Text Mining with HathiTrust

An Introduction for Librarians

HATHI
TRUST
RESEARCH
CENTER

Set up instructions

Handout p. 1

- Create HTRC account: analytics.hathitrust.org
- Find workshop materials:
  - https://uofi.box.com/v/HTRC-fall2019
  - https://go.illinois.edu/htrc-workshop

Eleanor Koehl

Associate Director for Outreach & Education, HTRC

efdkoehl@hathitrust.org

Ryan Dubnicek

Digital Humanities Specialist, HTRC

rdubnic2@illinois.edu

Introduce yourself to your neighbor

Workshop outline/structure

- Introduction
- Text as data
- Break (20 min)
- Research with text data
- Lunch (60 min)
- Text analysis methods
- Break (20 min)
- Text analysis workflows
- Other HTRC services

Introduction

In this section we will…

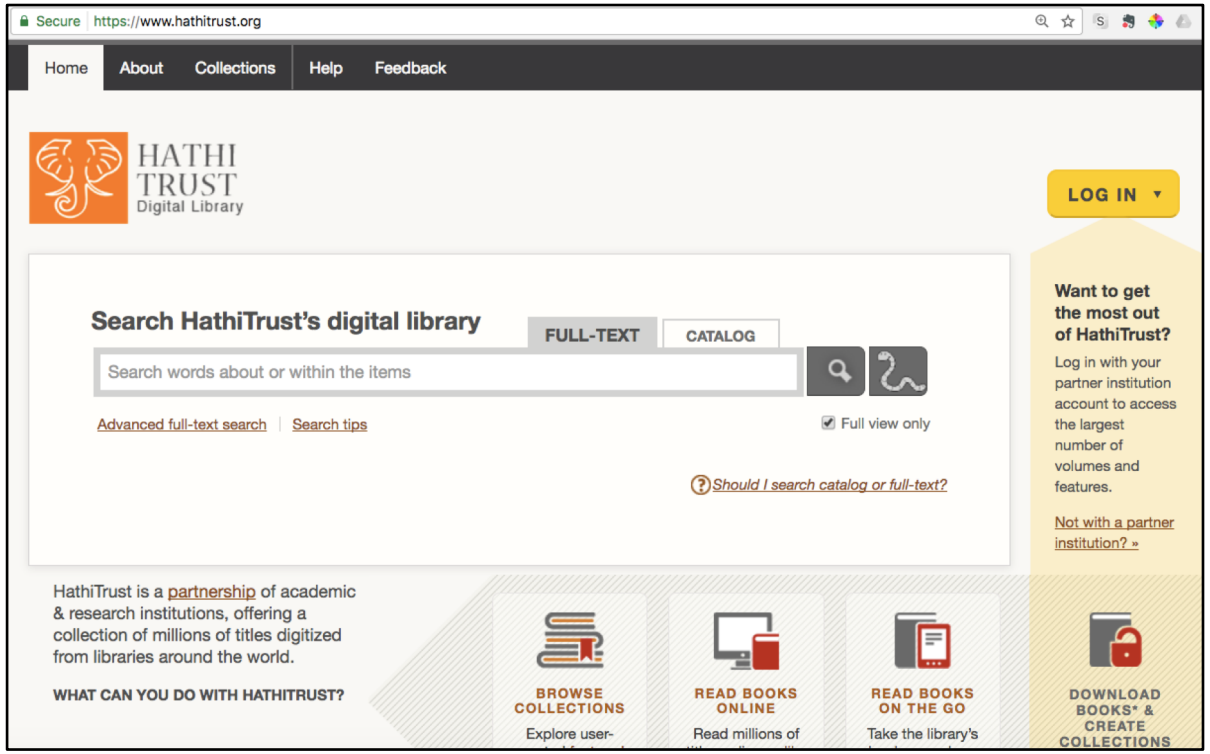
- Briefly introduce HathiTrust and its Research Center
- Define text analysis
- Introduce our case studies

HathiTrust

- Digital library partnership
  - More than 150 member institutions
- Founded in 2008
- Grew out of large-scale digitization initiative at research libraries
  - Roots in Google Books project

HathiTrust Digital Library

- 17+ million volumes
  - 62% in copyright or of undetermined status
- Search and read books in the public domain



HathiTrust Research Center

- Facilitates text analysis of HathiTrust content
- Research & Development
- Located at Indiana University and the University of Illinois
  - With support from HathiTrust

HTRC Analytics

https://analytics.hathitrust.org

What is text analysis?

- Using computers to reveal information in and about text
  - Algorithms discern patterns
  - Text may be “unstructured”
  - More than just search
- Everyday examples
  - Seeking out patterns in scientific literature
  - Identifying spam e-mail

How text analysis works (generally)

- Break textual data into smaller pieces
- Abstract (reduce) text so that a computer can crunch it
- Counting!
  - Words, phrases, parts of speech, etc.
- Computational statistics
  - Develop hypotheses based on counts of textual features

Non-consumptive research

Research in which computational analysis is performed on text, but not research in which a researcher reads or displays substantial portions of the text to understand the expressive content presented within it.

- Complies with copyright law
- Foundation of HTRC work
- Other terms: non-expressive use

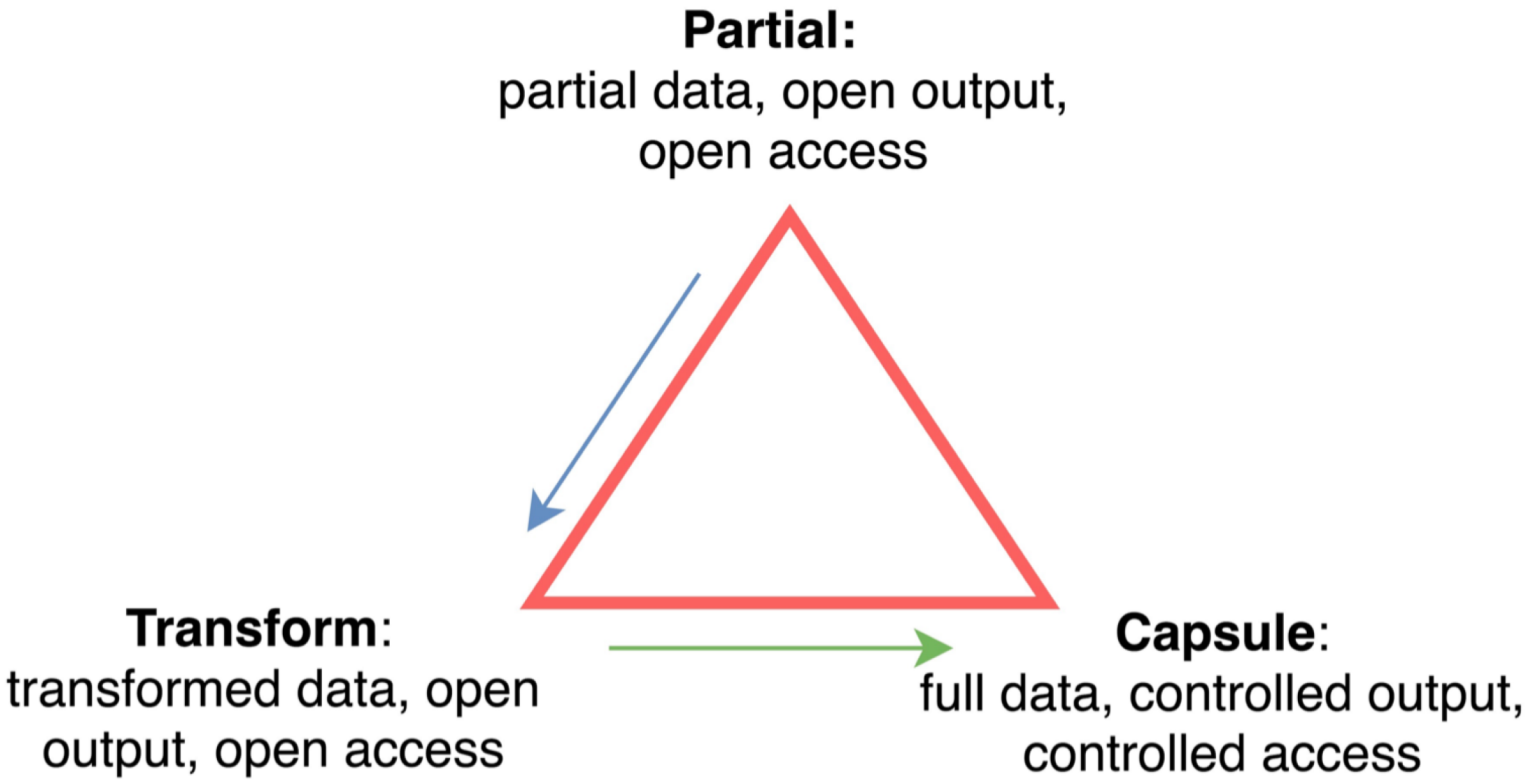
Non-consumptive paradigm

Includes such computational tasks as:

- text extraction
- textual analysis and information extraction
- linguistic analysis
- automated translation
- image analysis
- file manipulation
- OCR correction
- indexing and search

More here: https://www.hathitrust.org/htrc_ncup

HTRC's non-consumptive framework



Three Approaches

Partial

- Web-based tools: To analyze and visualize text data

Transform

- Derived Datasets: Including Extracted Features dataset

Capsule

- Secure Data Capsules: For flexible, self-directed research

HTRC Extracted Features

Derived Data

Transform

- Downloadable JSON dataset
- Per-volume files of metadata and data elements
- Periodically updated

Web tools

Off-the-shelf analysis and visualization tools

Partial

HathiTrust + Bookworm

HTRC Analytics

Data Capsule service

Secure compute environments

Capsule

- Researcher imports tools
- Access to data
- Export only derived results

feeds

Converted into

Feeds into

Feeds into

HathiTrust text corpus

synched nightly, lives at IU

Case studies

1. Inside the Creativity Boom (Samuel Franklin)

2. The Transformation of Gender in English-Language Fiction (Ted Underwood, David Bamman, and Sabrina Lee)

3. How Capitalism Changed American Literature (Dan Sinykin)

Discussion

- *What examples have you seen of text analysis?*
- *In what contexts do you see yourself using text analysis?*

  What about the researchers you support?

Text as data

In this section we will…

- Conceptualize text as data for analysis
- Explore HathiTrust as a source for textual data
- Get hands-on experience with HathiTrust+Bookworm
- Consider the data analyzed in the case studies

Humanities data

- Data is material generated or collected while conducting research
- Examples of humanities data:
  - Citations
  - Code/Algorithms
  - Databases
  - Geospatial coordinates

Can you think of others?

Definition adapted from: NEH, https://www.neh.gov/files/grants/data_management_plans_2018.pdf

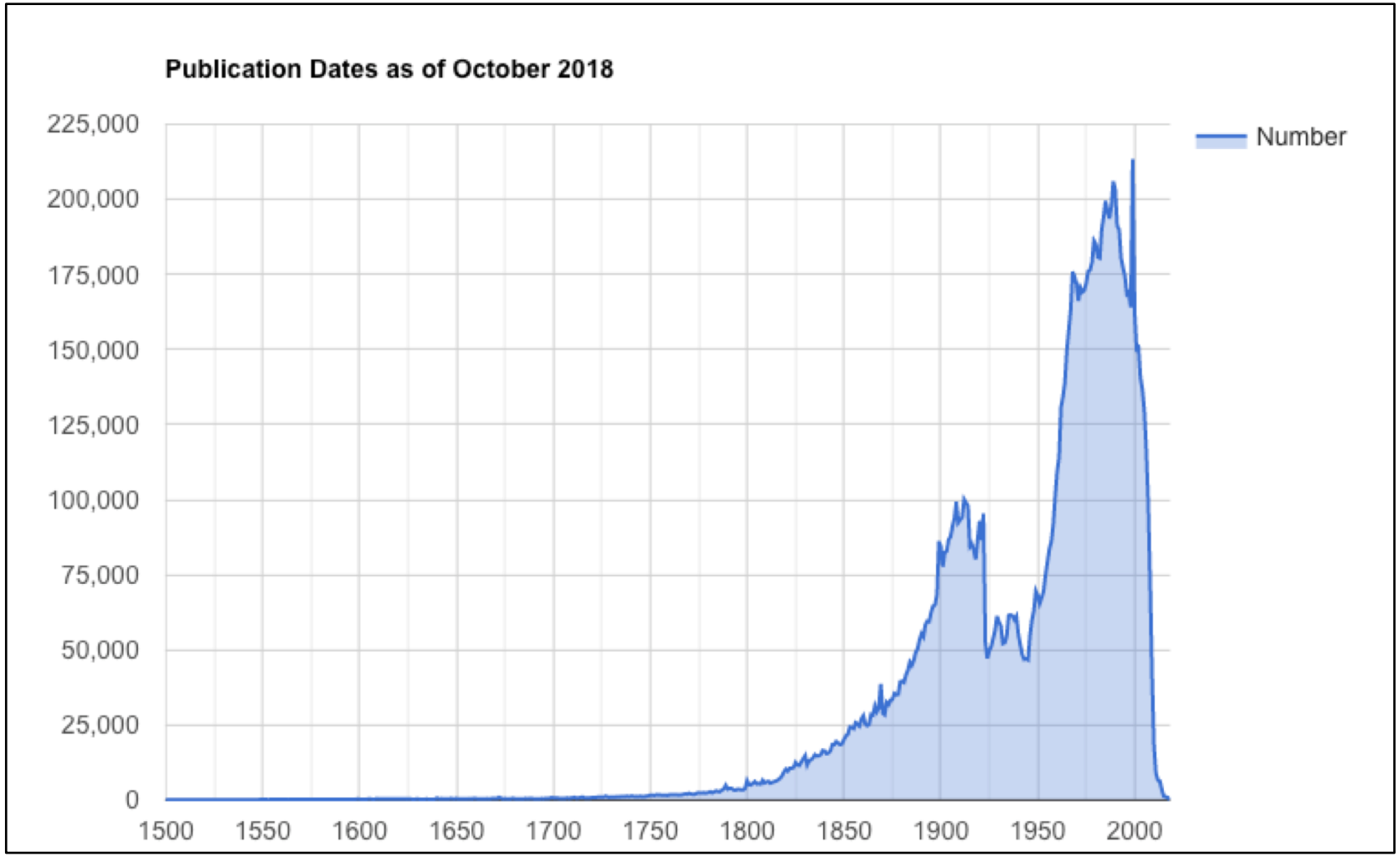
Text as data

- Data quality
  - Clean vs. dirty OCR
- Analyzed by corpus/corpora
  - Text corpus: a digital collection OR an individual's research text dataset
  - Text corpora: "bodies" of text

