Text Mining with HathiTrust: An Introduction for Librarians

Set up checklist

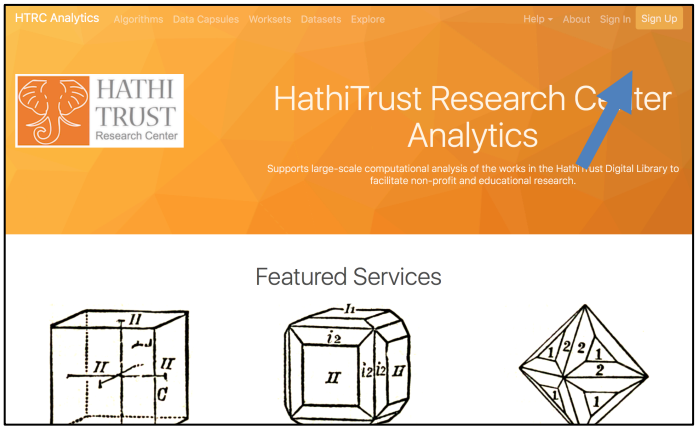
☐ **Access workshop materials**

1. https://uofi.box.com/v/HTRC-fall2019
2. Click through to the 'Librarian_workshop' folder.
3. (Optional) Download the folder, or whichever files you like by clicking the "More options" button (···) on the right.
4. Select "**Download**".
5. The folder will download as a zipped file to your computer.

☐ **Create an HTRC Analytics account**

http://analytics.hathitrust.org/

1. Click "Sign Up" in the top right corner.
2. Use an email address from an academic institution and follow security guidelines for the password.
3. Activate your account from the link you will be sent via email.



☐ Load GitHub page with links to the Jupyter Notebook activities: https://go.illinois.edu/htrc-workshop

☐ **Verify you** aren't **using Internet Explorer**

Text as data

KEY TOOLS & PLATFORMS

.

HathiTrust + Bookworm

A tool that visualizes word frequencies over time in the HathiTrust Digital Library

ACTIVITY: Visualize word trends ***Slide 38***

Use the HT+BW tool at: https://bookworm.htrc.illinois.edu/develop to visualize unigrams

- Experiment with the settings and faceting.
- Get creative

CASE STUDIES: Characterize the data used ***Slide 45***

Read all 3 case studies

Then characterize the data used, such as:

- What criteria did they use to build their corpus?
- What was the period of study?
- Did they have access to the full text?

READ AND REFLCT: Collections as data ***Slides 50-52***

Read excerpts from the Santa Barbara Statement on Collections as Data (full statement: https://collectionsasdata.github.io/statement/)

Then answer:

- Does your library provide access to digital collections as data?
- How so? Why not? How could it?

Research with text data

KEY TOOLS & PLATFORMS

HTRC Analytics

An interface for working with HTRC worksets, which are collections of text from HathiTrust that can be analyzed using the non-consumptive tools and environments in HTRC Analytics.

HTRC algorithms

A set of off-the-shelf text analysis algorithms provided via HTRC Analytics for users to analyze their worksets, such as algorithms for extracting named entities and doing topic modeling.

ACTIVITY: Teach your neighbor about text pre-processing ***Slide 59***

Term	Definition
Punctuation	"The first choice a researcher must make when deciding how to preprocess a corpus is what classes of characters and markup to consider as valid text. The most inclusive approach is simply to choose to preprocess all text, including numbers, any markup (html) or tags, punctuation, special characters ($, %, &, etc), and extra white-space characters. These non-letter characters and markup may be important in some analyses (e.g. hashtags that occur in Twitter data), but are considered uninformative in many applications. It is therefore standard practice to remove them. The most common of these character classes to remove is punctuation."
Numbers	"While punctuation is often considered uninformative, there are certain domains where numbers may carry important information. For example, references to particular sections in the U.S. Code ('Section 423', etc.) in a corpus of Congressional bills may be substantively meaningful regarding the content legislation. However, there are other applications where the inclusion of numbers may be less informative."
Lowercasing	"Another preprocessing step taken in most applications is the lowercasing of all letters in all words. The rationale for doing so is that that whether or not the first letter of a word is uppercase (such as when that words starts a sentence) most often does not affect its meaning. For example, 'Elephant' and 'elephant' both refer to the same creature, so it would seem odd to count them as two separate word types for the sake of corpus analysis. However, there are some instances where a word with the same spelling may have two different meanings that are distinguished via capitalization, such as 'rose' (the flower), and 'Rose' the proper name."

