Getting Started with the HathiTrust Research Center

SET UP

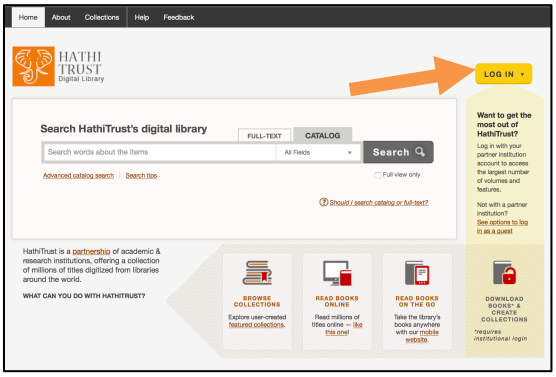
- ☐ Access workshop materials
- ☐ Create an HTDL account
- ☐ Create an HTRC Analytics account
- ☐ Verify you **are not** using Internet Explorer

Access workshop materials

1. https://uofi.box.com/v/digital-initiatives-htrc

Create an HTDL account

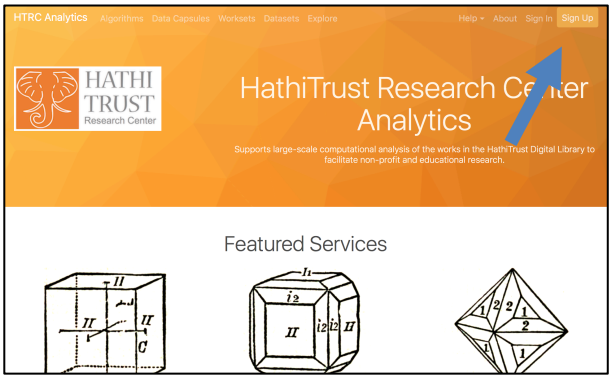
https://www.hathitrust.org



1. Click “LOG IN” in the top right corner.
2. If you are affiliated to an HT partner institution, select your institution and then follow the directions for institutional log in.
3. If you are not affiliated to an HT partner institution, you can log in as a guest.
    - Click “See options to log in as a guest”.
    - You can log in with a Google, Facebook, Twitter, AOL, LinkedIn, Windows Live (Hotmail), Yahoo!, or University Michigan Friend Account.
    - Click on an option of your choice and follow the directions.

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.

SECTION 1 Introduction

KEY TOOLS & PLATFORMS

HathiTrust

A library consortium founded in 2008. HathiTrust is a community of research libraries committed to the long-term curation and availability of the cultural record.

The HathiTrust Digital Library (HTDL)

HathiTrust's digital preservation repository and access platform for public domain and in-copyright content from a variety of sources, including Google, the Internet Archive, Microsoft, and in-house partner institution initiatives. Overall, the content mostly consists of digitized books from libraries.

The HathiTrust Research Center (HTRC)

HathiTrust's center for facilitating computational, scholarly research using the 16+ million volumes in the HathiTrust Digital Library. The HTRC provides mechanisms for non-consumptive access to content in the HathiTrust corpus, as well as tools for computational text analysis.

ACTIVITY: Explore sample research projects

☞ ***Slide 9***

In pairs or small groups, review the summarized research projects available at **http://go.illinois.edu/ddrf-research-examples**.

Then discuss the following questions:

- *How do the projects involve change over time, pattern recognition, or comparative analysis?*
- *What kind of text data do they use (time period, source, etc.)?*
- *What are their findings?*

SECTION 2 Gathering Textual Data

KEY TOOLS & PLATFORMS

HT Collection Builder

An interface for creating collections via the HathiTrust Digital Library.

HTRC Analytics

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

HTRC Extracted Features

A downloadable dataset of text data and metadata extracted and abstracted from volumes in the HathiTrust Digital Library.

ACTIVITY: Evaluating different sources for textual data

☞ ***Slide 38***

If we are building a corpus for political history, what are the strengths and weaknesses of each of the following broad sources for textual data? Please discuss and take some notes in the chart below.