Data in HathiTrust

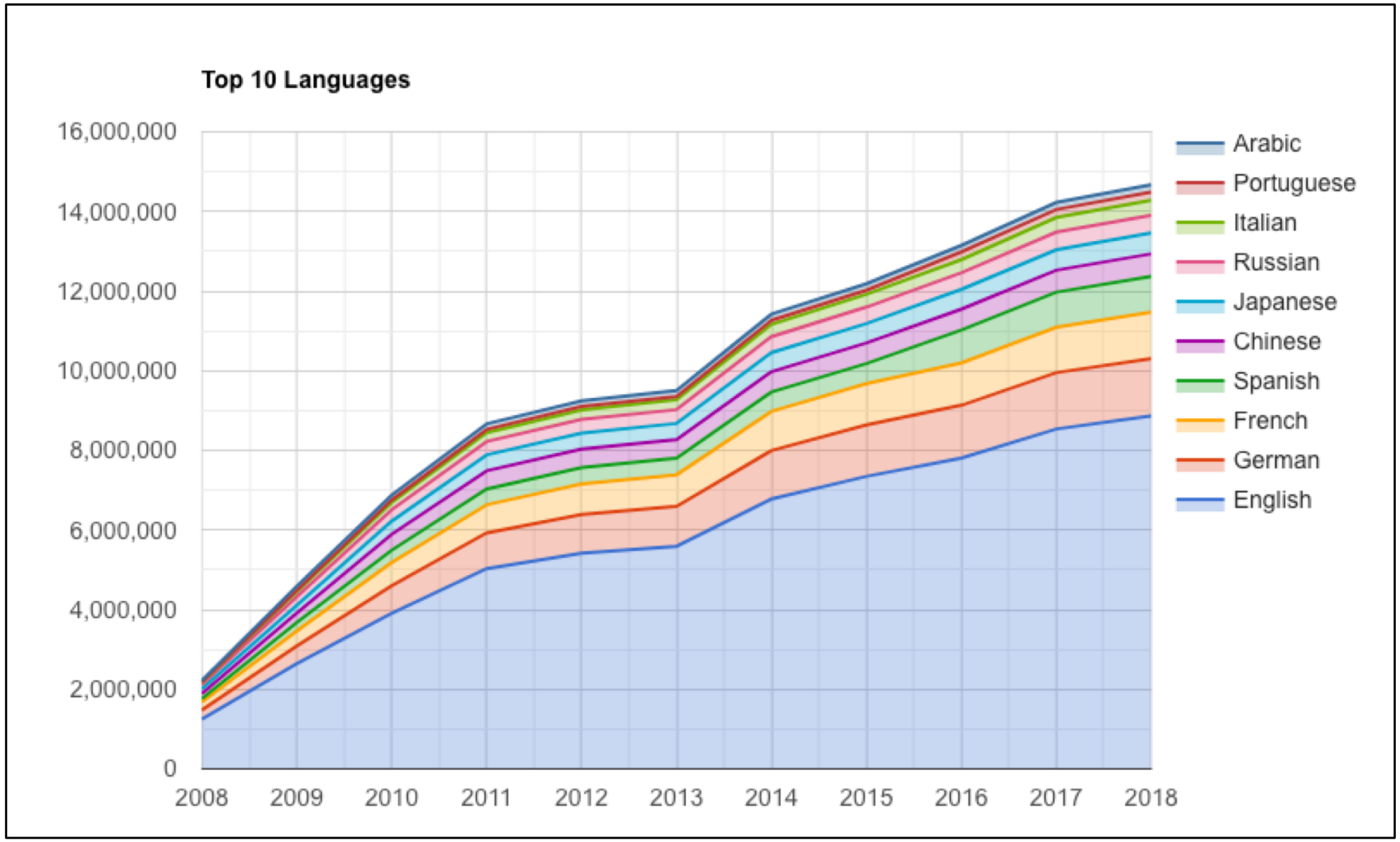
- Metadata
  - Primarily bibliographic (MARC) metadata
    - Example
  - Structural metadata (METS)
- Full text data
  - OCR text
  - Generated automatically during the digitization process
  - It’s dirty (uncorrected)

Publication dates in HTDL



https://www.hathitrust.org/snapshot-of-hathitrust-collection-10-years

Languages in HTDL



https://www.hathitrust.org/snapshot-of-hathitrust-collection-10-years

HathiTrust sources

Contributor	Volumes	Percentage
University of Michigan	4,727,160	31%
University of California	3,902,003	26%
Harvard University	841,981	6%
Cornell University	586,023	4%
University of Illinois at Urbana-Champaign	562,023	4%
University of Wisconsin - Madison	561,985	4%
Indiana University	530,775	4%
University of Minnesota	520,774	4%
The University of Texas	460,151	3%
Pennsylvania State University	390,364	3%
Remaining 42 Contributors	**1,993,400**	**13%**

Top 10 Contributors to HathiTrust as of March 2017

HathiTrust collection: US Federal Gov Docs

- 1,232,294 separate digital objects
- Fed Docs Registry: Attempt to id all US fed docs ever created
  - https://www.hathitrust.org/usdocs_registry
- Federal Documents collection description

  https://www.hathitrust.org/blogs/perspectives-from-hathitrust/federal_documents_collective_collection

Collection visualization: HathiTrust + Bookworm

Brings together:

- Text data (unigrams)
- Bibliographic metadata

HathiTrust

- Visualization tool
- Track trends in a repository

Bookworm

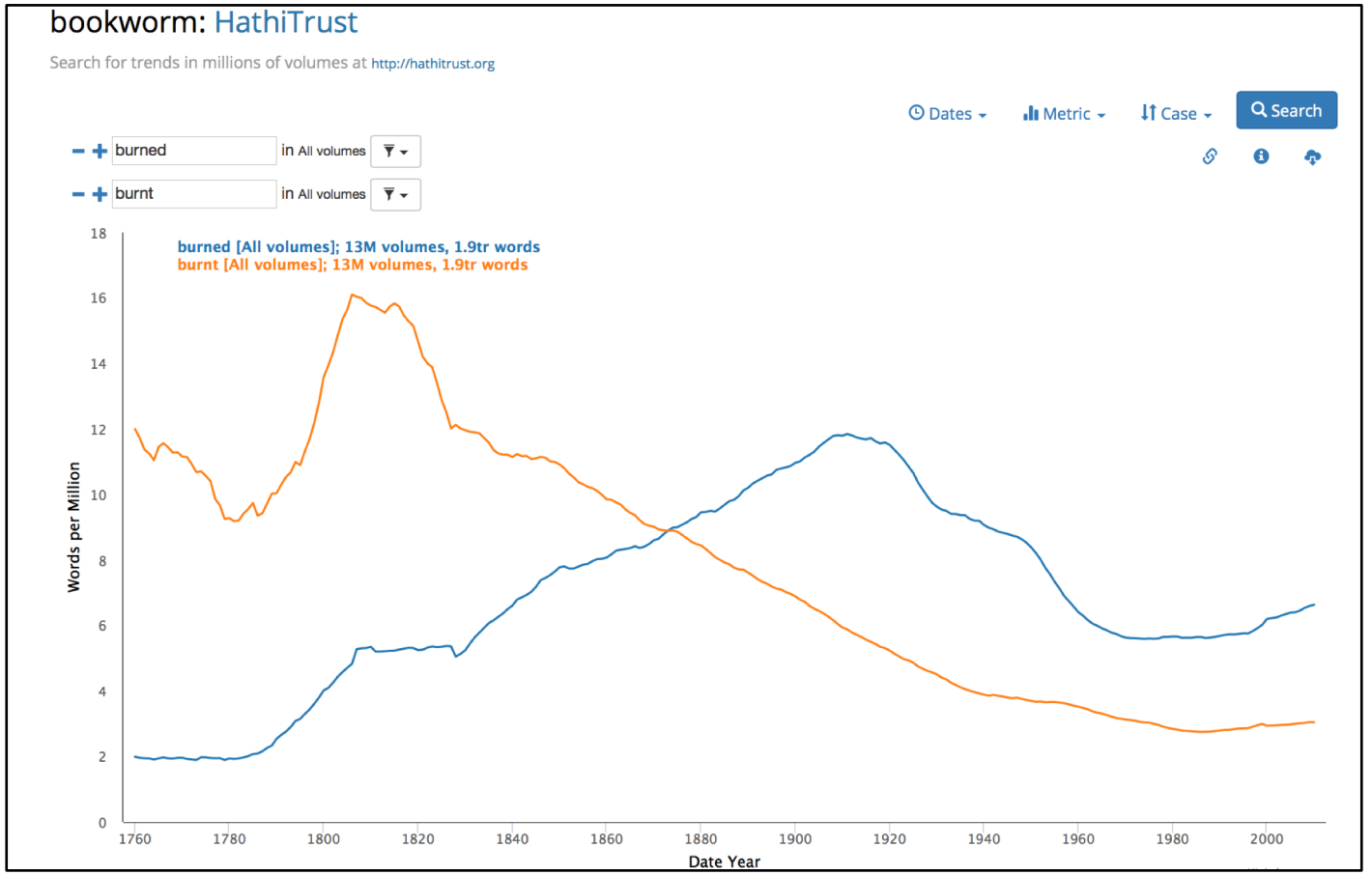
Bookworm visualization framework

- Visualizes categories
- The category is plotted along the x-axis
  - Often plot years along the x-axis
  - Can plot other things!
- HathiTrust+Bookworm is just one implementation of the framework

Adapted from Ben Schmidt, "Bookworm API Philosophy"

Reading an HT+BW graph

- Burned vs. Burnt



Do you see any trends?

Key terms in text analysis

N-gram

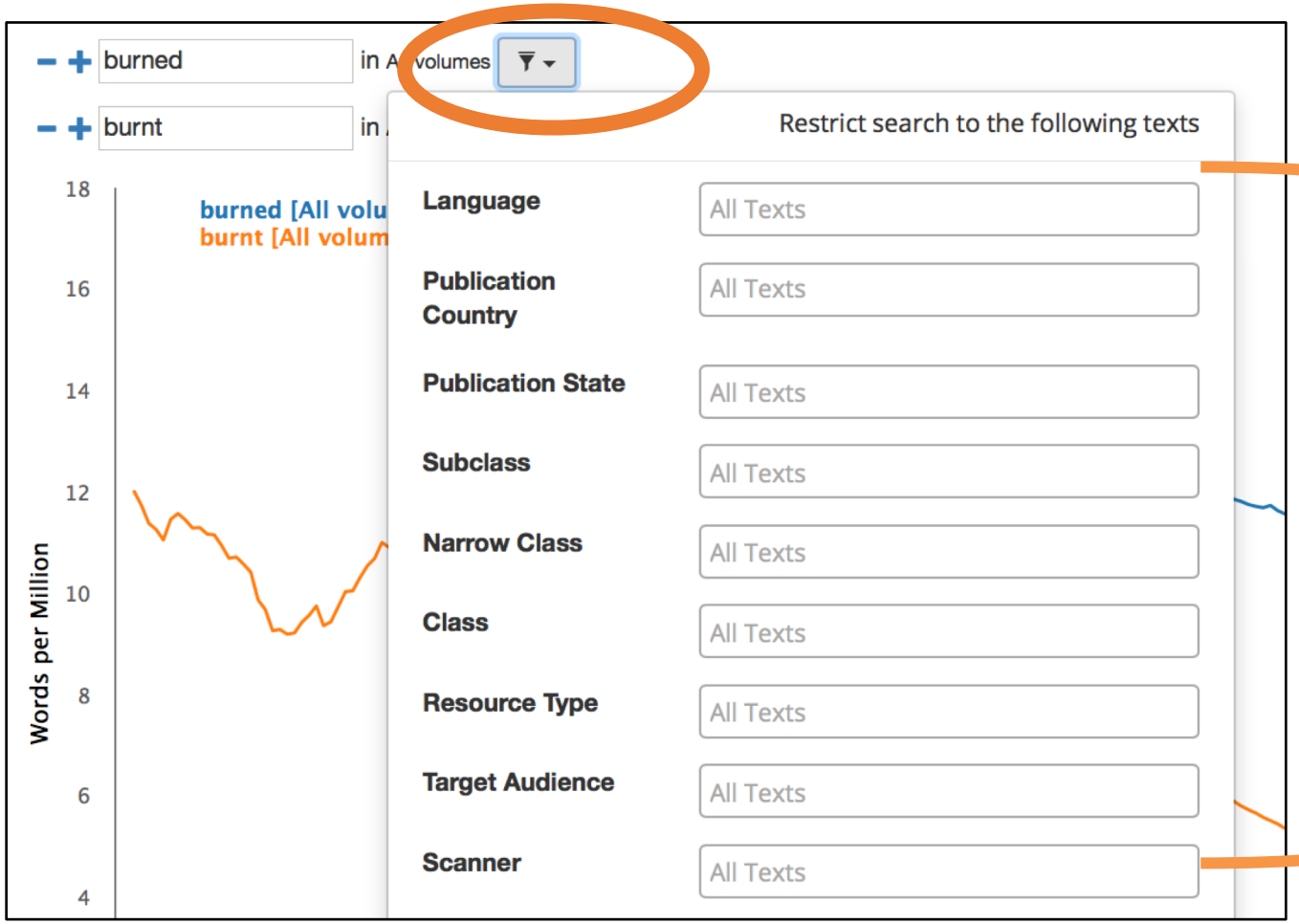
A contiguous chain of n items from a sequence of text where n is the number of items. Example: Bigram.

[it is], [is navigable], [navigable for], [for such], [such vessels], [vessels as], [as can], [can pass], [pass the], [the bar], [bar to], [to within], [within a] [a very], [very short], [short distance], [distance of], [of the], [the town], [town beyond], [beyond which], [which it], [it is], [is too], [too shallow], [shallow even], [even for], [for boats]

Bookworm functionality

- Search for unigrams (single words)
- Facet on metadata
- Adjust the years on the X-axis
- Visualize word trends across the corpus

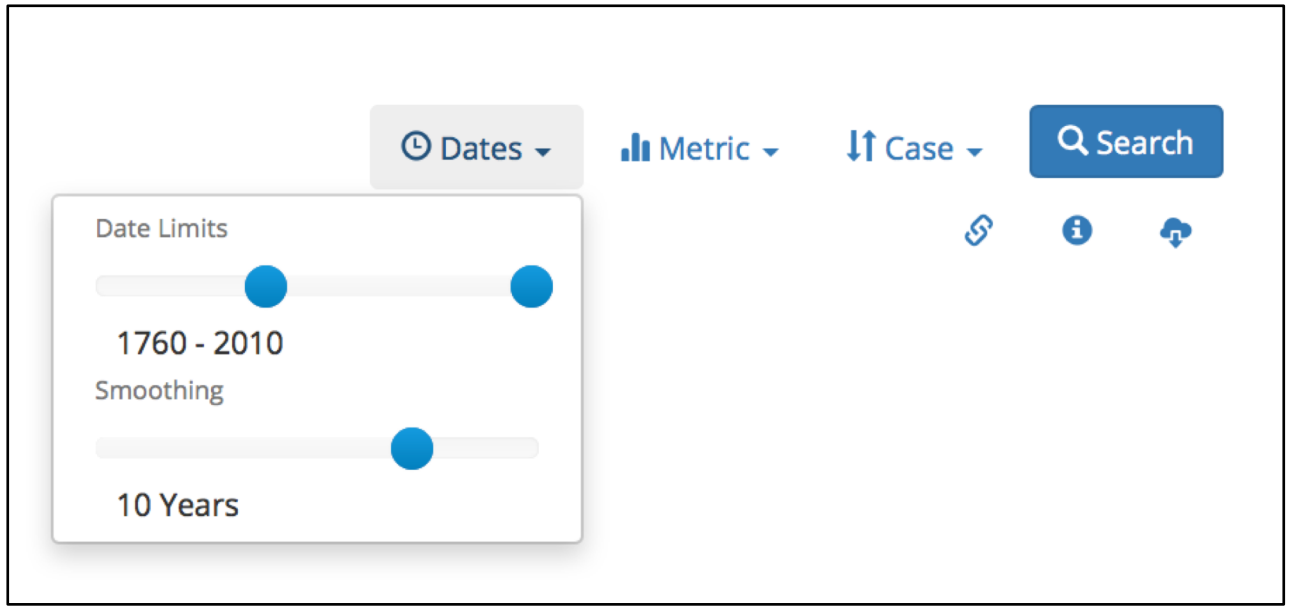
Bookworm interface



Limit your search with facets

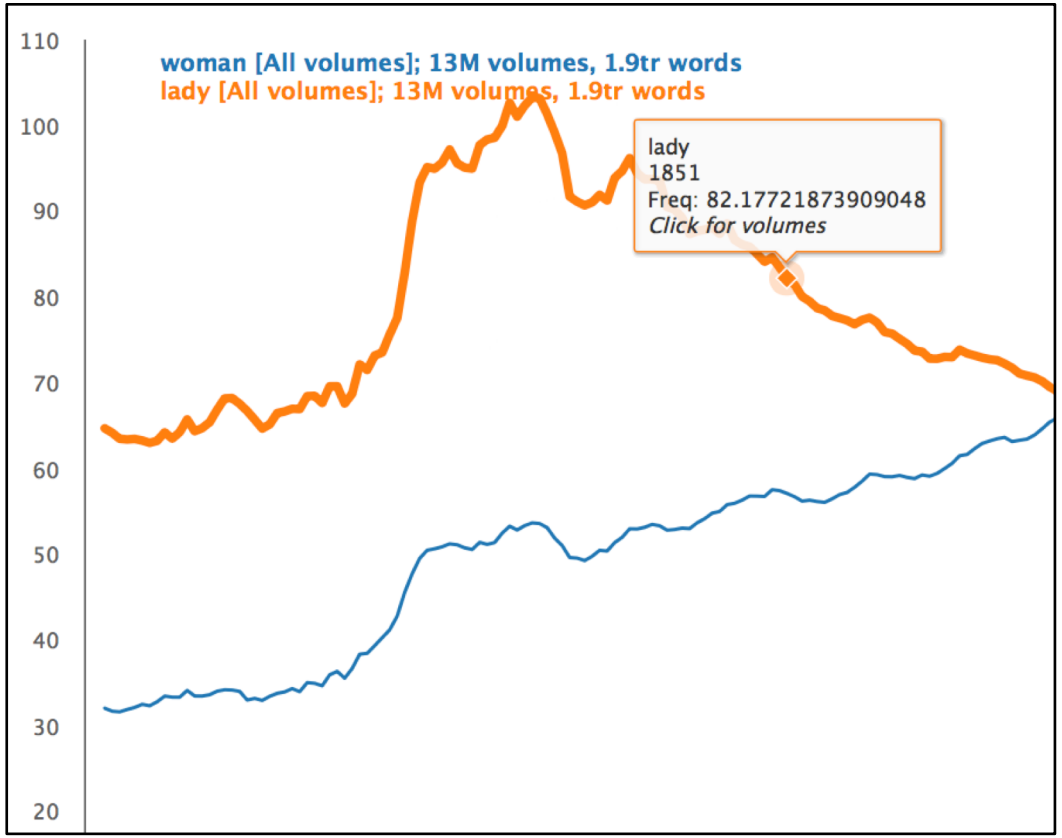
https://bookworm.htrc.illinois.edu/develop