Stemming	"The next choice a researcher is faced with in a standard text preprocessing pipeline is whether or not to stem words. Stemming refers to the process of reducing a word to its most basic form (Porter, 1980). For example the words 'party', 'partying', and 'parties' all share a common stem 'parti'. Stemming is often employed as a vocabulary reduction technique, as it combines different forms of a word together. However, stemming can sometimes combine together words with substantively different meanings ('college students partying', and 'political parties'), which might be misleading in practice."
Stopword Removal	"…some words, often referred to as "stop words", are unlikely to convey much information. These consist of function words such as 'the', 'it', 'and', and 'she', and may also include some domain-specific examples such as 'congress' in a corpus of U.S. legislative texts. There is no single gold-standard list of English stopwords, but most lists range between 100 and 1,000 terms."
n-gram Inclusion	"While it is most common to treat individual words as the unit of analysis, some words have a highly ambiguous meaning when taken out of context. For example the word 'national' has substantially different interpretations when used in the multi-word expressions: "national defense", and "national debt". This has led to a common practice of including n-grams from documents where an n-gram is a contiguous sequence of tokens of length n (Manning and Schutze, 1999). For example, the multi-word expression 'a common practice' from the previous sentence would be referred to as a 3-gram or tri-gram."
Infrequently Used Terms	"In addition to removing common stopwords, researchers often remove terms that appear very infrequently as part of corpus preprocessing. The rationale for this choice is often two-fold; (1) theoretically, if the researcher is interested in patterns of term usage across documents, very infrequently used terms will not contribute much information about document similarity. And (2) practically, this choice to discard infrequently used terms may greatly reduce the size of the vocabulary, which can dramatically speed up many corpus analysis tasks."

From: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2849145 (Denny and Spirling, 2017)

CASE STUDIES: Characterize the data used ***Slide 60***

Look back at the Creativity Boom *case study*

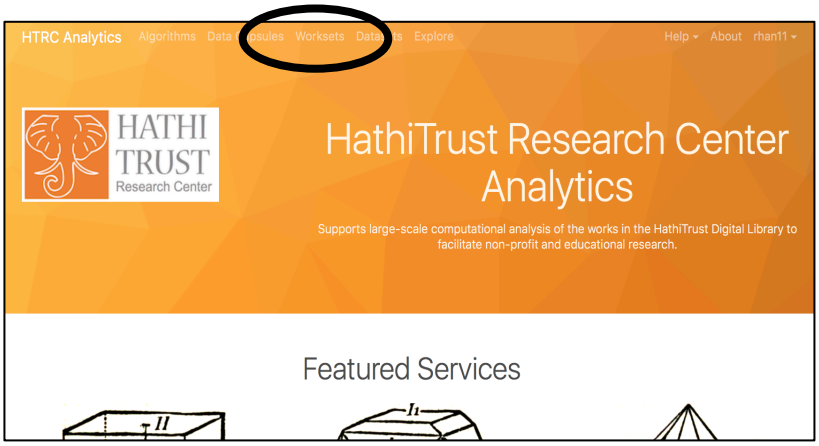
Consider the following questions:

- What steps did he take to prepare his data?
- What assumptions did he make while preparing his data?

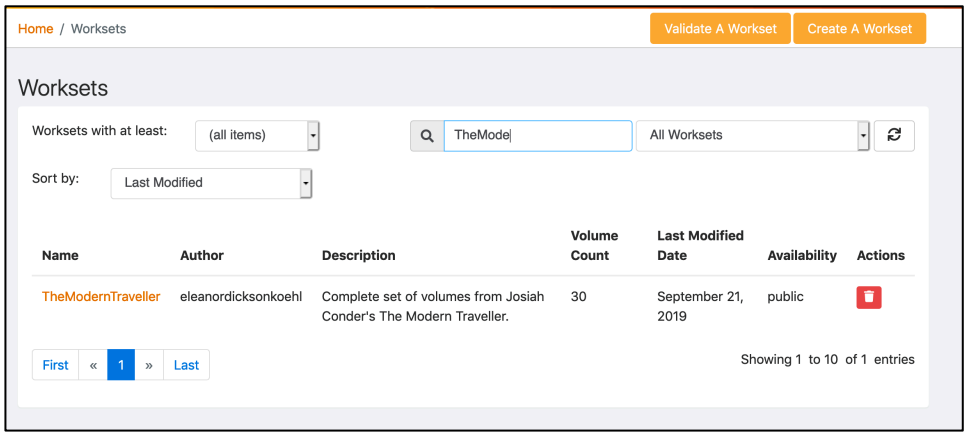
ACTIVITY: Run an HTRC Algorithm ***Slide 62***

Let's try performing a popular text analysis method, topic modeling, using a web-based tool.