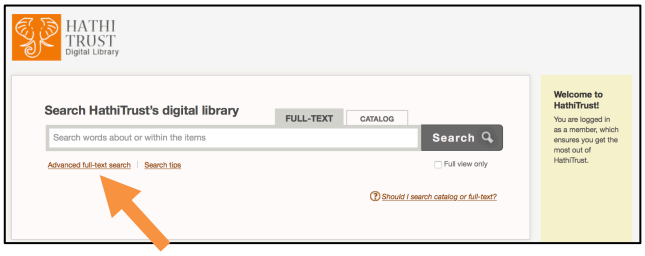
	Strengths	Weaknesses
Vendor database		
Library and archives digital collections		
Social media		

ACTIVITY: Building and uploading a workset

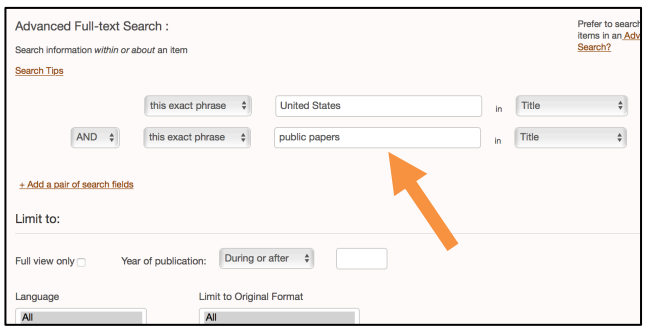
☞ ***Slide 47***

Let's create a workset for a political science student. As an example, we'll do an advanced full-text search for volumes that contain both "public papers" and "United States" in their titles

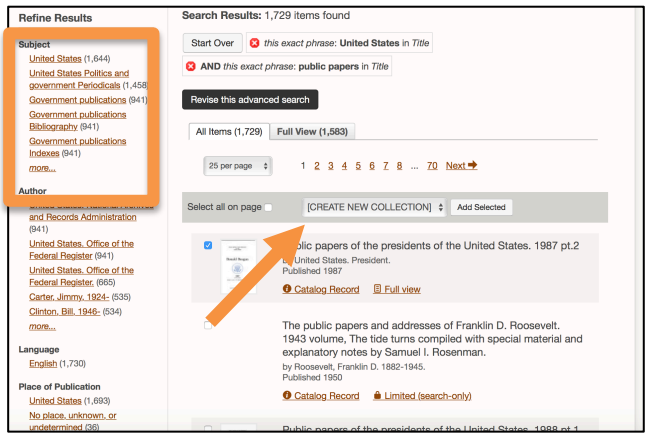
1. Make sure you are still logged in to the HTDL: https://www.hathitrust.org/
2. Click the "FULL-TEXT" tab to search in full text.
3. Click "Advanced full-text search" under the search bar.



4. Search "public papers" and "United States" in the titles, and select "this exact phrase" and "Title" to limit our search.
5. Click "Search" at the bottom of the page.



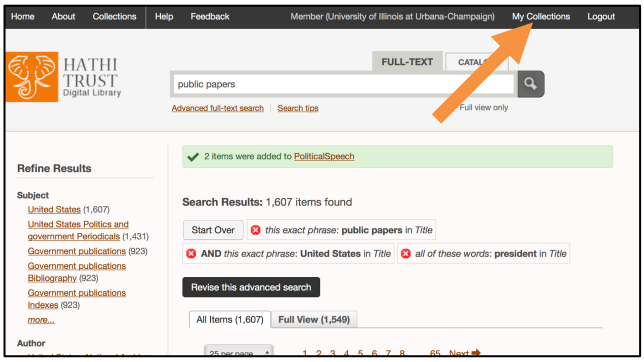
6. Filter your results using the fields on the left.
7. Click the checkbox next to the titles you would like to add to your collection.
8. When ready to create your collection, click on the “Select Collection” bar and choose “[CREATE NEW COLLECTION]” from the drop-down menu. Click “Add Selected”.



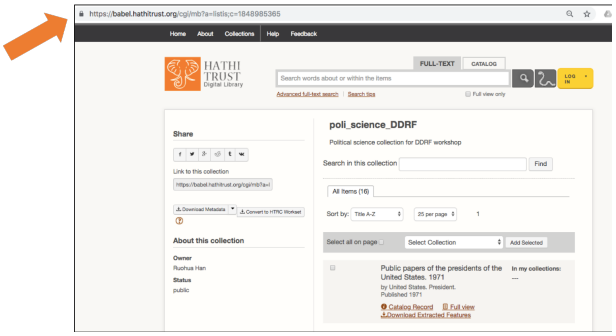
9. Enter information about your collection in the pop-up window. Click “Save Changes”.