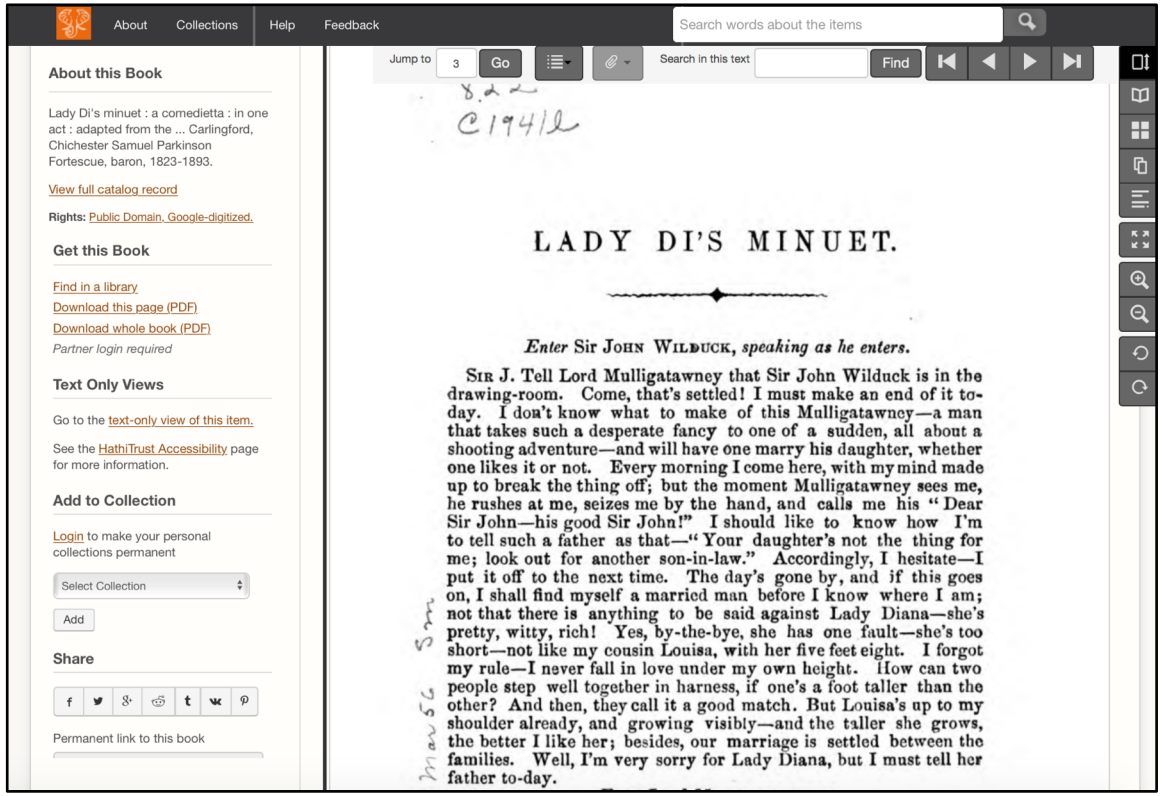
Bookworm interface



Fine-tune your results

Bookworm interface





Links directly to texts in the HTDL

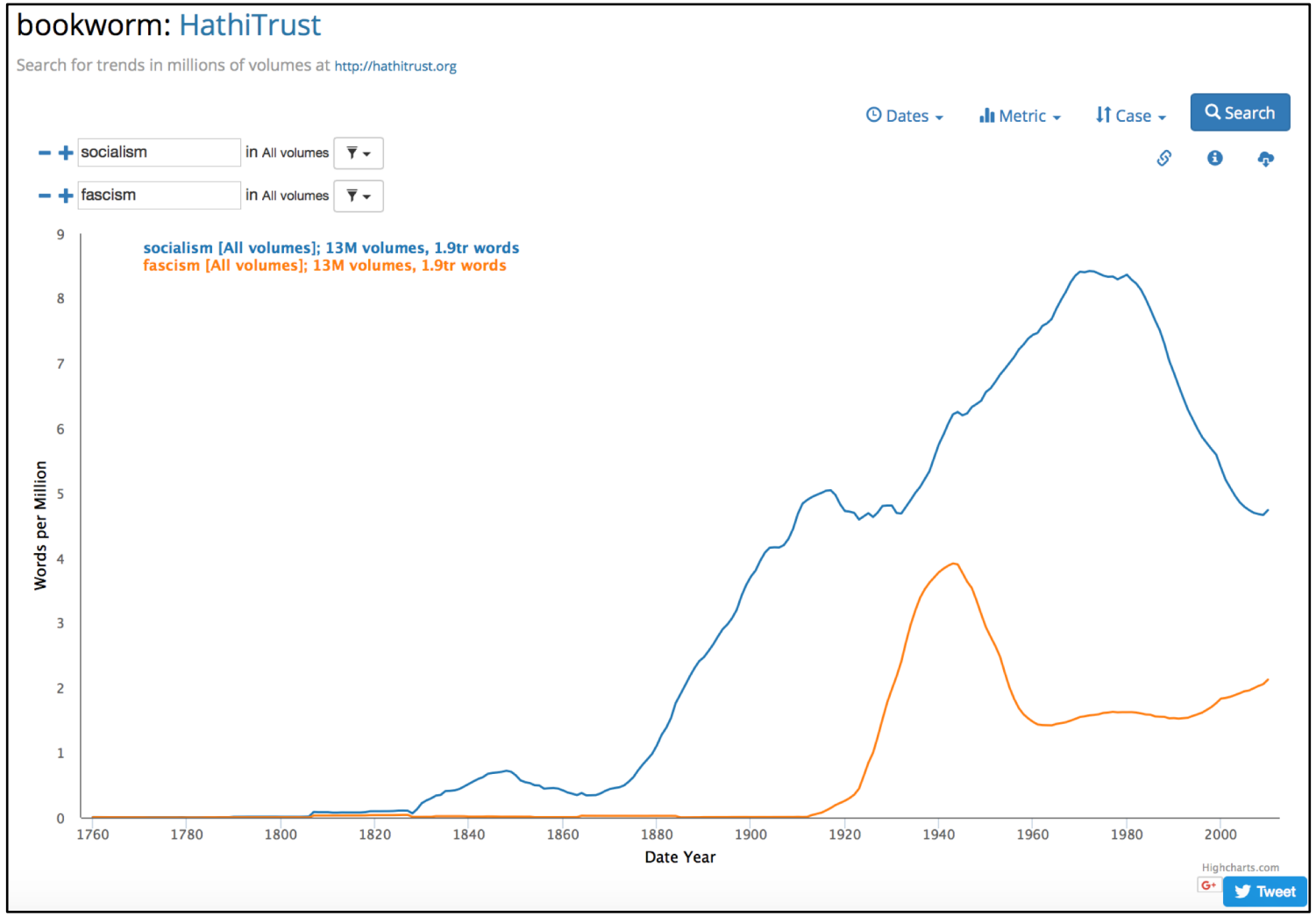
Hands-on activity

Handout p. 1

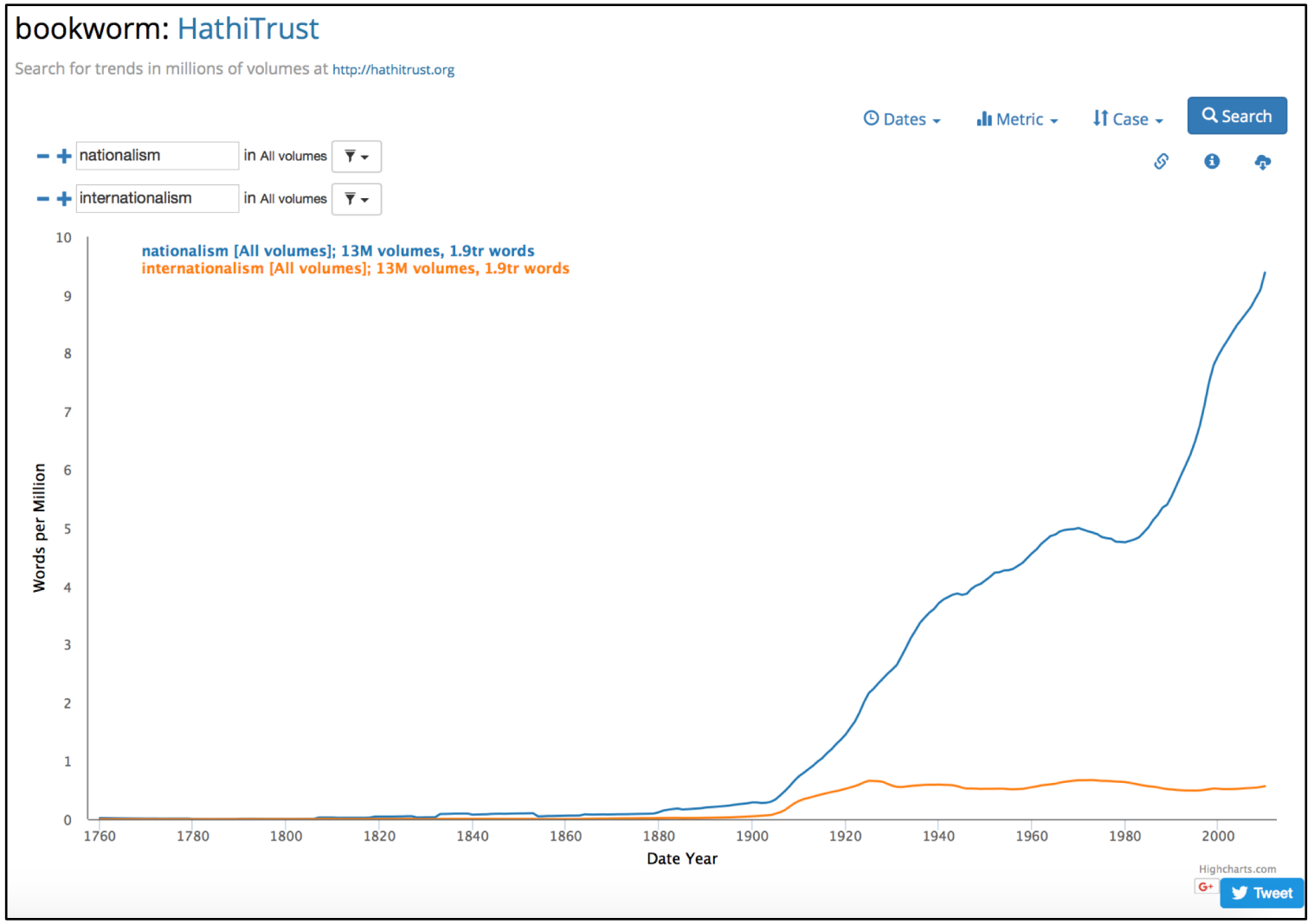
Use HT+BW to explore lexical trends

Website: https://bookworm.htrc.illinois.edu/develop

Examples


bookworm: HathiTrust
Search for trends in millions of volumes at http://hathitrust.org
Dates
Metric
Case
Search
socialism
in All volumes
fascism
in All volumes
socialism [All volumes]; 13M volumes, 1.9tr words
fascism [All volumes]; 13M volumes, 1.9tr words
Words per Million
0
1
2
3
4
5
6
7
8
9
1760
1780
1800
1820
1840
1860
1880
1900
1920
1940
1960
1980
2000
Date Year
Highcharts.com
Tweet

Examples


bookworm: HathiTrust
Search for trends in millions of volumes at http://hathitrust.org
Dates
Metric
Case
Search
nationalism
in All volumes
internationalism
in All volumes
nationalism [All volumes]; 13M volumes, 1.9tr words
internationalism [All volumes]; 13M volumes, 1.9tr words
Words per Million
0
1
2
3
4
5
6
7
8
9
10
1760
1780
1800
1820
1840
1860
1880
1900
1920
1940
1960
1980
2000
Date Year
Highcharts.com
Tweet

HathiTrust data access options

Method	Data	Description	Rights status	Restrictions
HT dataset request	Full text OCR	Download plain text OCR	Public domain	Depends on your university
HT Data API	Full text OCR, page images	Download page images and plain text OCR	Public domain	Non-Google digitized only
HTRC Algorithms	Full text OCR (not viewable)	Analyze a workset using off-the-shelf tools	All	Data can be computed on, but is not exposed
HTRC Extracted Features	Abstracted text and metadata	Download JSON files for each of 15.7 million volumes in HathiTrust	All	Data is preprocessed
HTRC Data API	Full text OCR	Analyze plain text OCR	All for HT members; else public domain	For use in a Data Capsule only

Building corpora

- Identify texts through full text search
  - Use a key term or phrase
- Identify texts through metadata
  - Date range, author’s names, etc
- Match to a list of known items, such as a bibliography

Building corpora

- Identify the best dataset available to the researcher, with minimal bias
- Techniques (Bode, 2019)
  - Statistical – primarily rely on sampling and statistical modeling to identify gaps and outliers to reduce bias
  - Scholarly – engage in a practice of historicizing the data prior to analysis to assess potential biases imposed by what is available, and in what format

Building corpora

- Process usually involves deduplication
- What to keep/discard is project dependent
- Examples of deduplication:
  - OCR quality
  - Earliest edition
  - Editions without forewords or afterwords

Case studies: Characterize the data used

Handout p. 2

1. Read all 3 case studies

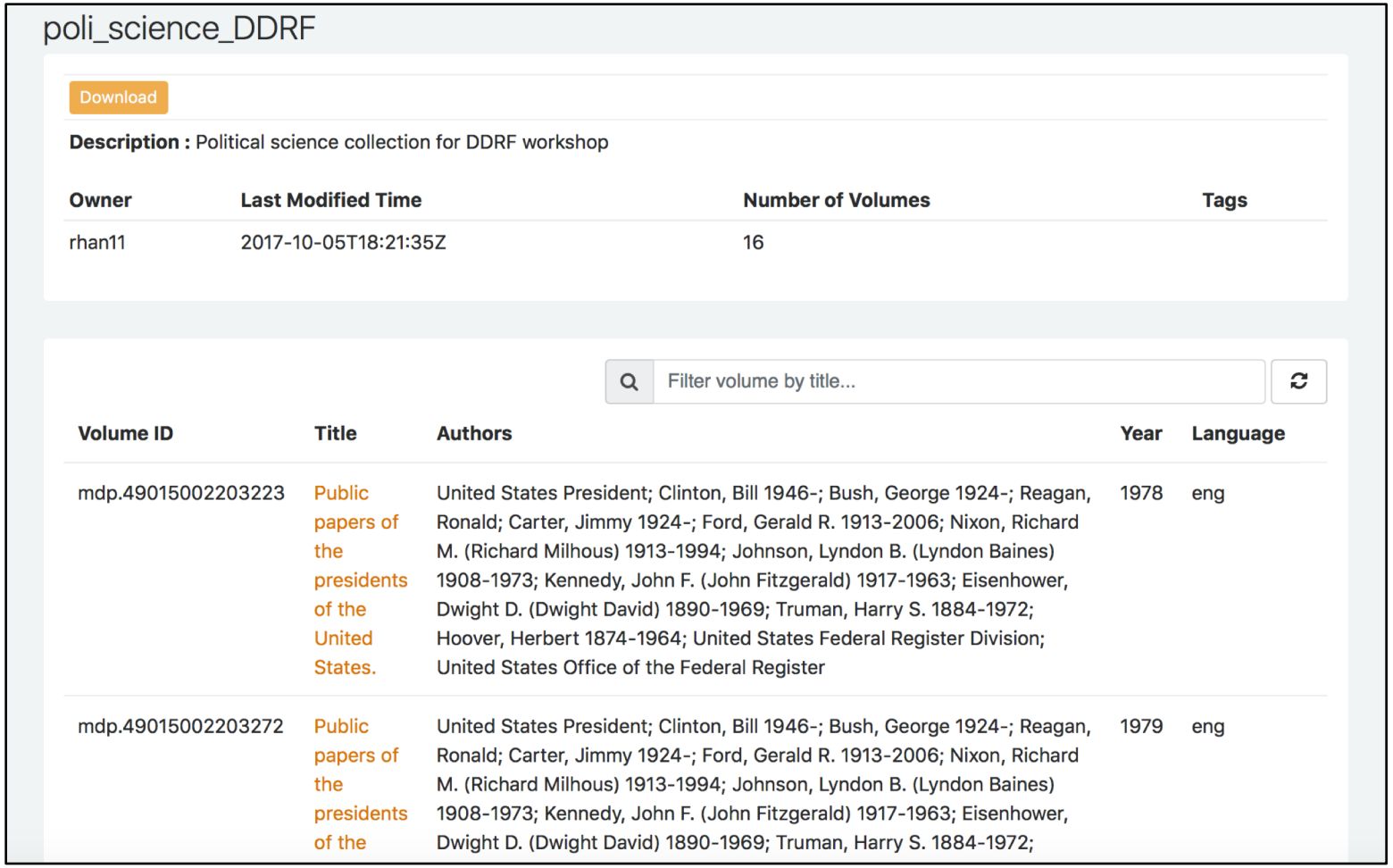
2. Then characterize the data used, such as:

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

HTRC Worksets

- User-created collections of text from the HathiTrust Digital Library
  - Think of them as textual datasets
- Can be shared and cited
- Suited for non-consumptive access

HTRC Worksets



poli_science_DDRF

Download

Description : Political science collection for DDRF workshop

Owner	Last Modified Time	Number of Volumes	Tags
rhan11	2017-10-05T18:21:35Z	16	

Filter volume by title...

