Text Mining with HathiTrust: Empowering Librarians to Support Digital Scholarship Research

Eleanor Dickson Koehl

HathiTrust

Follow this and additional works at: https://digital.sandiego.edu/symposium

https://digital.sandiego.edu/symposium/2019/2019/4

This Workshop is brought to you for free and open access by Digital USD. It has been accepted for inclusion in Digital Initiatives Symposium by an authorized administrator of Digital USD. For more information, please contact digital@sandiego.edu.
Text Mining with HathiTrust: Empowering Librarians to Support Digital Scholarship Research

Presenter 1 Title
Digital Scholarship Librarian

Session Type
Workshop

Abstract
This workshop will introduce attendees to text analysis research and the common methods and tools used in this emerging area of scholarship, with particular attention to the HathiTrust Research Center. The workshop's "train the trainer" curriculum will provide a framework for how librarians can support text data mining, as well as teach transferable skills useful for many other areas of digital scholarly inquiry. Topics include: introduction to gathering, managing, analyzing, and visualizing textual data; hands-on experience with text analysis tools, including the HTRC’s off-the-shelf algorithms and datasets, such as the HTRC Extracted Features; and using the command line to run basic text analysis processes. No experience necessary! Attendees must bring a laptop.

Location
KIPJ Room EF

Creative Commons License
This work is licensed under a Creative Commons Attribution 4.0 License.

This workshop is available at Digital USD: https://digital.sandiego.edu/symposium/2019/2019/4
Text Mining with HathiTrust: Empowering Librarians to Support Digital Scholarship Research
Getting started

- Workshop materials and resources:
  - http://go.illinois.edu/ddrf-curriculum

- HTRC Analytics and the HTDL:
  - https://analytics.hathitrust.org  |  https://www.hathitrust.org

Handout p. 1 for instructions
1. Introduction
In this section we’ll…

- Introduce text analysis and broad text analysis workflows
  → Make sense of digital scholarly research practices
- Introduce HathiTrust and the HathiTrust Research Center
  → Understand the context for one text analysis tool provider
- Introduce our hands-on example and case study
  → Recognize research questions text analysis can answer
What is text analysis?

- Using computers to reveal information in and about text
  (Hearst, 2003)
  - Algorithms discern patterns
  - Text may be “unstructured”
  - More than just search

- What is it used for?
  - Seeking out patterns in scientific literature
  - Identifying spam e-mail
How does it work?

- 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
How does it impact research?

- Shift in perspective, leads to shift in research questions
  - Scale-up to “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
Activity

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?
Example: *Rowling and “Galbraith”: an authorial analysis*

**Question:**

*Did JK Rowling write *The Cuckoo’s Calling* under the pen name Robert Galbraith?*

Would be impossible to prove through human reading alone!

*comparative | patterns*

**Read more:** Rowling and “Galbraith”: an authorial analysis (Juola, 2013)
Example: *Rowling and “Galbraith”: an authorial analysis*

**Approach:**

- Reading led to hunch about authorship
- Computational comparison of diction between this book and others written by Rowling
- Statistical ‘proof’ of authorial fingerprint

**Read more:** Rowling and “Galbraith”: an authorial analysis (Juola, 2013)
Example: *Significant Themes in 19th Century Literature*

**Question:**

*What themes are common in 19th century literature?*

Answering this question requires a very large corpus and an impossible amount of human reading!

*patterns | comparative*

**Read more:** Significant Themes in 19th Century Literature (Jockers and Mimno, 2012)**
Example: *Significant Themes in 19th Century Literature*

**Approach:**
- Run large quantities of text through a statistical algorithm
- Words that co-occur are likely to be about the same thing
- Co-occurring words are represented as topics

*Read more: Significant Themes in 19th Century Literature (Jockers and Mimno, 2012)*
Example: *Significant Themes in 19th Century Literature*

From paper - Figure 3: Word cloud of topic labeled “Female Fashion.”
Example: *The Emergence of Literary Diction*

**Question:**

*What textual characteristics constitute “literary language”?*

This question covers a very large time span!

change over time | patterns

**Read more:** The Emergence of Literary Diction (Underwood and Sellers, 2012)
Example: *The Emergence of Literary Diction*

**Approach:**

- Train a computational model to identify literary genres
- Compare which words are most frequently used over time in non-fiction prose versus “literary” genres
- Demonstrated tendency for poetry, drama, and fiction to use older English words

*Read more:* The Emergence of Literary Diction (Underwood and Sellers, 2012)
Example: The Emergence of Literary Diction

Y axis: Yearly ratio of words that entered English before 1150 / words that entered from 1150-1699

From paper: graph of diction patterns between genres, using frequency counts

X axis: Year
HathiTrust

- Founded in 2008
- Grew out of large-scale digitization initiative at academic research libraries
  - With roots in Google Books project
- Over 120 partner institutions continue to contribute
HathiTrust Digital Library

- Contains over 16 million volumes
  - ~ 50% English
  - From the 15th to 21st century, 20th century concentration
  - ~ 63% in copyright or of undetermined status
- Search and read books in the public domain
HathiTrust Research Center

- Facilitates text analysis of HTDL content
- Research & Development
- Located at Indiana University and the University of Illinois
HTRC for text analysis

Digitized text
- Scanned & OCR-ed

Computational methods
- E.g. Word counts, classification, topic modeling

Analysis

HathiTrust Research Center
- at scale from the digital library
- provided tools and services
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
Workshop outline

- Follow the research process:
  - Gathering textual data
  - Working with textual data
  - Analyzing textual data
  - Visualizing textual data

- Hands-on activities around a central research question & case study example at each step
  - Using both HTRC and non-HTRC tools
Sample Reference Question

Question:

I’m a student in history who would like to incorporate digital methods into my research. I study American politics, and in particular I’d like to examine how concepts such as liberty change over time.

