Text Mining with HathiTrust
An Introduction for Librarians
Set up instructions

• 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

HathiTrust Research Center Analytics
Supports large-scale computational analysis of the works in the HathiTrust Digital Library to facilitate non-profit and educational research.

Featured Services
- Extracted Features
- Text Analysis Algorithms
- Data Capsules

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

**Partial:**
- partial data, open output, open access

**Transform:**
- transformed data, open output, open access

**Capsule:**
- full data, controlled output, controlled access
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
HathiTrust text corpus
synched nightly, lives at IU

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
HATHITRUST

Feeds into
HTRC Analytics
Feeds into
Data Capsule service
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?*

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

[Graph showing publication dates as of October 2018]

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

https://www.hathitrust.org/snapshot-of-hathitrust-collection-10-years
### Top 10 Contributors to HathiTrust as of March 2017

<table>
<thead>
<tr>
<th>Contributor</th>
<th>Volumes</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>University of Michigan</td>
<td>4,727,160</td>
<td>31%</td>
</tr>
<tr>
<td>University of California</td>
<td>3,902,003</td>
<td>26%</td>
</tr>
<tr>
<td>Harvard University</td>
<td>841,981</td>
<td>6%</td>
</tr>
<tr>
<td>Cornell University</td>
<td>586,023</td>
<td>4%</td>
</tr>
<tr>
<td>University of Illinois at Urbana-Champaign</td>
<td>562,023</td>
<td>4%</td>
</tr>
<tr>
<td>University of Wisconsin - Madison</td>
<td>561,985</td>
<td>4%</td>
</tr>
<tr>
<td>Indiana University</td>
<td>530,775</td>
<td>4%</td>
</tr>
<tr>
<td>University of Minnesota</td>
<td>520,774</td>
<td>4%</td>
</tr>
<tr>
<td>The University of Texas</td>
<td>460,151</td>
<td>3%</td>
</tr>
<tr>
<td>Pennsylvania State University</td>
<td>390,364</td>
<td>3%</td>
</tr>
<tr>
<td><strong>Remaining 42 Contributors</strong></td>
<td><strong>1,993,400</strong></td>
<td><strong>13%</strong></td>
</tr>
</tbody>
</table>
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
Collection visualization: HathiTrust + Bookworm

*Brings together:*

- Text data (unigrams)
- Bibliographic metadata
- Visualization tool
- Track trends in a repository
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

Use HT+BW to explore lexical trends

Website: https://bookworm.htrc.illinois.edu/develop
Examples
Examples
# HathiTrust data access options

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

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

Workset viewed on HTRC Analytics

Workset manifest
Creating HTRC Worksets

How would you like to create your workset?

Upload File

Create a workset from a file of HathiTrust volume IDs

Import From HathiTrust

Create a workset from an existing, public HathiTrust collection
HathiTrust user-created collections

To find, go to www.hathitrust.org

Then, click ‘Collections’ at top
Read and reflect

• 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

• “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

• 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

• 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.

<table>
<thead>
<tr>
<th>Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>Punctuation</td>
</tr>
<tr>
<td>Numbers</td>
</tr>
<tr>
<td>Lowercasing</td>
</tr>
<tr>
<td>Stemming</td>
</tr>
<tr>
<td>Stopword Removal</td>
</tr>
<tr>
<td>n-gram Inclusion</td>
</tr>
<tr>
<td>Infrequently Used Terms</td>
</tr>
</tbody>
</table>
Case studies: preparing text data

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

Run the HTRC’s Named Entity Recognizer algorithm

Website: [https://analytics.hathitrust.org/](https://analytics.hathitrust.org/)
The Modern Traveller workset