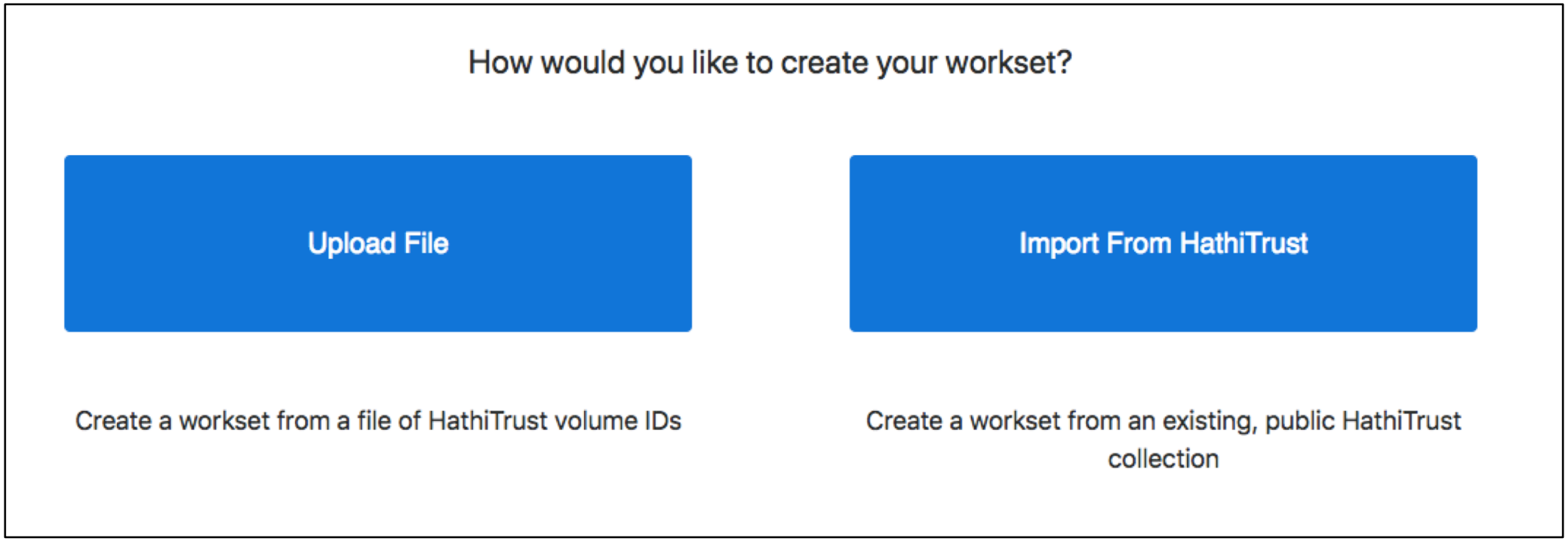
Volume ID	Title	Authors	Year	Language
mdp.49015002203223	Public papers of the presidents of the United States.	United States President; Clinton, Bill 1946-; Bush, George 1924-; Reagan, Ronald; Carter, Jimmy 1924-; Ford, Gerald R. 1913-2006; Nixon, Richard M. (Richard Milhous) 1913-1994; Johnson, Lyndon B. (Lyndon Baines) 1908-1973; Kennedy, John F. (John Fitzgerald) 1917-1963; Eisenhower, Dwight D. (Dwight David) 1890-1969; Truman, Harry S. 1884-1972; Hoover, Herbert 1874-1964; United States Federal Register Division; United States Office of the Federal Register	1978	eng
mdp.49015002203272	Public papers of the presidents of the	United States President; Clinton, Bill 1946-; Bush, George 1924-; Reagan, Ronald; Carter, Jimmy 1924-; Ford, Gerald R. 1913-2006; Nixon, Richard M. (Richard Milhous) 1913-1994; Johnson, Lyndon B. (Lyndon Baines) 1908-1973; Kennedy, John F. (John Fitzgerald) 1917-1963; Eisenhower, Dwight D. (Dwight David) 1890-1969; Truman, Harry S. 1884-1972;	1979	eng

Workset viewed on HTRC Analytics

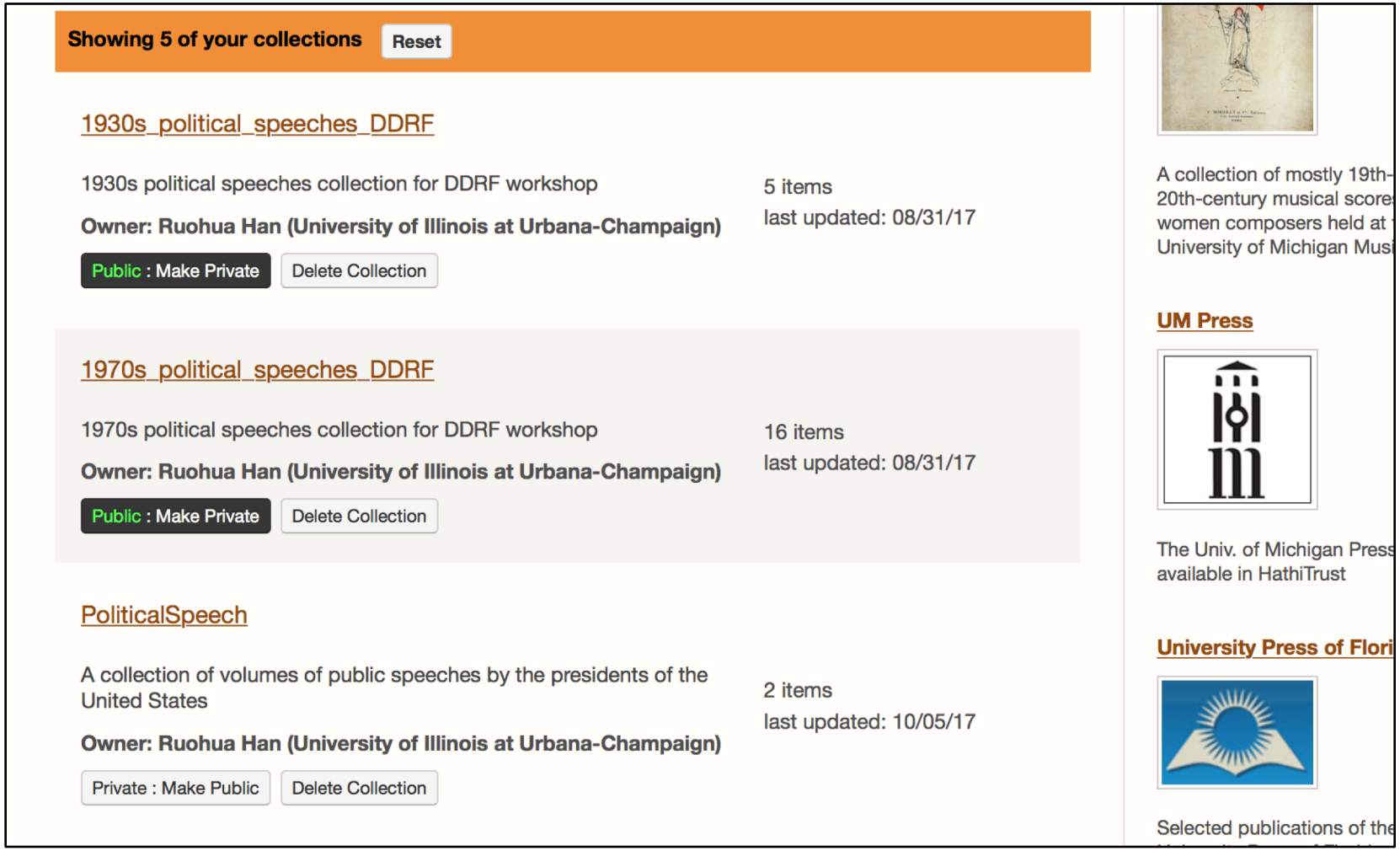
mdp.49015002221845
mdp.49015002221837
mdp.49015002221829
mdp.49015002221787
mdp.49015002221811
mdp.49015002221761
mdp.49015002221779
mdp.49015002203140
mdp.49015002203157
mdp.49015002203033
mdp.49015002203231
mdp.49015002203249
mdp.49015002203223
mdp.49015002203405
mdp.49015002203272
mdp.49015002203215

Workset manifest

Creating HTRC Worksets



HathiTrust user-created collections



To find, go to www.hathitrust.org

Then, click ‘Collections’ at top

Read and reflect

Handout p. 2

- Santa Barbara Statement on Collections as Data (Collections as Data National Forum, 2017)

  https://collectionsasdata.github.io/statement/

- Provides a set of high level principles to guide collections as data work

Read and reflect

Handout p. 2

- “With a few exceptions, cultural heritage institutions have rarely built digital collections or designed access with the aim to support computational use.”
- “Any digital material can potentially be made available as data that are amenable to computational use. Use and reuse is encouraged by openly licensed data in non-proprietary formats made accessible via a range of access mechanisms that are designed to meet specific community needs.”
- “Ethical concerns are integral to collections as data.”

-- Santa Barbara Statement on Collections as Data

Read and reflect

Handout p. 2

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

Break (20 minutes)

Research with text data

In this section we will…

- Recognize the steps taken to prepare text data
- Run the HTRC's Named Entity Recognizer
- Explore text analysis research questions

Moving beyond the book-like object

- Text decomposition/recomposition (Rockwell, 2003)
  - Cleaning data involves discarding data
  - Prepared text may be illegible to the human reader
- Text may be chunked or grouped
  - Chunking = dividing the text into smaller pieces (paragraphs, 1000 words, pages)
  - Grouping = combining smaller pieces of data together

Steps to prepare text data

- Correct OCR errors
- Remove title, header information
- Remove html or xml tags
- Split or combine files
- Remove certain words, punctuation marks
- Lowercase text
- Tokenize the words

Key concepts in text analysis

Tokenization

Breaking text into pieces called tokens. Often certain characters, such as punctuation marks, are discarded in the process

[it] [is] [navigable] [for] [such] [vessels] [as] [can] [pass] [the] [bar] [to] [within] [a] [very] [short] [distance] [of] [the] [town] [beyond] [which] [it] [is] [too] [shallow] [even] [for] [boats]

Hands-on Activity

Handout p. 3-4

- In groups of 2 or 3, assign each person several of the text preparation actions seen in the table to the right (Denny and Spirling, 2017).
- Read the descriptions. Then take turns explaining each to your group.

Term
Punctuation
Numbers
Lowercasing
Stemming
Stopword Removal
n-gram Inclusion
Infrequently Used Terms

Case studies: preparing text data

Handout p. 5

1. Look back at the *Creativity Boom* case study
2. Consider the following questions:
   - What steps did he take to prepare his data?
   - What assumptions did he make while preparing his data?

HTRC Named Entity Recognizer algorithm

Named Entity Recognizer (v2.0)

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.

How it works:

- performs header/body/footer identification
- extracts body text 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 language specified by the user
- performs entity recognition/extraction using the Stanford Named Entity Recognizer
- shuffles the entities found on each page (to prevent aiding page reconstruction)
- saves the resulting entities to a file

Result of job: table of the named entities found in a workset.