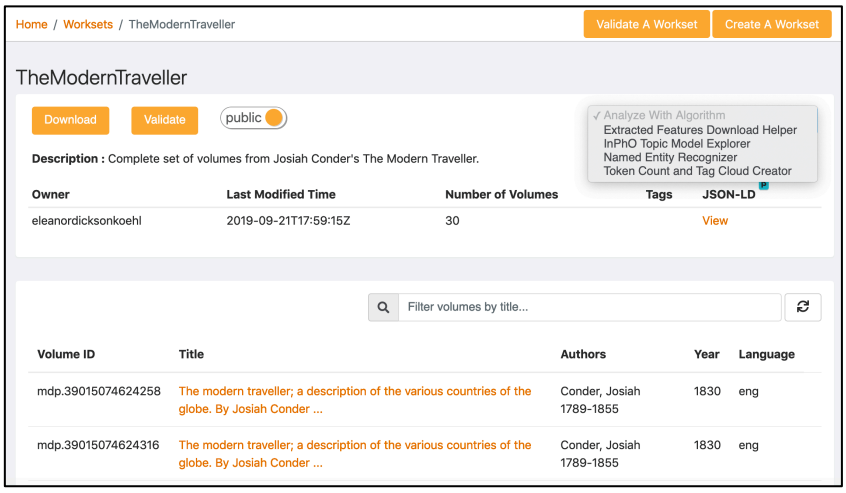
1. From the homepage of HTRC Analytics, click "Worksets."



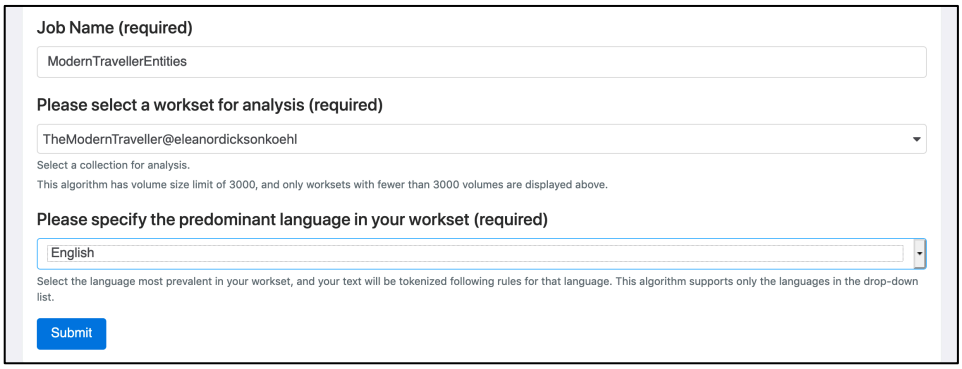
2. Choose "All Worksets" from the dropdown menu, and then filter by typing the name of the workset "**TheModernTraveller**".
3. Select the workset from the list by clicking on the name.



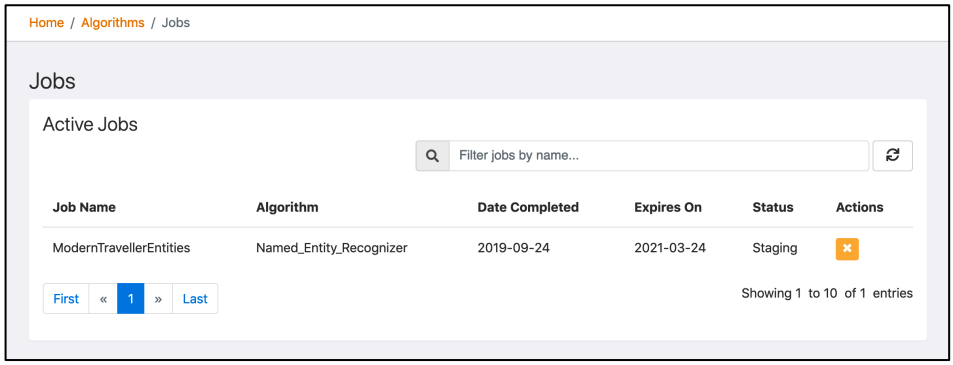
4. Select “Named Entity Recognizer” from the “Analyze with Algorithm” menu.