10. You’ll see a confirmation that your collection was created. Click “My Collections” in the top right.

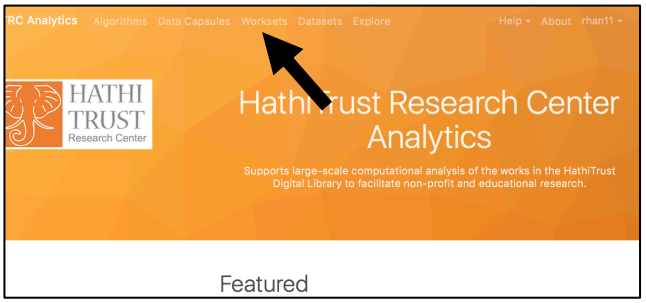
11. Click the title of the collection you just created to view it.



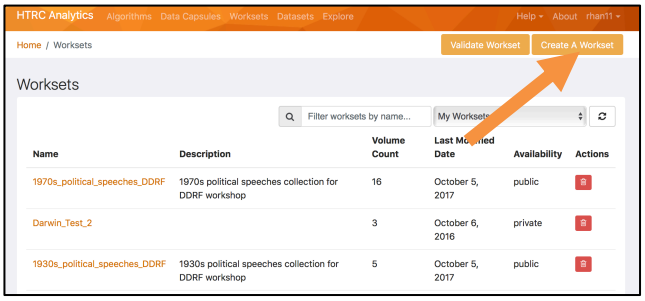
12. You will be able to see the title and description of your collection, as well as all the items in it. Copy (highlight and ctrl-C/cmd-c) the URL in the bar. You will need this to create your workset.



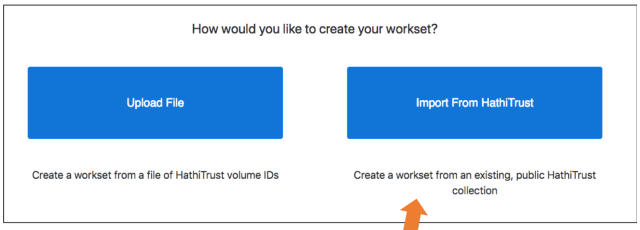
13. Log in to HTRC Analytics: https://analytics.hathitrust.org
14. Go to “Worksets” in the header menu.



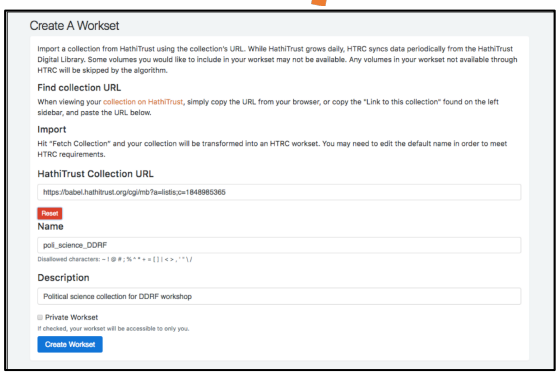
15. Click on “Create a workset” near the top right.



16. There are 2 creation options for HTRC worksets: either upload a file containing a list of HathiTrust volume IDs if curated outside of the HTDL interface, or import from the HTDL using the collection URL. This activity uses the “import from HT” method.



17. Paste (ctrl-v/cmd-v) your HT collection URL in the Collection URL field and click to retrieve the information for the collection.
18. Enter information about your workset. Only characters A-Z 0-9 () - or _ are allowed for the name of your workset.
19. Click “Create Workset”. You should be able to find your new workset on your Worksets page.



SECTION 3 Preparing Textual Data

ACTIVITY: Teach your neighbor about text pre-processing

☞ ***Slide 96***

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)

SECTION 4 Analyzing Textual Data

KEY TOOLS

HTRC algorithms

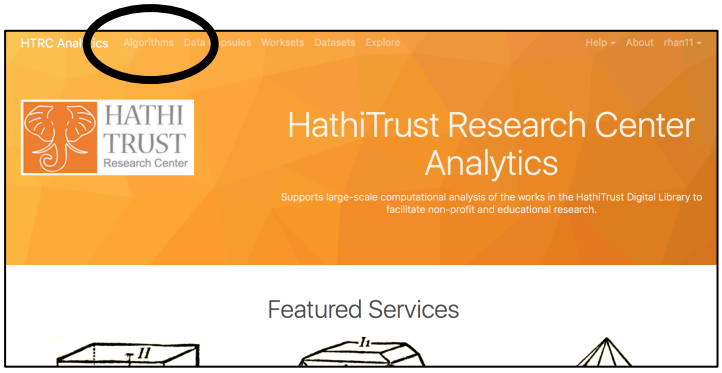
A set of off-the-shelf text analysis algorithms provided via HTRC Analytics for users to analyze their worksets, such as algorithms for extracting named entities and doing topic modeling.