https://analytics.hathitrust.org/statisticalalgorithms

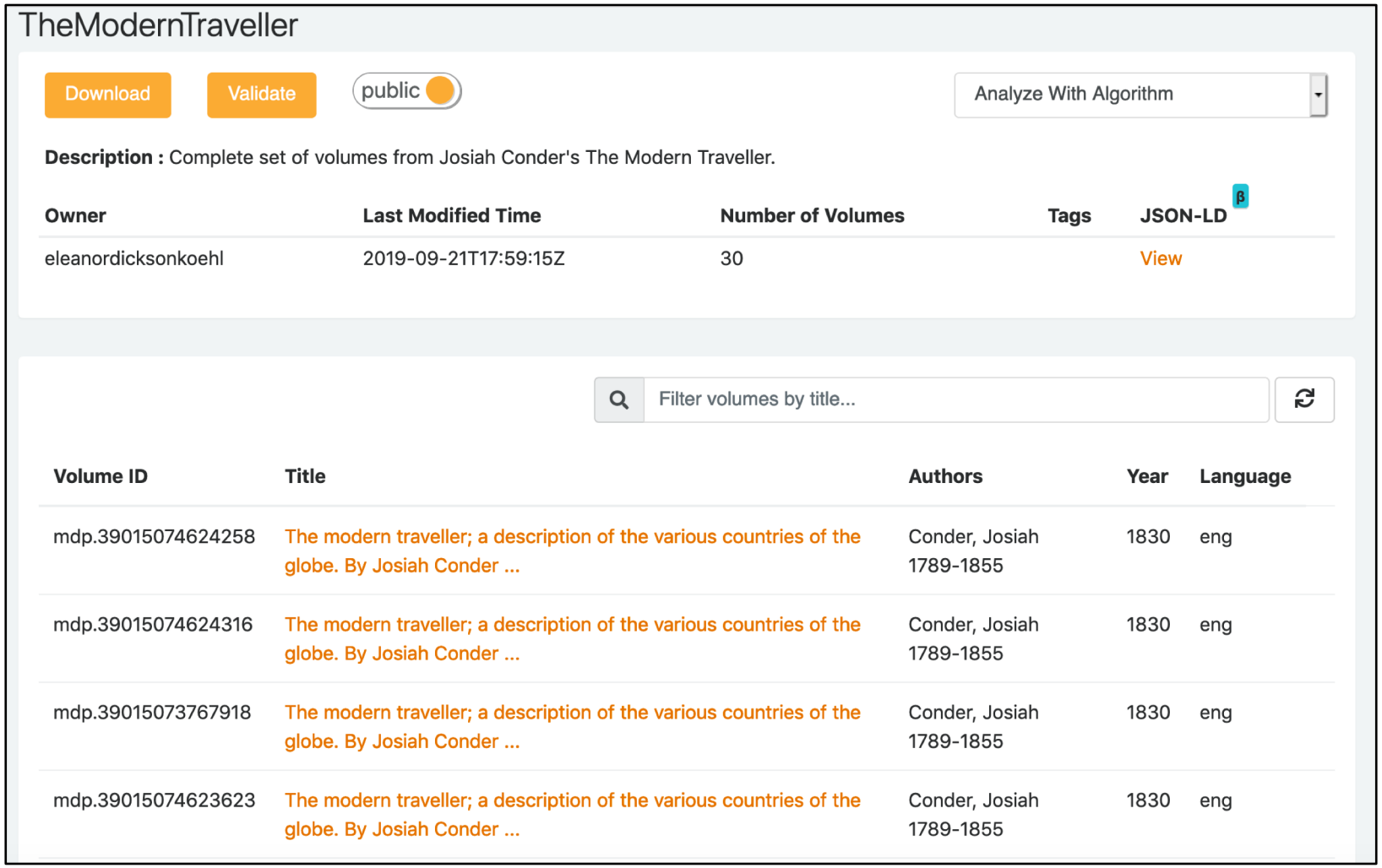
Hands-on activity

Handout p. 5

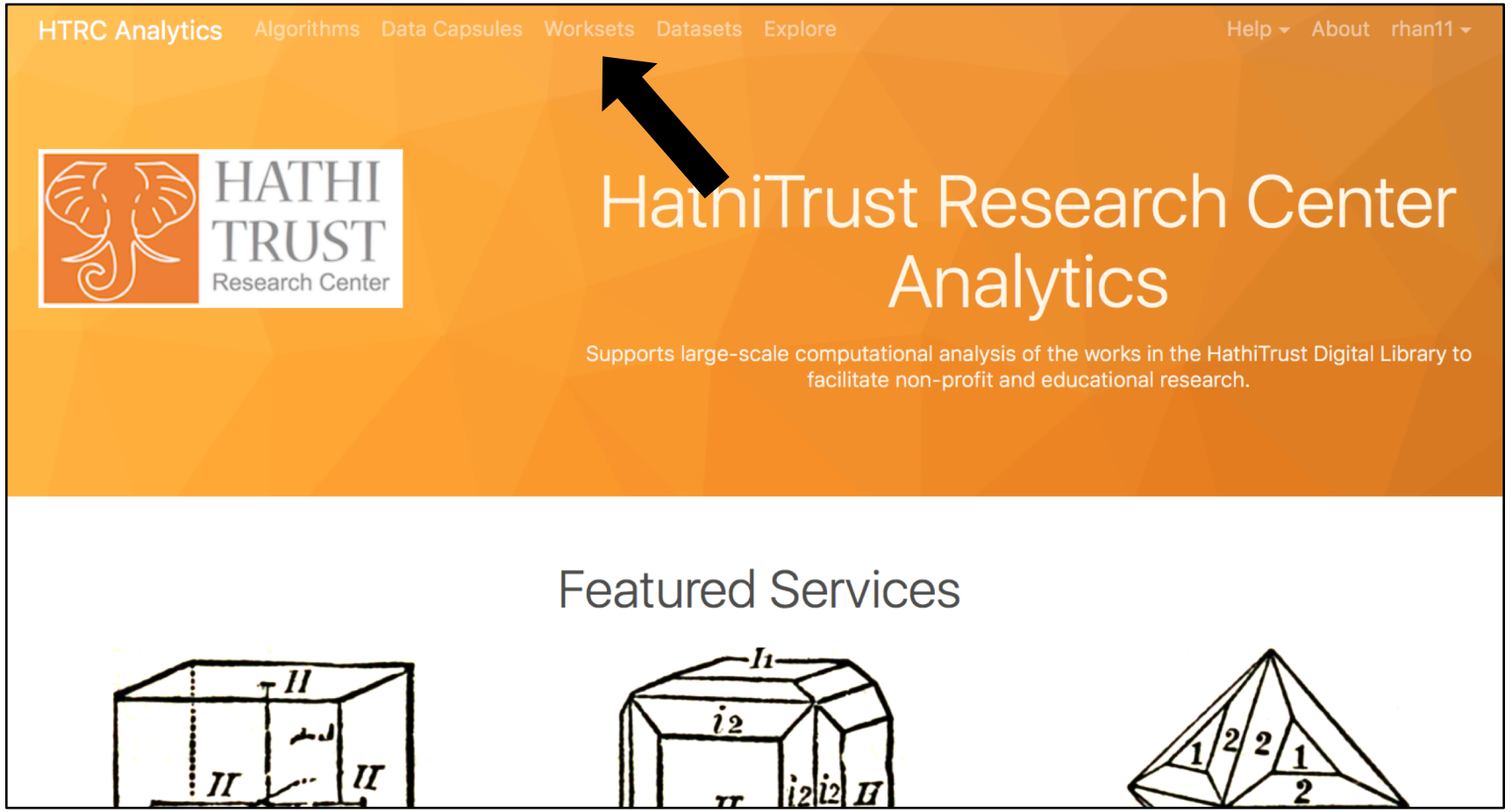
Run the HTRC's Named Entity Recognizer algorithm