<table>
<thead>
<tr>
<th>Owner</th>
<th>Last Modified Time</th>
<th>Number of Volumes</th>
<th>Tags</th>
<th>JSON-LD</th>
</tr>
</thead>
<tbody>
<tr>
<td>eleanordicksonkoehl</td>
<td>2019-09-21T17:59:15Z</td>
<td>30</td>
<td></td>
<td>View</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Volume ID</th>
<th>Title</th>
<th>Authors</th>
<th>Year</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>mdp.39015074624258</td>
<td>The modern traveller; a description of the various countries of the globe. By Josiah Conder ...</td>
<td>Conder, Josiah 1789-1855</td>
<td>1830</td>
<td>eng</td>
</tr>
<tr>
<td>mdp.39015074624316</td>
<td>The modern traveller; a description of the various countries of the globe. By Josiah Conder ...</td>
<td>Conder, Josiah 1789-1855</td>
<td>1830</td>
<td>eng</td>
</tr>
<tr>
<td>mdp.39015073767918</td>
<td>The modern traveller; a description of the various countries of the globe. By Josiah Conder ...</td>
<td>Conder, Josiah 1789-1855</td>
<td>1830</td>
<td>eng</td>
</tr>
<tr>
<td>mdp.39015074623623</td>
<td>The modern traveller; a description of the various countries of the globe. By Josiah Conder ...</td>
<td>Conder, Josiah 1789-1855</td>
<td>1830</td>
<td>eng</td>
</tr>
</tbody>
</table>
Navigate to the HTRC algorithms

https://analytics.hathitrust.org
Find the workset

<table>
<thead>
<tr>
<th>Name</th>
<th>Author</th>
<th>Description</th>
<th>Volume Count</th>
<th>Last Modified Date</th>
<th>Availability</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>000mixed_grill</td>
<td>researcher603</td>
<td>ducks and squirrels</td>
<td>16</td>
<td>April 23, 2015</td>
<td>public</td>
<td></td>
</tr>
</tbody>
</table>
Filter by name & select the workset

<table>
<thead>
<tr>
<th>Name</th>
<th>Author</th>
<th>Description</th>
<th>Volume Count</th>
<th>Last Modified Date</th>
<th>Availability</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>TheModernTraveller</td>
<td>eleanordicksonkoehl</td>
<td>Complete set of volumes from Josiah Conder's The Modern Traveller.</td>
<td>30</td>
<td>September 21, 2019</td>
<td>public</td>
<td></td>
</tr>
</tbody>
</table>
View workset & choose NER algorithm from menu
Input job details & submit

Job Name (required)
ModernTravellerEntities

Please select a workset for analysis (required)
TheModernTraveller@eleanordicksonkoehl
Select a collection for analysis.
This algorithm has volume size limit of 3000, and only worksets with fewer than 3000 volumes are displayed above.

Please specify the predominant language in your workset (required)
English
Select the language most prevalent in your workset, and your text will be tokenized following rules for that language. This algorithm supports only the languages in the drop-down list.

Submit
See your job in progress

### Jobs

#### Active Jobs

<table>
<thead>
<tr>
<th>Job Name</th>
<th>Algorithm</th>
<th>Date Completed</th>
<th>Expires On</th>
<th>Status</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>ModernTravellerEntities</td>
<td>Named_Entity_Recognizer</td>
<td>2019-09-24</td>
<td>2021-03-24</td>
<td>Staging</td>
<td><img src="https://via.placeholder.com/15" alt="X" /></td>
</tr>
</tbody>
</table>

Showing 1 to 10 of 1 entries
View your results

### Modern Traveller Entities

<table>
<thead>
<tr>
<th>Name</th>
<th>Job ID</th>
<th>Algorithm</th>
<th>Date Completed</th>
<th>Expires On</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modern Traveller Entities</td>
<td>80fd60fe-c9a8-47ba-a896-c39bc313a5b1</td>
<td>Named_Entity_Recognizer</td>
<td>2019-09-24</td>
<td>2021-03-24</td>
<td>Finished</td>
</tr>
</tbody>
</table>

### Input Parameters

<table>
<thead>
<tr>
<th>Name</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>language</td>
<td>en</td>
</tr>
<tr>
<td>input_collection</td>
<td>TheModernTraveller@eleanordicksonkoehl</td>
</tr>
</tbody>
</table>

