    - We use provide a base regular expression that limits the display to only words that are made of letters or that contain a hyphen.
    - This parameter does not affect what words will appear in the token count file.
- Tokens are counted for the entire workset, and then sorted in descending order for the resulting token count file.
    - The top tokens are displayed in a tagcloud. Choose how many tokens to display in the visualization when the job is run.

How to utilize results

- View the tag cloud visualization
    - Most-frequently used words in the workset are displayed larger, while the less-frequently used words are displayed smaller.
    - View the token count list (token_counts.csv)
    - Download token_counts.csv and compare the token counts for one workset to the counts of another workset, or create your own visualization

Author

Boris Capitanu

Version

1

Named Entity Recognizer

Description

Generate a list of all of the names of people and places, as well as dates, times, percentages, and monetary terms, found in a workset. You can choose which entities you would like to extract.

Result of job

Table of the named entities found in a workset (entities.csv)

Volume limit

3000 volumes or 3GB

Specifications and Parameters

- Prepares the data by identifying page header, body and footer and extracting page body only for analysis.
- Combines of end-of-line hyphenated words in order to de-hyphenate the text
- Tokenizes the text using the Stanford NLP model for the most prominent language in the workset, identified by a 2-letter code for that language
    - This algorithm only supports English (en), French (fr), Arabic (ar), Chinese (zh), German (de), or Spanish (es). Entering the code for any other language will fail.
- Performs entity recognition/extraction using the Stanford Named Entity Recognizer, and then shuffles the entities found on each page in order to prevent aiding page reconstruction.

How to utilize results

- View the named entities list (entities.csv), which shows the volume ID where the entity was found, the page sequence on which the entity occurred, the entity, and the entity type (person, place, etc.)
- Download entities.csv for further analysis locally.
    - Compare the results of multiple worksets.
    - Create a visualization of the entities.

Author

Boris Capitanu

Version

2

Deprecated algorithms

These algorithms are no longer available in HTRC Analytics:

- Naive-Bayes classification
- MARC Downloader
- Meandre Dunning Log-likelihood to Tagcloud
- Simple Deployable Word Count
- Meandre Topic Modeling
- Meandre Tagcloud
- Meandre Tagcloud with Cleaning
- Meandre Spellcheck Report Per Volume
- Meandre OpenNLP Entities List
- Meandre OpenNLP Date Entities To Simile