Website: https://analytics.hathitrust.org/

The Modern Traveller workset

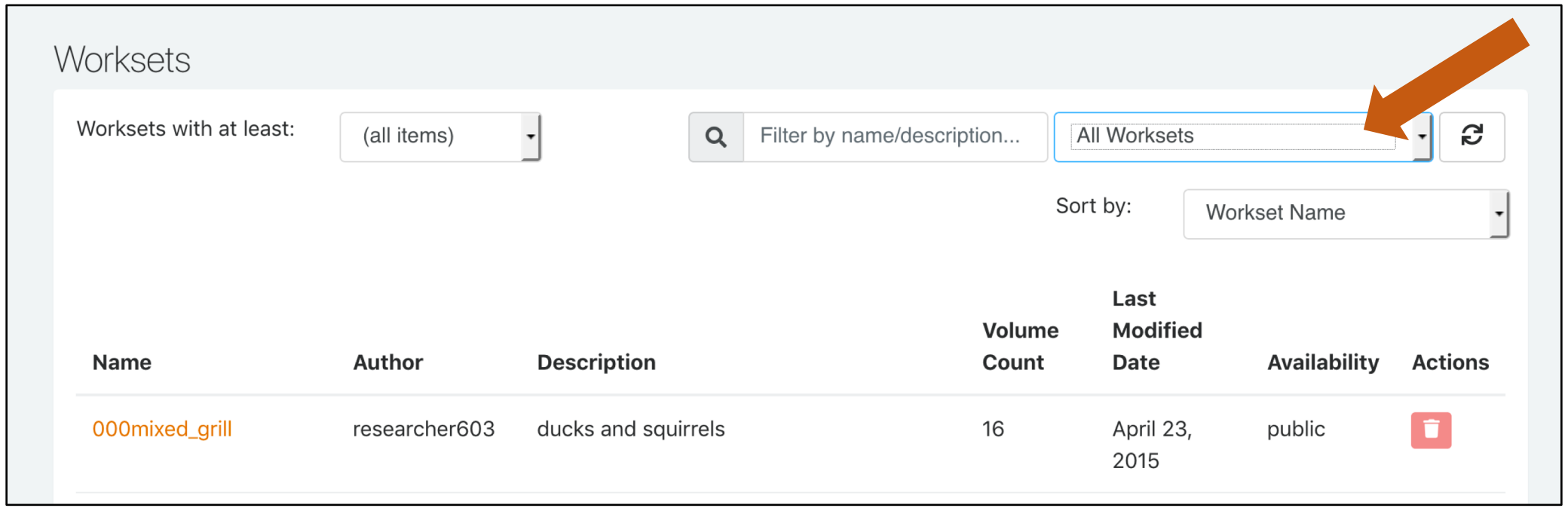


Navigate to the HTRC algorithms

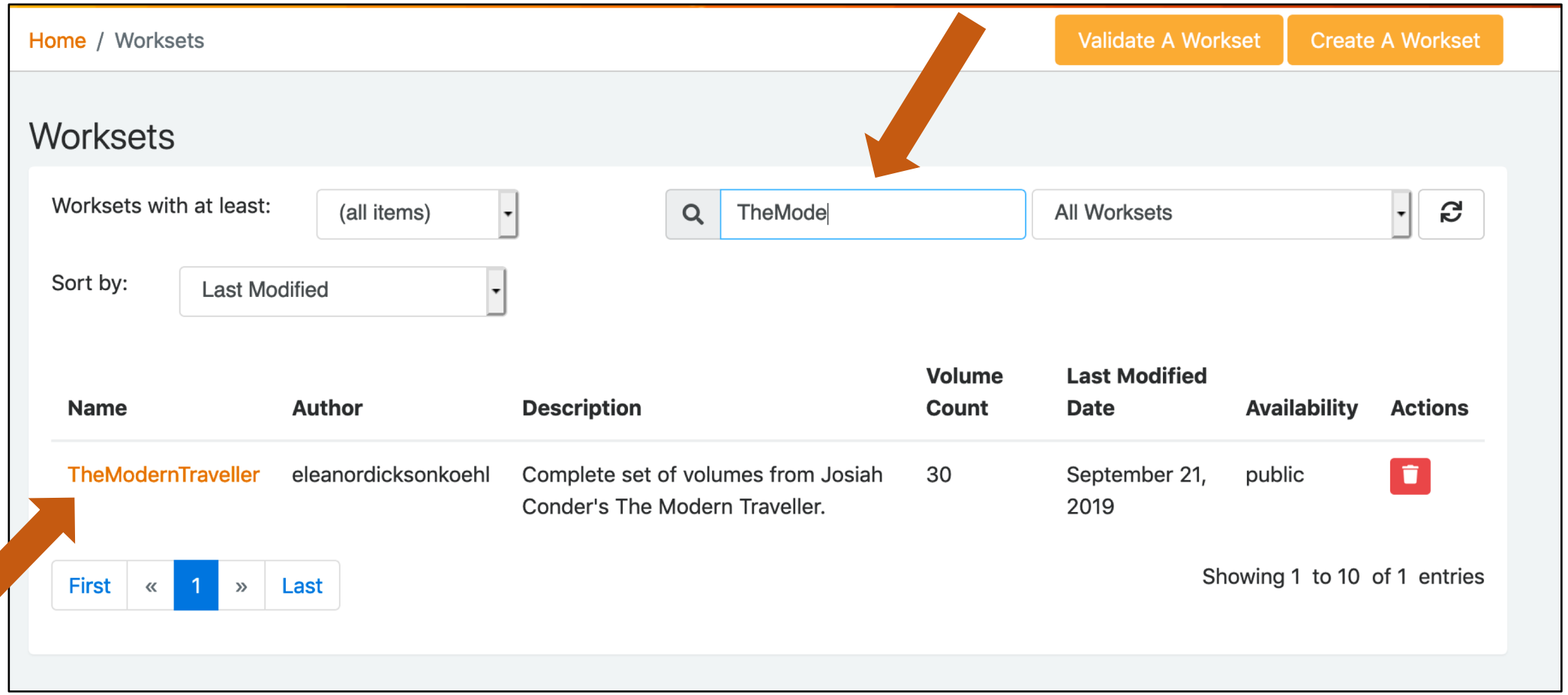


https://analytics.hathitrust.org

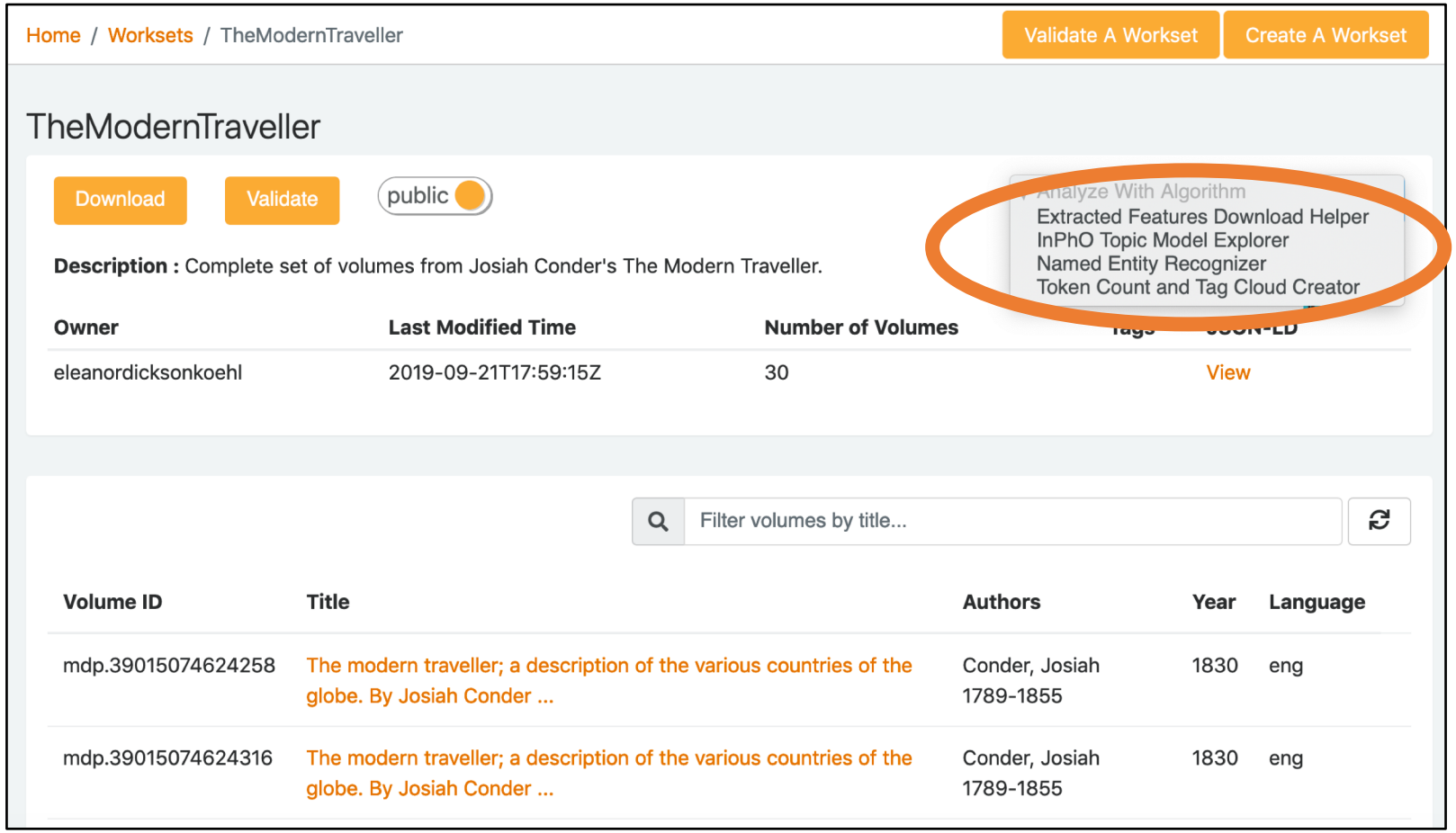
Find the workset



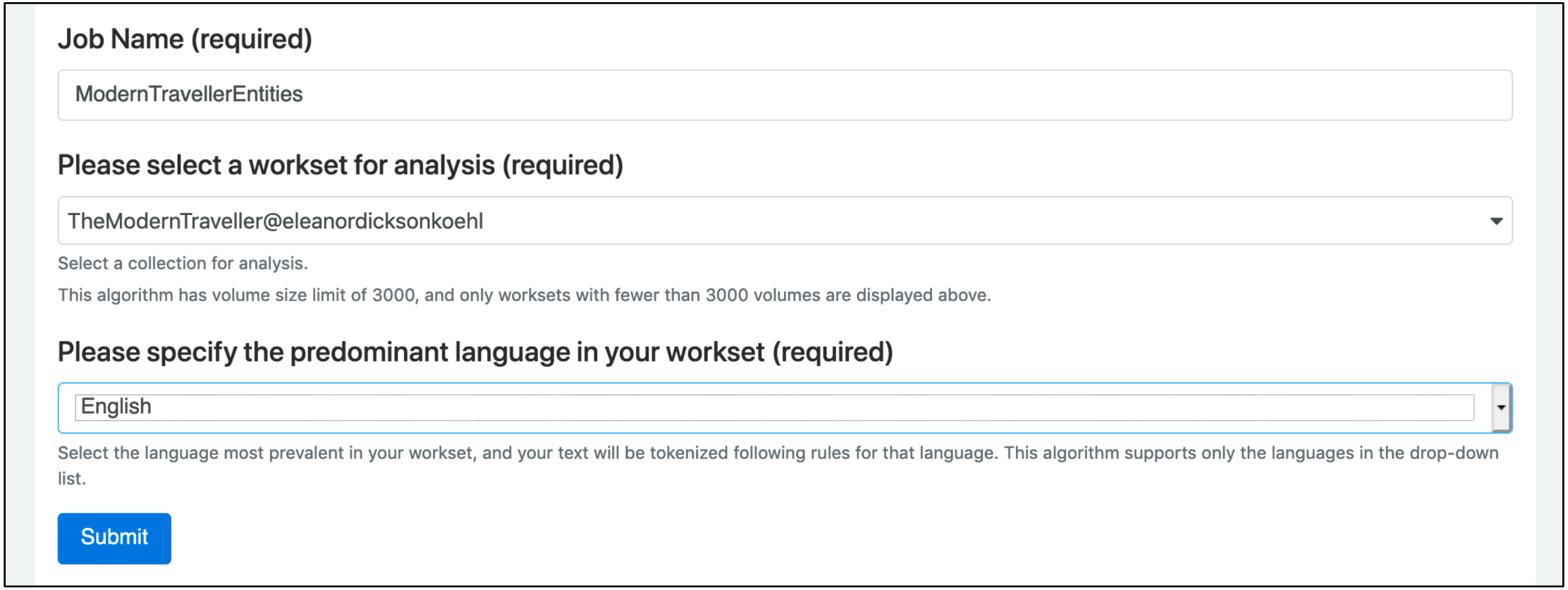
Filter by name & select the workset



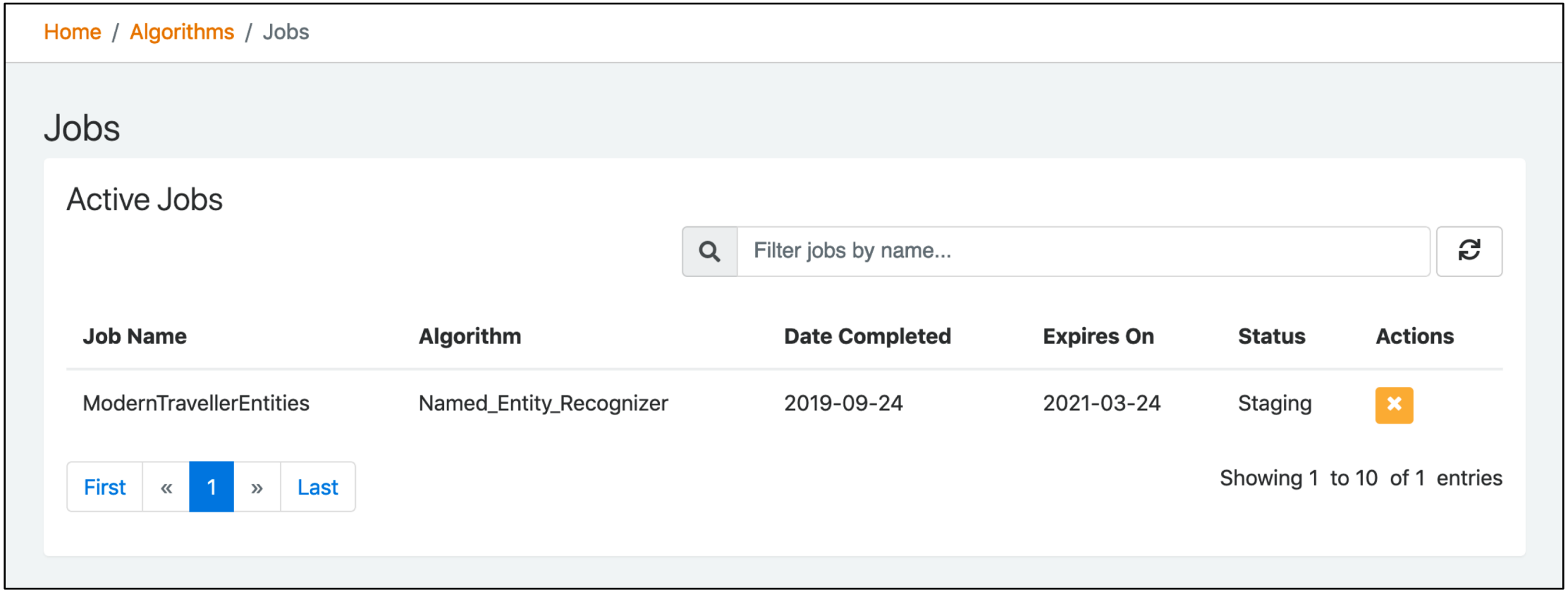
View workset & choose NER algorithm from menu



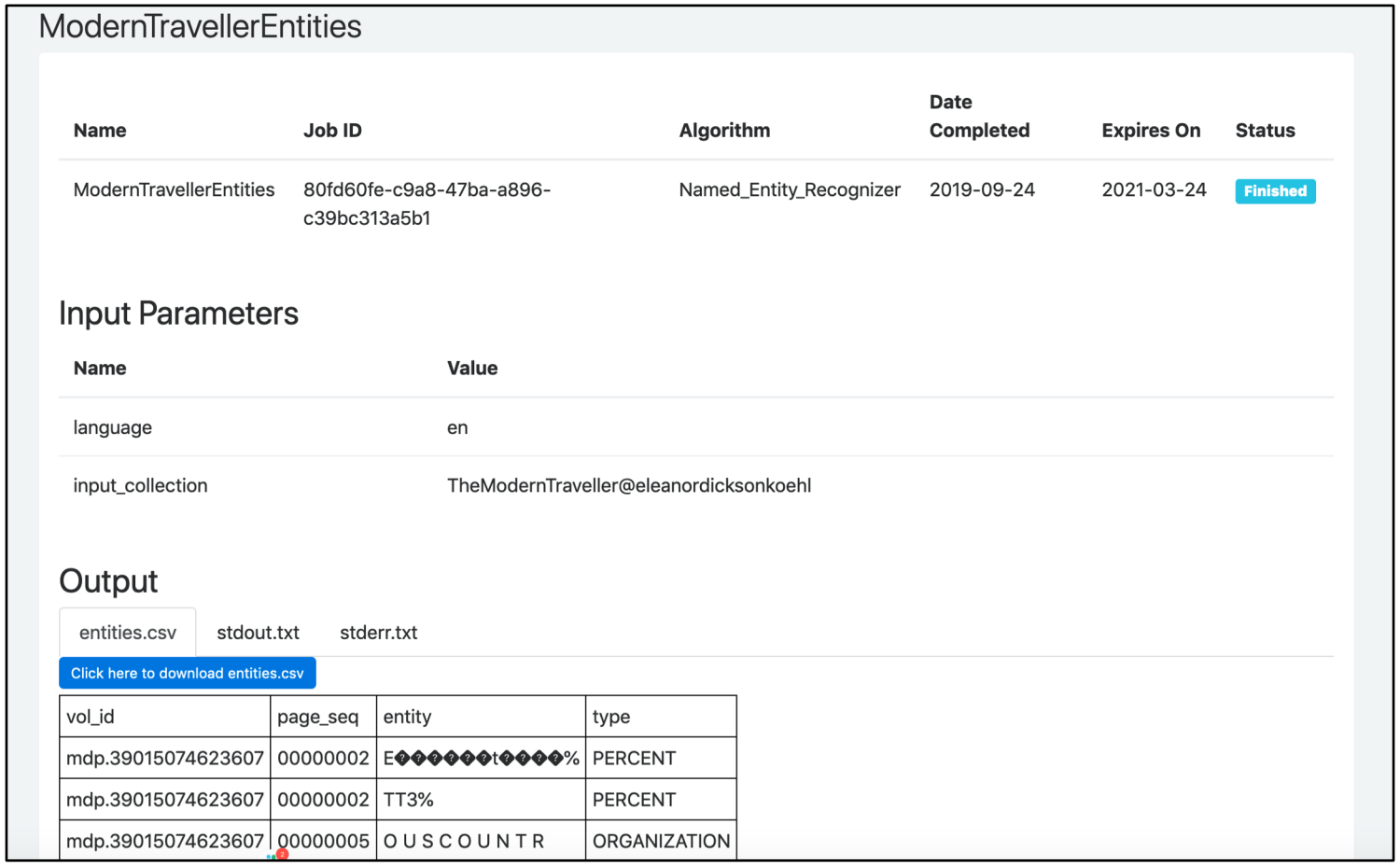
Input job details & submit



See your job in progress



View your results



ModernTravellerEntities

Name	Job ID	Algorithm	Date Completed	Expires On	Status
ModernTravellerEntities	80fd60fe-c9a8-47ba-a896-c39bc313a5b1	Named_Entity_Recognizer	2019-09-24	2021-03-24	Finished

Input Parameters

Name	Value
language	en
input_collection	TheModernTraveller@eleanordicksonkoehl

Output

entities.csv | stdout.txt | stderr.txt

Click here to download entities.csv

vol_id	page_seq	entity	type
mdp.39015074623607	00000002	E������t����%	PERCENT
mdp.39015074623607	00000002	TT3%	PERCENT
mdp.39015074623607	00000005	O U S C O U N T R	ORGANIZATION

Hands-on activity

Run 1 or more HTRC algorithms

- Use either TheModernTraveller workset or one of your choosing
- What does the algorithm do? What results do you get?

Website: https://analytics.hathitrust.org/

How does text analysis impact research?

- Shift in perspective, leads to shift in research questions
  - “Distant reading” (Moretti, 2013)
- One step in the research process
  - Can be combined with close reading
- Opens up:
  - Questions not provable by human reading alone
  - Larger corpora for analysis
  - Studies that cover longer time spans

Text analysis research questions

- May involve:
  - Change over time
  - Pattern recognition
  - Comparative analysis

Case studies: explore the research question

Handout p. 7

1. Return to the *How Capitalism Changed American Literature* case study
2. Answer the following questions:
   - What was his research question?
   - Why is it a good question for use of text analysis?

Lunch break

Text analysis methods

In this section we will…

- Learn about different text analysis methods
- Get started with Python via Jupyter Notebooks
- Map the geographic entities you generated earlier

How does Named Entity Recognition work?

- How does the computer know a name is a name?
- Not just search
  - Because the algorithm isn't looking for a list of possible entities, but any entity
- Uses semantic, context clues to determine entities
  - Sentence boundaries
  - Parts-of-speech

How text analysis works (generally)

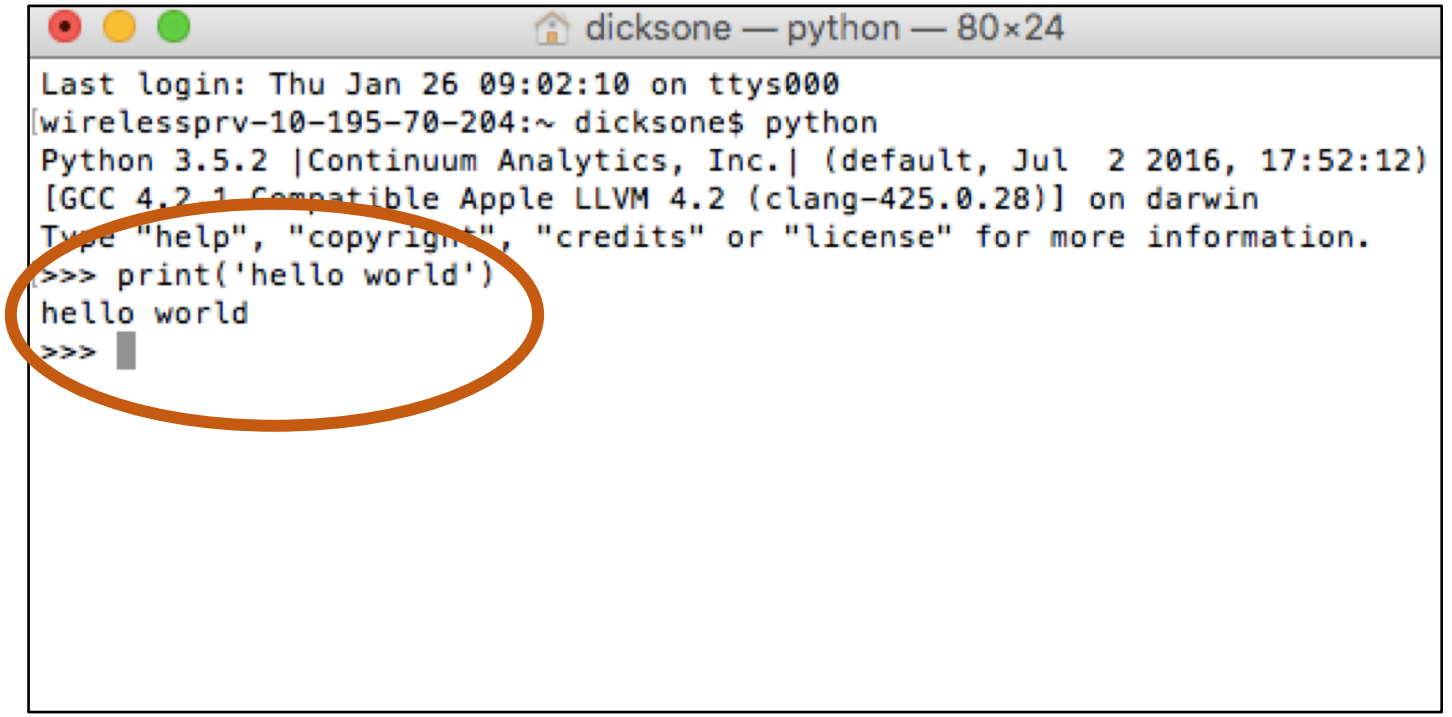
- Break textual data into smaller pieces
  - Chunking, tokenizing, or generating n-grams
- Abstract (reduce) text so that a computer can crunch it
- Counting!
  - Words, phrases, parts of speech, etc.
- Computational statistics
  - Develop hypotheses based on counts of textual features

Introduction to Python

- Python is a scripting language
- Good for working with data
  - Interpreted language → follows step-by-step directions
- Relatively straightforward syntax
  - Avoids excess punctuation

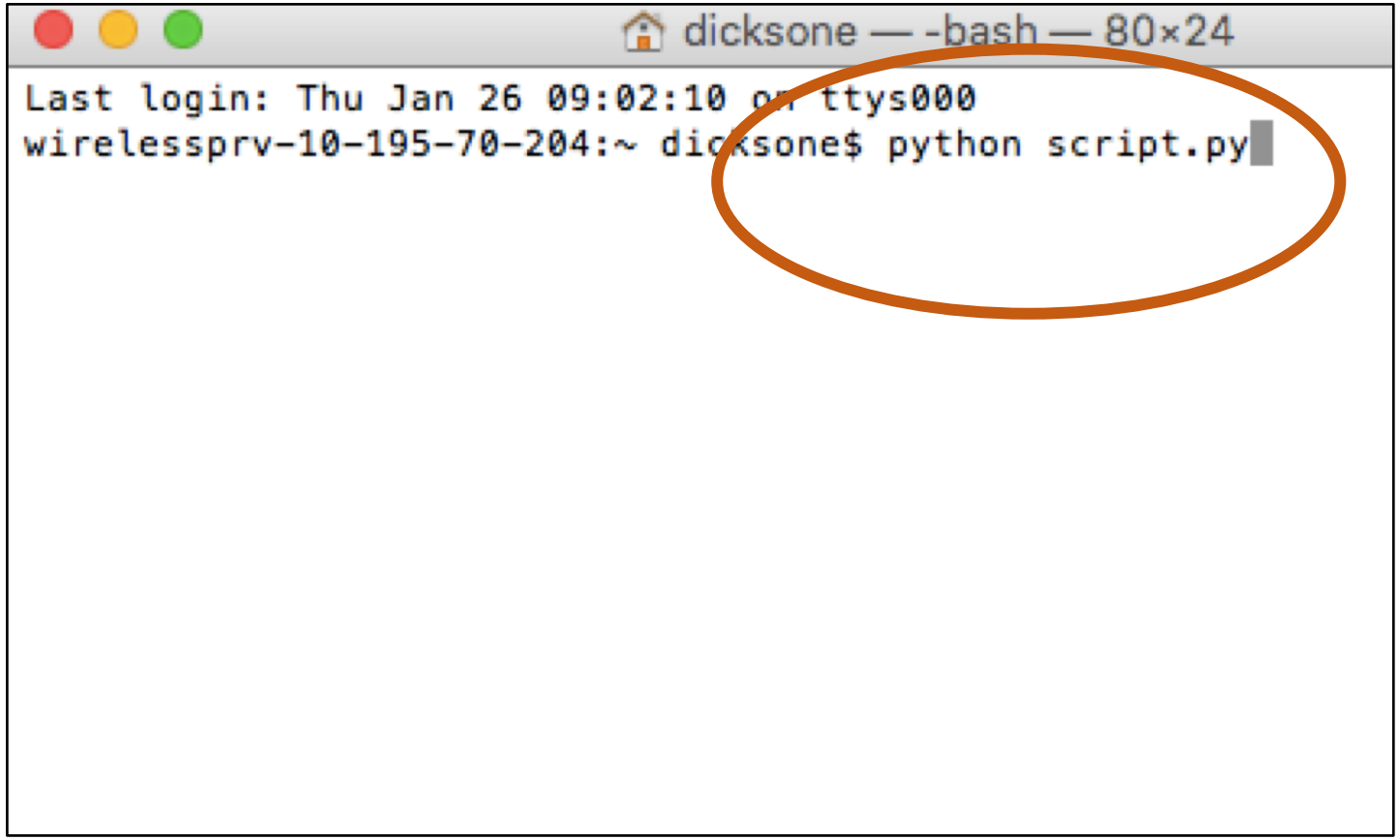
Using Python: Interactive programming

- Using a **python interpreter**
- Run each step at the prompt
- If you enter “Python” on the command line, you will start the interpreter
- *We aren’t using it today!*