ACTIVITY: Running an algorithm in HTRC Analytics

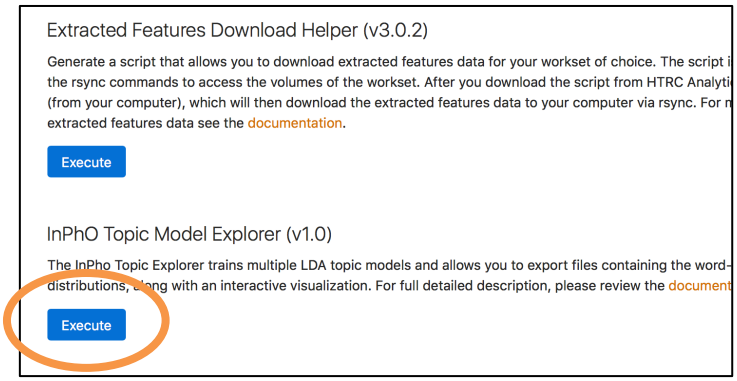
☞ ***Slide 115***

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

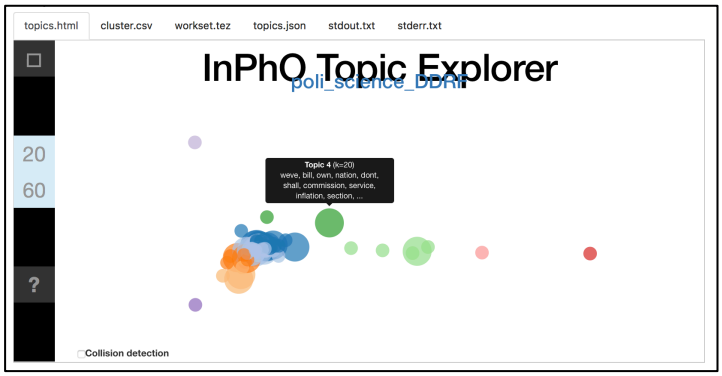
1. From the homepage of HTRC Analytics, click “Algorithms.”

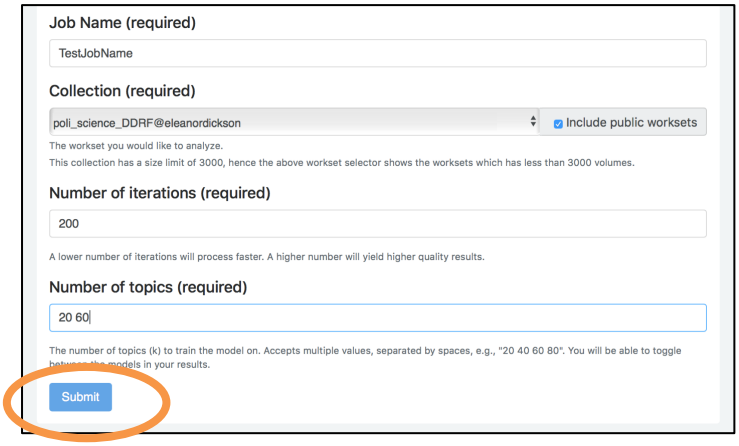


2. Click on the “Execute” button under the name and description of the algorithm you want to run. Select “InPhO Topic Model Explorer (v1.0)” for this activity.

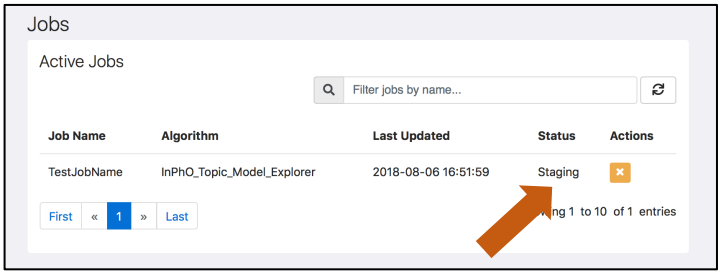


3. Choose a workset from either all worksets or just your private worksets.
4. For this example exercise, check the “Include public worksets” option and select “**poli_science_DDRF@eleanordicksonkoehl**”.
5. To navigate to the workset more quickly, after clicking on the arrow button to expand the list of worksets, type “EF” and the down arrow and the workset that we need will appear at the bottom of the list.
6. Enter a name for your job, type “200” for the number of iterations, and type “20 60” for the number of topics to be created. Click “Submit.”