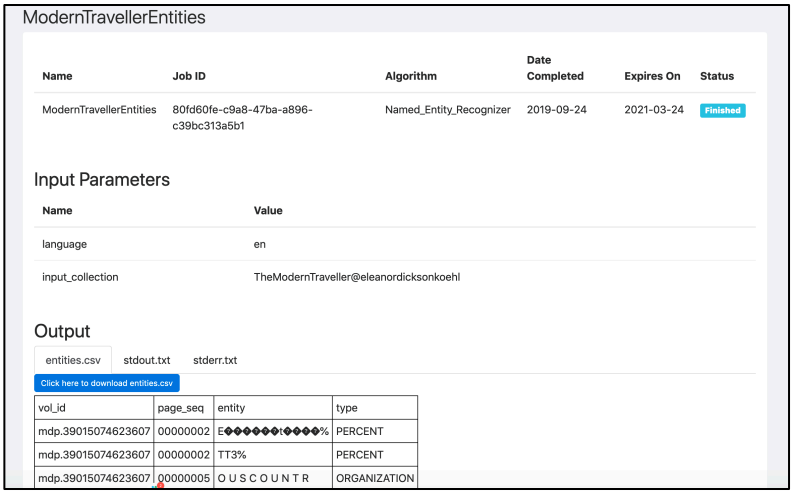
5. Enter a name for your job and select “English” as the predominate language.
6. Click “Submit.”



7. See the current job in “Active Jobs” and refresh your screen to see the status change.
8. You may have to be patient while it finishes, especially if the workset is large.



9. Once the job is done, it will be listed under “Completed Jobs.”
10. Click on the job name to see the results. Scroll to the “output” area to see the preview of the CSV file that was generate.
11. You can download the file if you like.



CASE STUDIES: Explore the research question ***Slide 74***

Return to the How Capitalism Changed American Literature *case study.*

Answer the following questions:

- What was his research question?
- Why is it a good question for use of text analysis?

Text Analysis Methods

KEY TOOLS & PLATFORMS

> **Jupyter Notebook**
> A format for storing live code and documentation.

ACTIVITY: Get started with Jupyter notebooks ***Slide 84***

Practice exercises with HTRC Extracted Features files in a Jupyter notebook.

1. Go to https://go.illinois.edu/htrc-workshop and click "Launch Binder"
2. Open 01-learning-jupyter-environment.ipynb

ACTIVITY: Map location entities ***Slide 85***

Work in a Jupyter notebook to map the location entities you extracted from The Modern Traveller.

1. Go to https://go.illinois.edu/htrc-workshop and click "Launch Binder" OR directly from Binder,
2. Open 02-geocoding-module.ipynb

CASE STUDIES: Identify the methods ***Slide 90***

Look back at all 3 case studies

Then characterize the methods use:

- What methods did they each use?
- Can you tell how their text data was represented (bag of words, word embeddings)?

Text Analysis Workflows

KEY TOOLS/PLATFORMS

> **HTRC Extracted Features**
> A downloadable dataset of text data and metadata extracted from volumes in HathiTrust.

HTRC Feature Reader

A Python library for working with HTRC Extracted Features.

ACTIVITY: Examine an Extracted Features file ***Slide 104***

Download a sample Extracted Features file and explore the data format.

1. Go to the Box folder.
2. Either download and open the file or preview it in Box. (If you download it and open it in Firefox, it will be nicely formatted.)
3. See if you can find:
    - The OCLC number
    - How many lines are in the body of page sequence 0005

ACTIVITY: Work with HTRC Extracted Features ***Slide 109***

Practice exercises with HTRC Extracted Features files in a Jupyter notebook.

1. Go to https://go.illinois.edu/htrc-workshop and click “Launch Binder” OR directly from Binder,
2. Open 03-ef-activity.ipynb

CASE STUDIES: Identify the methods ***Slide 110***

Go back to the Transformation of Gender case stud, and also open go to the documentation for BookNLP: https://github.com/dbamman/book-nlp

Then can you:

- Explain the research pipeline to your neighbor?
- What tools or steps does it include?

More resources

Email htrc-help@hathitrust.org

Office hours: every 3rd Wednesday at 3 ET go.illinois.edu/htrchelp-live

HTRC Wiki: wiki.htrc.illinois.edu