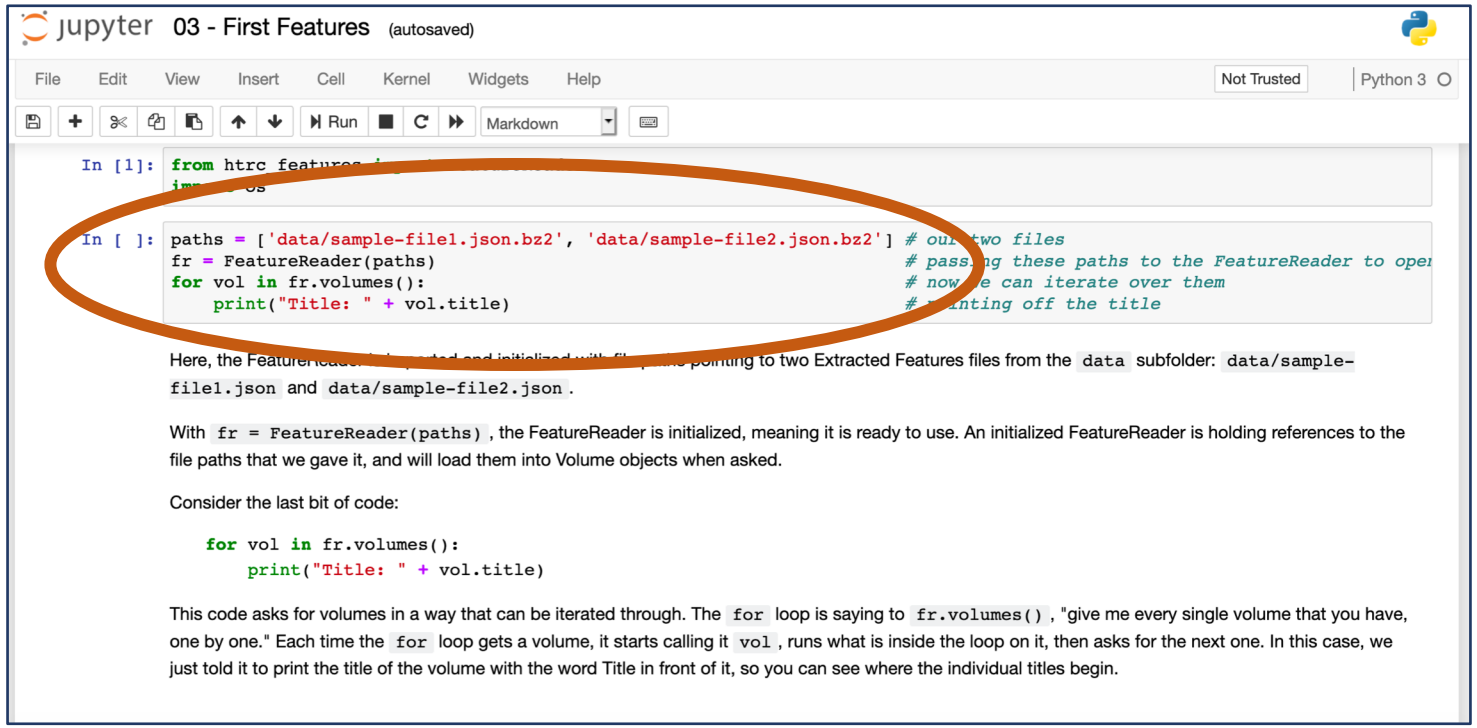
Using Python: Write & run scripts

- Scripts are directions for your computer to follow
- Save the script as a file ending in .py
- On the command line, run the script
- *Also not running this today!*



Using Python: Jupyter Notebooks

- Write and view code an an interactive environment
- Mixes Python (and other commands) with space for additional text
- Good for teaching and sharing code
- *We're using these today!*



Hands-on: Learning the Jupyter environment

Handout p. 7

Get comfortable with Jupyter notebooks with the “Learning the Environment” notebook

Website: https://go.illinois.edu/htrc-workshop

- Launch Binder
- 01-learning-jupyter-environment.ipynb

Hands-on: Map location entities

Handout p. 7

Geocode and map the location entities output from the HTRC Named Entity Recognizer.

Website: https://go.illinois.edu/htrc-workshop

- Launch Binder (if not already launched)
- 02-geocoding-module.ipynb

Key methods in text analysis

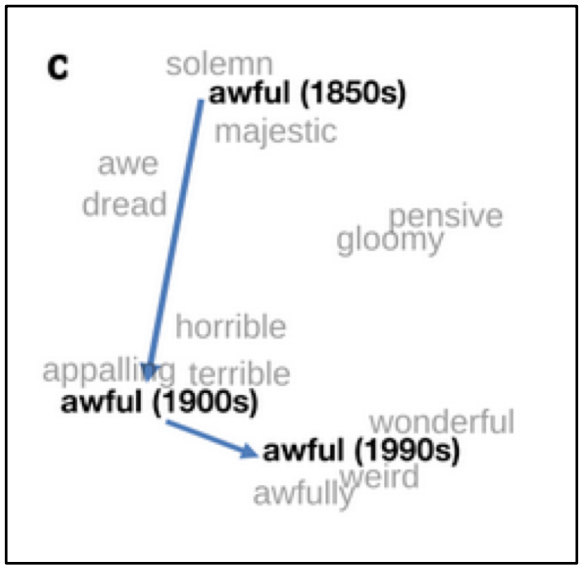
Representing text data – Bag of words

- Text is represented by the words and number of time they occurred, disregarding grammar and word order

Key methods in text analysis

Representing text data – Word vector embeddings

- Each word is represented by a number that reflects multi-dimensional information about its use in the text (i.e. to measure similarity between 2 words)



https://nlp.stanford.edu/projects/histwords/

Key methods in text analysis

Analyzing text data – Natural Language Processing

- Using computers to understand the meaning, relationships, and semantics within human-language text
- **Named entity extraction:** what names of people, places, and organizations are in the text?
- **Sentiment analysis**: what emotions are present in the text?
- **Stylometry**: what can we learn from measuring features of style?

Key methods in text analysis

Analyzing text data – Machine Learning

- Training computers to recognize patterns
- **Topic modeling:** What thematic topics are present in the text?
  - Unsupervised machine learning
- **Naïve Bayes classification:** Which of the categories that I have named does the text belong to?
  - Supervised machine learning

Case studies: identify the methods

Handout p. 7

1. Look back at all 3 case studies

2. Then characterize the methods used:

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

Break (20 minutes)

Text analysis workflows

In this section we will…

- Get experience with HTRC Extracted Features
- Perform exploratory data analysis
- Examine a text analysis workflow/pipeline

The toolkit

- Researcher-dependent
- Requires understanding of statistics
- Often draws on expert collaborators
- Consists of command line tools and programming languages

Exploratory data analysis

- Approach for getting familiar with data
- Easier to recognize patterns (and problems)
  - Hard to see trends in a spreadsheet or text file!
- There are whole books about it
- Strategies:
  - Plot raw data
  - Plot simple statistics
  - Compare plots to look for patterns

Features in the HTRC

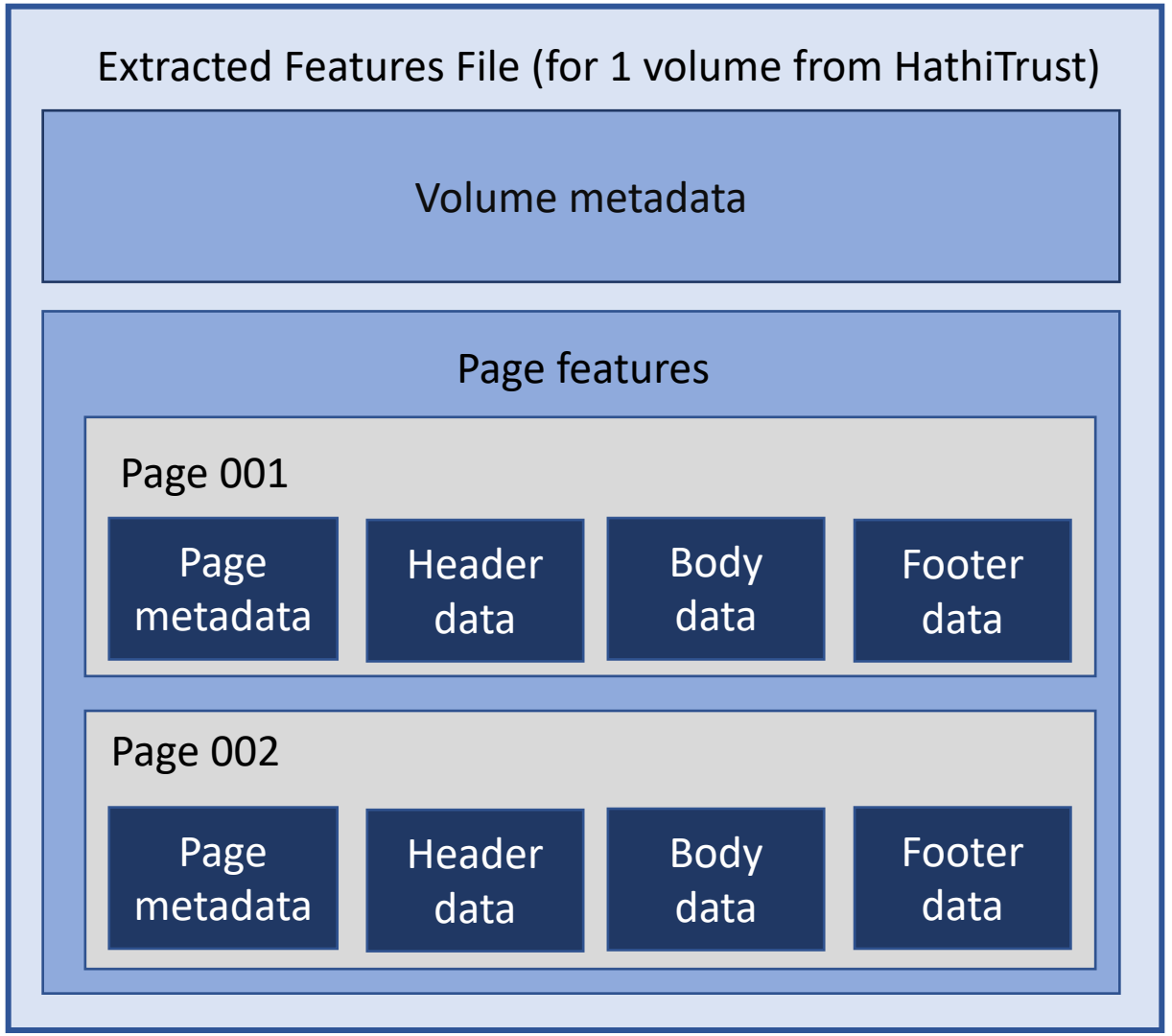
- HTRC Extracted Features dataset
- Downloadable
- Structured data consisting of features
- From 15.7 million volumes

- https://analytics.hathitrust.org/datasets#ef

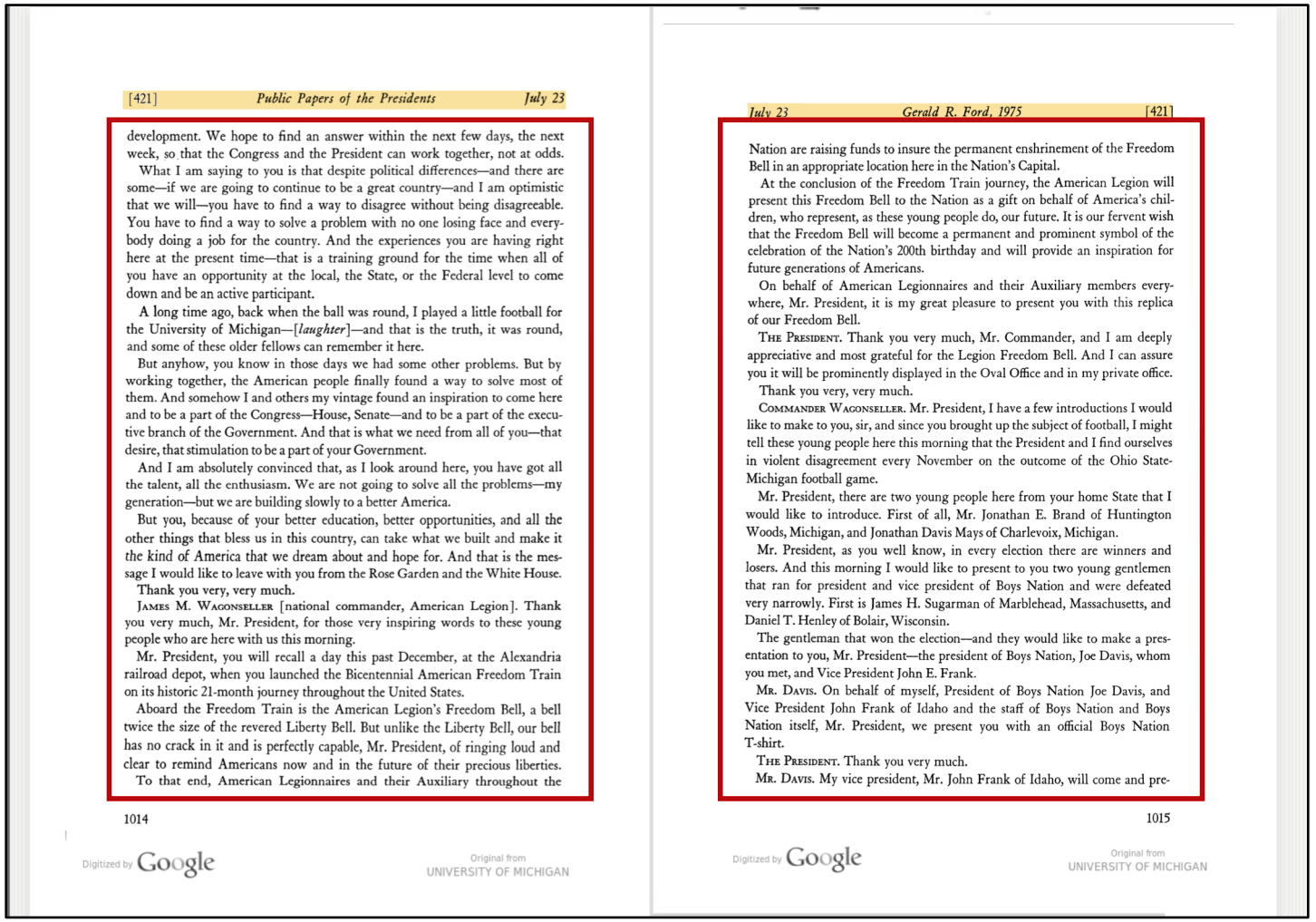
HTRC Extracted Features (EF)

- The features are
  - Selected data and metadata
  - Extracted from raw text
- Position the researcher to begin analysis
  - Some of the preprocessing is already done
- Form of non-consumptive access

Extracted Features model



Header, body, footer?



[421] *Public Papers of the Presidents* *July 23*

development. We hope to find an answer within the next few days, the next week, so that the Congress and the President can work together, not at odds.

What I am saying to you is that despite political differences—and there are some—if we are going to continue to be a great country—and I am optimistic that we will—you have to find a way to disagree without being disagreeable. You have to find a way to solve a problem with no one losing face and everybody doing a job for the country. And the experiences you are having right here at the present time—that is a training ground for the time when all of you have an opportunity at the local, the State, or the Federal level to come down and be an active participant.

A long time ago, back when the ball was round, I played a little football for the University of Michigan—[*laughter*]—and that is the truth, it was round, and some of these older fellows can remember it here.

But anyhow, you know in those days we had some other problems. But by working together, the American people finally found a way to solve most of them. And somehow I and others my vintage found an inspiration to come here and to be a part of the Congress—House, Senate—and to be a part of the executive branch of the Government. And that is what we need from all of you—that desire, that stimulation to be a part of your Government.