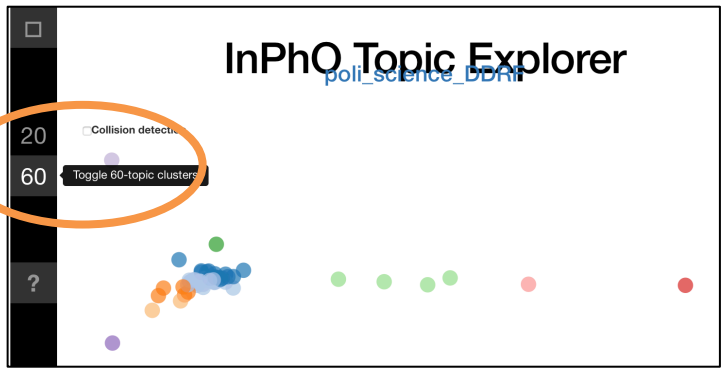


7. See the current job in “Active Jobs” and refresh your screen to see the status change.
8. You may have to be patient while it finishes, especially if the workset is large.
9. Once the job is done, it will be listed under “Completed Jobs.”

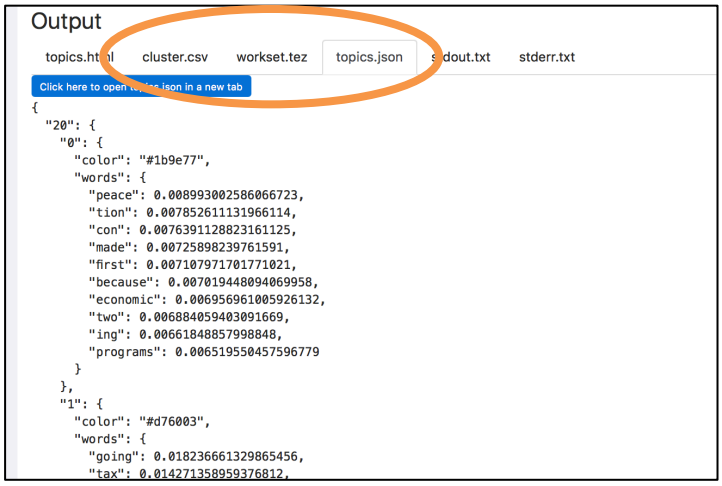


10. Click on the job name to see the results. Scroll to the “output” area to see the bubble visualization of the generated topics. Hover over a bubble to see the top terms in a topic.

11. The numbers on the side relate to the number of topics generated, as do the size of the bubbles. Toggle the display of the n-topic clusters by clicking on those numbers.



12. You can also view and download 3 results files: topics.json, cluster.csv, and workset.tez. These files can be used to play with the visualization in more depth outside HTRC Analytics.



ACTIVITY: Identify the method

☞ *Slide 132*

What are the broad areas and methods used for the research examples we read earlier?

Project summaries: http://go.illinois.edu/ddrf-research-examples

	Broad area	Specific method
Rowling and "Galbraith": an authorial analysis		
Significant Themes in 19th Century Literature		
The Emergence of Literary Diction		

SECTION 5 Visualizing Textual Data

KEY TOOLS/PLATFORMS

HathiTrust + Bookworm

A tool that visualizes word frequencies over time in the HathiTrust Digital Library. It can be accessed at: https://bookworm.htrc.illinois.edu/develop.

ACTIVITY: Which visualization technique?

☞ ***Slide 154***

Match the type of use to type of visualization:

Visualization	What would it be good for?
Word cloud	
Trees/hierarchies	
Networks	
Timeline	
Map	
Bubble chart	
Heat map	

Uses
Change over time
Spatial
Topical density
Relationships
Word distribution

Bonus: *what kinds of variables (i.e. data points) you would need for each visualization?*

ACTIVITY: Visualize word trends

☞ ***Slide 166***

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

- As examples, you could try: fascism, socialism, nationalism, or internationalism.
- Experiment with the settings and faceting.

Get creative!