Approach:

- We’ll practice approaches for answer this question throughout the workshop
Case Study

*Inside the Creativity Boom* | Researcher: Samuel Franklin

**Question:**

*How do the use and meaning of creative and creativity change over the 20th century?*

**Approach:**

- We’ll discuss how this researcher approached his question throughout the workshop

Learn more: [https://wiki.htrc.illinois.edu/x/CADiAQ](https://wiki.htrc.illinois.edu/x/CADiAQ)
A word of caution…

Workshop outline suggests research workflow like:

1. Find text
2. Prepare text
3. Apply algorithm
4. Visualize results
A word of caution...

Actual research workflow like:

- Search for text
- Get access to text
- Clean text
- Exploratory visualization
- Prepare text
- Apply algorithm
- Visualize results
Discussion

- What examples have you seen of text analysis?

- What makes a research question conducive to data mining methods?
Questions?
2. Gathering Textual Data
In this section we’ll…

- Explore the concept of a text data and where to find it
  - Provide data reference for researchers
- Build a HathiTrust workset
  - Gain experience in building a textual dataset
- Learn how Sam built a Creativity Corpus of HathiTrust volumes
  - Understand real-world data collection strategies
Where we’ll end up

Create a collection of volumes from the HathiTrust Digital Library and prepare it for analysis in HTRC Analytics as a workset.
“Text analysis projects share in common 3 challenges. First, data of interest must be found. Second, data must be gettable. Third, if it’s not already formed according to wildest dreams, ways must be known of getting data into a state that they are readily usable with desired methods and tools.”

Kludging: Web to TXT (Padilla, 2015)

http://www.thomaspadilla.org/2015/08/03/kludge/
Finding text

- Not always easy
  - copyright restrictions
  - licensing restrictions
  - format limitations
  - hard-to-navigate systems

** issues more pronounced at scale**
Vendor databases

- Be aware of licensing restrictions

- Strategies
  - Addendums to libraries’ contracts
  - Vendor-provided services
  - Asking for special permission case-by-case

- Example: JSTOR Data for Research
Library/archives digital collections

- Wealth of material, but:
  - Often siloed
  - Access not formulated for research at scale

- Things to look for:
  - Plain text
  - Bulk download

- Example: UNC’s DocSouth Data
Social media

- Popular with social science researchers

- To access:
  - Some provide systems to access text
  - Or there are 3rd-party tools on the market

- Example: Twitter API (Application Programming Interface)
Activity

Building a corpus for political history, what are the strengths and weaknesses of each of these broad sources for textual data?

<table>
<thead>
<tr>
<th>Source</th>
<th>Strengths</th>
<th>Weaknesses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vendor database</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Library/archives digital collections</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social media</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Handout p. 3
Evaluating sources of text data

<table>
<thead>
<tr>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Does the researcher already have a data source in mind?</td>
</tr>
<tr>
<td>Is the text they want to use already digitized?</td>
</tr>
<tr>
<td>Are there copyright and licensing concerns?</td>
</tr>
<tr>
<td>How technically experienced is the researcher?</td>
</tr>
<tr>
<td>What is the period, place, person of interest?</td>
</tr>
<tr>
<td>How much flexibility is needed for working with the data?</td>
</tr>
<tr>
<td>Does the researcher have funding?</td>
</tr>
<tr>
<td>What format does the researcher expect the data in?</td>
</tr>
</tbody>
</table>
Building corpora

- Identify texts through full text search
  - Use a key term or phrase

- Identify texts through metadata
  - Search by certain author(s)
  - Search within a date range
  - Search for a specific genre

- Or some combination of the two!
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
Discussion

- What expertise do librarians already have to help with building a corpus for textual analysis?
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

<table>
<thead>
<tr>
<th>Owner</th>
<th>Last Modified Time</th>
<th>Number of Volumes</th>
<th>Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>man11</td>
<td>2017-10-05T18:21:35Z</td>
<td>16</td>
<td></td>
</tr>
</tbody>
</table>

## Volume ID: mdp.49015002221845
- **Authors**
- **Year**: 1978
- **Language**: eng

## Workset viewed on the web

## Workset manifest
Building worksets

- Stored in HTRC
  - Require account with university email address
- Ways to build:
  - Import from HT Collection Builder
  - Compile volume IDs elsewhere
I’m a student in history who would like to incorporate digital methods into my research. I study American politics, and in particular I’d like to examine how concepts such as liberty change over time.