### Output

- entities.csv
- stdout.txt
- stderr.txt

<table>
<thead>
<tr>
<th>vol_id</th>
<th>page_seq</th>
<th>entity</th>
<th>type</th>
</tr>
</thead>
<tbody>
<tr>
<td>mdp.39015074623607</td>
<td>00000002</td>
<td>E</td>
<td>PERCENT</td>
</tr>
<tr>
<td>mdp.39015074623607</td>
<td>00000002</td>
<td>TT3%</td>
<td>PERCENT</td>
</tr>
<tr>
<td>mdp.39015074623607</td>
<td>000000005</td>
<td>O</td>
<td>ORGANIZATION</td>
</tr>
</tbody>
</table>
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

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 in 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

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

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

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

Extracted Features File (for 1 volume from HathiTrust)

Volume metadata

Page features

Page 001
- Page metadata
- Header data
- Body data
- Footer data

Page 002
- Page metadata
- Header data
- Body data
- Footer data
development. We hope to find an answer within the next few days, the next
work, 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 every-
body doing a job for the country. And the experience you are having right
here at the present time—that is the 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—(laugh)—and that is the truth, it was round,
and some of these older fellows can remember it here.
But anyhow, you know it is three days we had some other problems. But by
working together, the American people finally found a way to solve some of
those. And somehow I and others my vintage found an inspiration to come here
and be part of the Congress—House, Senate—and to be part of the execu-
tive 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 all the
skills, all the abilities, 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 mes-
gle I would like to share with you from the Rose Garden and the White House.
Thank you very, very much.
James M. Watson (national commander, American Legion). Thank
you very much, Mr. President, for these 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.
Abroad the Freedom Train is the American Legion’s Freedom Bell, a bell
twice the size of the original 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
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 chil-
dren who represent, as those 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 every-
where, Mr. President, it is my great pleasure to present you with this replica
of our Freedom Bell.
Test 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.
Constance Wagoner (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 those 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 R. Board of Huntington
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 soundly. First is James H. Sugerman of Marlboro, Massachusetts, and
Daniel T. Hatchet of Beloit, Wisconsin.
The gentlemen that won the election—and they would like to make a pre-
sentation to you, Mr. President—their vice president of Boys Nation, Joe Davis, when
you were, 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.
Test President. Thank you very much.
Mr. Davis. My vice president, Mr. John Frank of Idaho, will come and pre-
Volume metadata

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

```json
{
  "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."
  }
}
```
Page metadata

• Page sequence

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

```json
{  
  "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

```
"body":{
    "tokenCount":504,
    "lineCount":43,"
    "emptyLineCount":0,"
    "sentenceCount":12,
    "tokenPosCount":{
    "synthesis":{"NNP":1},
    "Laws":{"NNP":1},
    "beautiful":{"JJ":1},
    "philosopher":{"NN":1},
    "uponthe":{"IN":1},"for":{"IN":1},
```

Hands-on: examine an EF file

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 rsync
  • 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

https://wiki.ucop.edu/display/Curation/PairTree
Hands-on: Analyzing Extracted Features

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

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Data</th>
<th>Description</th>
<th>Rights status</th>
<th>Restrictions</th>
</tr>
</thead>
<tbody>
<tr>
<td>HT dataset request</td>
<td>Full text OCR</td>
<td>Download page images and plain text OCR</td>
<td>Public domain</td>
<td>Depends on your university</td>
</tr>
<tr>
<td>HT Data API</td>
<td>Full text OCR, page images</td>
<td>Download page images and plain text OCR</td>
<td>Public domain</td>
<td>Non-Google digitized only</td>
</tr>
<tr>
<td>HTRC Algorithms</td>
<td>Full text OCR (not viewable)</td>
<td>Analyze a workset using off-the-shelf tools</td>
<td>All</td>
<td>Data can be computed on, but is not exposed</td>
</tr>
<tr>
<td>HTRC Extracted Features</td>
<td>Abstracted text and metadata</td>
<td>JSON files for each of 15.7 million volumes in HathiTrust</td>
<td>All</td>
<td>Data is preprocessed</td>
</tr>
<tr>
<td>HTRC Data API</td>
<td>Full text OCR</td>
<td>Analyze plain text OCR</td>
<td>All for HT members; else public domain</td>
<td>For use in a Data Capsule only</td>
</tr>
</tbody>
</table>
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](https://wiki.htrc.illinois.edu/x/CADiAQ)
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