And I am absolutely convinced that, as I look around here, you have got all the talent, all the enthusiasm. We are not going to solve all the problems—my generation—but we are building slowly to a better America.

But you, because of your better education, better opportunities, and all the other things that bless us in this country, can take what we built and make it the kind of America that we dream about and hope for. And that is the message I would like to leave with you from the Rose Garden and the White House.

Thank you very, very much.

JAMES M. WAGONSELLER [national commander, American Legion]. Thank you very much, Mr. President, for those very inspiring words to these young people who are here with us this morning.

Mr. President, you will recall a day this past December, at the Alexandria railroad depot, when you launched the Bicentennial American Freedom Train on its historic 21-month journey throughout the United States.

Aboard the Freedom Train is the American Legion's Freedom Bell, a bell twice the size of the revered Liberty Bell. But unlike the Liberty Bell, our bell has no crack in it and is perfectly capable, Mr. President, of ringing loud and clear to remind Americans now and in the future of their precious liberties.

To that end, American Legionnaires and their Auxiliary throughout the

1014

July 23 *Gerald R. Ford, 1975* [421]

Nation are raising funds to insure the permanent enshrinement of the Freedom Bell in an appropriate location here in the Nation's Capital.

At the conclusion of the Freedom Train journey, the American Legion will present this Freedom Bell to the Nation as a gift on behalf of America's children, who represent, as these young people do, our future. It is our fervent wish that the Freedom Bell will become a permanent and prominent symbol of the celebration of the Nation's 200th birthday and will provide an inspiration for future generations of Americans.

On behalf of American Legionnaires and their Auxiliary members everywhere, Mr. President, it is my great pleasure to present you with this replica of our Freedom Bell.

THE PRESIDENT. Thank you very much, Mr. Commander, and I am deeply appreciative and most grateful for the Legion Freedom Bell. And I can assure you it will be prominently displayed in the Oval Office and in my private office.

Thank you very, very much.

COMMANDER WAGONSELLER. Mr. President, I have a few introductions I would like to make to you, sir, and since you brought up the subject of football, I might tell these young people here this morning that the President and I find ourselves in violent disagreement every November on the outcome of the Ohio State-Michigan football game.

Mr. President, there are two young people here from your home State that I would like to introduce. First of all, Mr. Jonathan E. Brand of Huntington Woods, Michigan, and Jonathan Davis Mays of Charlevoix, Michigan.

Mr. President, as you well know, in every election there are winners and losers. And this morning I would like to present to you two young gentlemen that ran for president and vice president of Boys Nation and were defeated very narrowly. First is James H. Sugarman of Marblehead, Massachusetts, and Daniel T. Henley of Bolair, Wisconsin.

The gentleman that won the election—and they would like to make a presentation to you, Mr. President—the president of Boys Nation, Joe Davis, whom you met, and Vice President John E. Frank.

MR. DAVIS. On behalf of myself, President of Boys Nation Joe Davis, and Vice President John Frank of Idaho and the staff of Boys Nation and Boys Nation itself, Mr. President, we present you with an official Boys Nation T-shirt.

THE PRESIDENT. Thank you very much.

MR. DAVIS. My vice president, Mr. John Frank of Idaho, will come and pre-

1015

Volume metadata

- Pulled from bibliographic metadata
- Title
- Author
- Language
- Identifiers


```
{
   "id":"uc1.b3419888",
   "metadata":{
      "schemaVersion":"1.2",
      "dateCreated":"2015-02-12T13:30",
      "title":"Zoonomia = or The laws of organic life / by Erasmus Darwin.",
      "pubDate":"1809",
      "language":"eng",
      "htBibUrl":"http://catalog.hathitrust.org/api/volumes/full/htid/uc1.b3419888.json",
      "handleUrl":"http://hdl.handle.net/2027/uc1.b3419888",
      "oclc":"3679915",
      "imprint":"Thomas and Andrews, 1809."
   },
   "features":{
      "schemaVersion":"2.0",
      "dateCreated":"2015-02-20T23:58",
      "pageCount":616,
      "pages":[
```

Page metadata

- Page sequence
- Computationally-inferred metadata
  - Word, line, and sentence counts
  - Empty line count
  - Language


```
{
    "seq":"00000035",
    "tokenCount":507,
    "lineCount":44,
    "emptyLineCount":0,
    "sentenceCount":14,
    "languages":[
    {
        "en":"1.00"}],
```

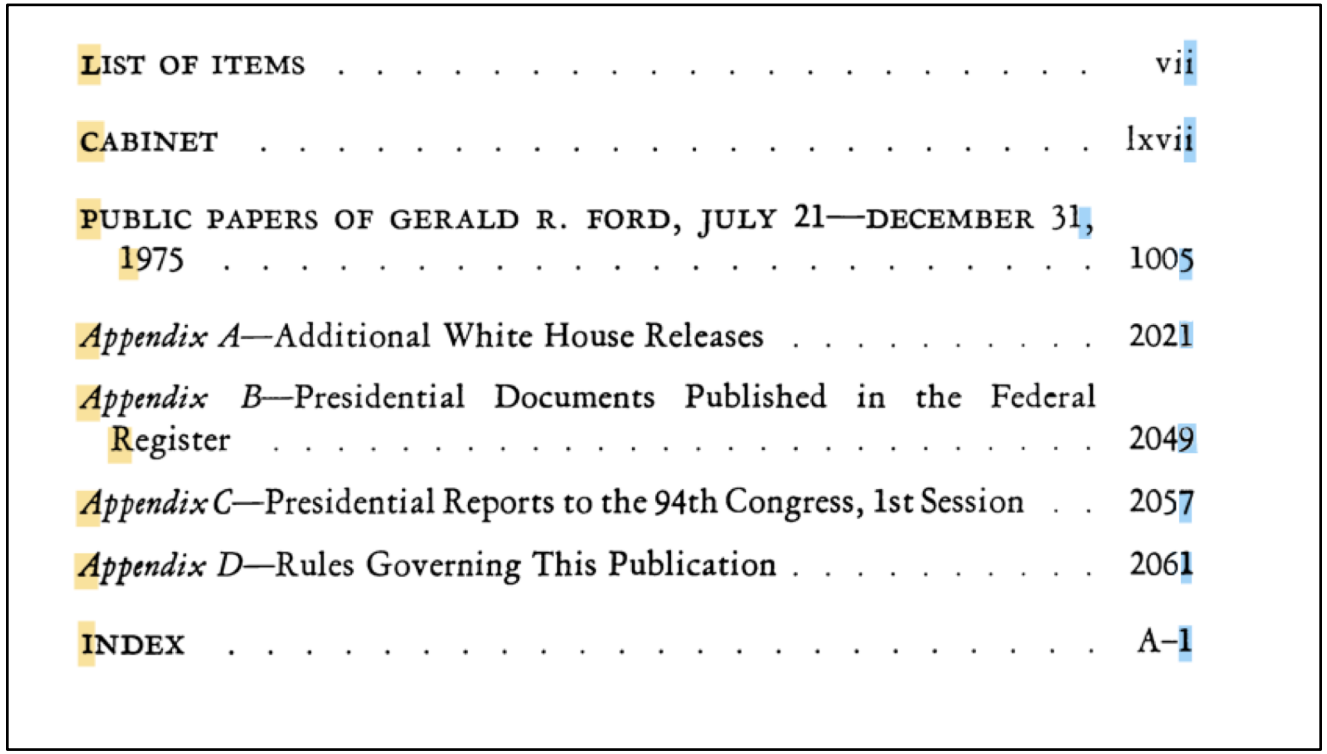
Page section features

Header, body, footer

- Line, empty line, and sentence count
- Counts of beginning- and end-line characters
- Token counts
  - Homonyms counted separately
  - Part-of-speech codes are from the Penn Tree Bank


```
40 "body":{
41     "tokenCount":504,
42     "lineCount":43,”
43     "emptyLineCount":0,”
44     "sentenceCount":12,
45     "tokenPosCount":{
46 "fynthefis":{"NNP":1},
47 "Laws":{"NNP":1},
48 "beautiful":{"JJ":1},
49 "philofopher":{"NN":1},
50                 "uponthe":{"IN":1},
51 "for":{"IN":1},
```

Begin line characters?



Hands-on: examine an EF file

Handout p. 8

Open the sample file in Box: mdp.39015073767769.json

Review the file and see if you can find:

- The OCLC number

- How many lines are in the body of page sequence 00000005

Using HTRC Extracted Features

- Identify parts of a book
  - From descriptive metadata
- Perform any method that works with bags-of-words
  - Topic modeling
  - Dunning’s log-likelihood
- Classify volumes
  - Compare with bibliographic metadata

HTRC Feature Reader Python library

- Python library for working with HTRC Extracted Features
  - Code to facilitate research using the JSON files
- Install using a package manager, like pip
  - Source code lives on Github
- Requires Pandas to run
  - pandas = Python library for working with data

Accessing Extracted Features

- Entire set is 4 TB; download what you need
- Need to know the rsync paths to the EF files you want to download
  - Use Feature Reader id_to_rsync function to get paths, then rysnc
  - Use the HTRC EF Download Helper algorithm to generate a shell script to rsync EF, then run the resulting file from the command line `sh polisci.sh`

Accessing Extracted Features

- Files sync in pairtree format
- File storage format
- Nested directories, broken down by characters in file name

```
abcd       -> ab/cd/
abcdefg    -> ab/cd/ef/g/
12-986xy4 -> 12/-9/86/xy/4/
```

https://wiki.ucop.edu/display/Curation/PairTree

Hands-on: Analyzing Extracted Features

Handout p. 8

Work with HTRC Extracted Features files to understand the data model and basic Python programming for analyzing text data.

Website: https://go.illinois.edu/htrc-workshop

- Launch Binder (if not already launched)
- 03-ef-activity.ipynb

Case studies: text analysis pipelines

Handout p. 8

1. Go back to the *Transformation of Gender* case study
2. Also open this page: https://github.com/dbamman/book-nlp
3. Then can you:
   - Explain the pipeline to your neighbor?
   - What tools or steps does it include?

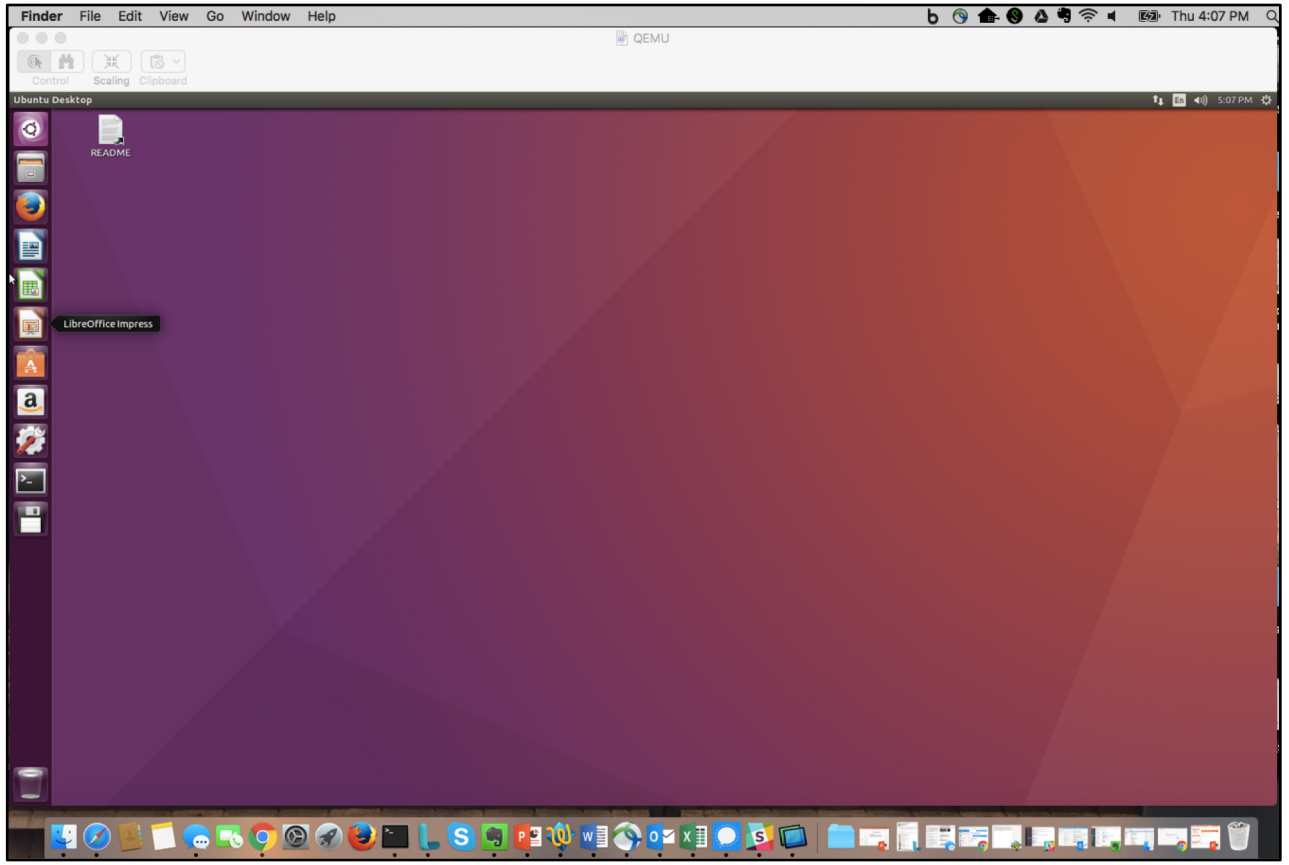
Other HTRC tools and services

HathiTrust data access options

Method	Data	Description	Rights status	Restrictions
HT dataset request	Full text OCR	Download page images and plain text OCR	Public domain	Depends on your university
HT Data API	Full text OCR, page images	Download page images and plain text OCR	Public domain	Non-Google digitized only
HTRC Algorithms	Full text OCR (not viewable)	Analyze a workset using off-the-shelf tools	All	Data can be computed on, but is not exposed
HTRC Extracted Features	Abstracted text and metadata	JSON files for each of 15.7 million volumes in HathiTrust	All	Data is preprocessed
HTRC Data API	Full text OCR	Analyze plain text OCR	All for HT members; else public domain	For use in a Data Capsule only

Secure Data Capsules

- Secure analysis environments
- Linux virtual desktop
- Protocols for data import/export



https://analytics.hathitrust.org/staticcapsules

Custom dataset request

- For researchers who need public domain data only
- Downloadable, OCR and metadata
- Request procedure
- Some limitation based on Google-digitization

https://www.hathitrust.org/datasets

Dataset help

- Assistance crafting lists of volume IDs
- For HathiTrust custom dataset requests
  - feedback@issues.hathitrust.org
- For HTRC worksets
  - htrc-help@hathitrust.org

Advanced Collaborative Support awards

- Competitively awarded “grants”
  - Time and resources awarded
- In the 5th round now
- Read the descriptions and reports:

  https://wiki.htrc.illinois.edu/x/CADiAQ

ACS project example

The Chicago School: Wikification as the First Step in Text Mining in Architectural History

- Explored history of the term “Chicago School” through the corpus
- Used wikifier tool to link named entities to Wikipedia entries
- Found different types of “Chicago Schools” (e.g., the Chicago School of bone breakers)
- Compute-intensive, relied on access to high performance computers
- Found use of the term as applied to architecture dating to 1889

Researcher: Dan Baciu

Documentation

- https://wiki.htrc.illinois.edu/
- Further information
- Technical documentation
- Step-by-step guides

Office hours

- Every 3rd Wednesday from 3-4 p.m. ET
- Ask questions, connect with other researchers
- go.illinois.edu/htrchelp-live

Help email

- htrc-help@hathitrust.org
- For general inquiries, troubleshooting, and research consultations

Questions?

htrc-help@hathitrust.org

https://teach.htrc.illinois.edu

Materials modified from curriculum funded by 

award #RE-00-15-0112-15

Bibliography

- Bode, K. (2019). Why you can't model away bias. Preprint: *Modern Language Quarterly 80.3.* https://katherinebode.files.wordpress.com/2019/08/mlq2019_preprintbode_why.pdf.
- Denny, M. J. and Spirling, A. (2017). Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What to Do about It. https://ssrn.com/abstract=2849145.
- Moretti, F. (2013). *Distant reading*. Verso Books.
- Rockwell, G. (2003). What is Text Analysis, Really? *Literary and Linguistic Computing*, *18*(2), 209–219. https://doi.org/10.1093/llc/18.2.209.