Approach:

- Create a textual dataset of volumes related to political speech in America with the HT Collection Builder, and upload it to HTRC Analytics as a workset for analysis.
Activity

In this activity, you will log in to HTDL and create a collection containing volumes of the public papers of the presidents of the United States, and import it into HTRC Analytics as a workset. Follow the instructions on the handout to build your workset.

Websites:

- HTDL: https://www.hathitrust.org
- HTRC Analytics: https://analytics.hathitrust.org
Go to HTDL interface

https://www.hathitrust.org
Log in
Log in
Search for volumes

- Click on “Advance full-text search”
Search for volumes
Filter results and select volumes

Filter results on the left sidebar

Select all or some of the returned search items for your collection.

An advanced search for volumes that contain all the words/phrases below in the title field: “public papers” and “United States”
Add volumes to collection

Once texts are selected, click “Select Collection” → choose “[CREATE NEW COLLECTION]” → click “Add Selected”
Add collection metadata
View your collection
View your collection
Grab the collection URL
Go to HTRC Analytics

https://analytics.hathitrust.org
Welcome! Returning users signing into the new HTRC Analytics interface for the first time must reset their password using the "Forgot Password" link below.

Sign In to HathiTrust Research Center

Username

Password

Remember me on this computer

SIGN IN

Forgot Password? | Forgot Username? | Create Account
Sign in

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
Go to Worksets page
Choose to create a workset
Choose creation method

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
Input workset information

Create A Workset

Import a collection from HathiTrust using the collection's URL. While HathiTrust grows daily, HTRC syncs data periodically from the HathiTrust Digital Library. Some volumes you would like to include in your workset may not be available. Any volumes in your workset not available through HTRC will be skipped by the algorithm.

Find collection URL
When viewing your collection on HathiTrust, simply copy the URL from your browser, or copy the "Link to this collection" found on the left sidebar, and paste the URL below.

Import
Hit "Fetch Collection" and your collection will be transformed into an HTRC workset. You may need to edit the default name in order to meet HTRC requirements.

HathiTrust Collection URL
https://babel.hathitrust.org/cgi/mb?x=collections:184985365

Reset
Name
poll_science_DDRF
Disallowed characters: - ! # $ % ^ & * ( ) [ ] { } \ / |

Description
Political science collection for DDRF workshop

☐ Private Workset
If checked, your workset will be accessible to only you.

Create Workset

Add collection URL here
View created workset

Worksets

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Volume Count</th>
<th>Last Modified Date</th>
<th>Availability</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1970s_political_speeches_DDRF</td>
<td>1970s political speeches collection for DDRF workshop</td>
<td>16</td>
<td>October 5, 2017</td>
<td>public</td>
<td></td>
</tr>
<tr>
<td>Darwin_Test_2</td>
<td></td>
<td>3</td>
<td>October 6, 2016</td>
<td>private</td>
<td></td>
</tr>
<tr>
<td>1930s_political_speeches_DDRF</td>
<td>1930s political speeches collection for DDRF workshop</td>
<td>5</td>
<td>October 5, 2017</td>
<td>public</td>
<td></td>
</tr>
<tr>
<td>poli_science_DDRF</td>
<td>Political science collection for DDRF workshop</td>
<td>16</td>
<td>October 5, 2017</td>
<td>public</td>
<td></td>
</tr>
<tr>
<td>Charles_Darwin_Test</td>
<td></td>
<td>32</td>
<td>August 24, 2016</td>
<td>private</td>
<td></td>
</tr>
</tbody>
</table>

Showing 1 to 10 of 5 entries
Workset review

- How did it go?
- What kind of search criteria did you use?
- Did you find any challenges?
Bulk retrieval

- Most researchers will need more than 1 or 10 texts
  - Hundreds, thousands, or millions of texts

- Getting lots of data could take lots of time!
  - Point-and-click is inefficient
  - Automate when possible
Automating retrieval

Transferring files

- FTP or SFTP: (Secure) File Transfer Protocol
  - moves files from one place to another on the internet

- rsync
  - Efficient: sends only the differences
  - Run from command line
  - Used by HathiTrust, can be used to download Extracted Features data
Web scraping (grabbing text on the web)

- Avoids tedious copying-and-pasting
- Some ways to scrape text from the web:
  - Run commands such as wget or curl in the command line
  - Write and run a script (a file of programming statements)
  - Use software such as webscraper.io or Kimono
Web scraping for the wise

- Web scraping puts a large workload on targeted server
  - This can upset the data holder

- Some data providers are more than willing to share
  - Ask for access
  - Check for an API

- Otherwise, time your requests to add a delay between server hits
  - It’s polite
  - Also signifies you are not a malicious attacker
Automating retrieval

APIs (Application Programming Interfaces)

- Digital pathways to or from content
  - Sometimes need a “key” for access
- Can be used to gain programmatic access
  - Usually need to write code to retrieve content
  - Sometimes have graphical user interface (GUI)
- Examples: A number of digital content providers have APIs
  - Twitter API: display tweets on a non-Twitter website
  - Chronicling America API: https://chroniclingamerica.loc.gov/about/api/
## Bulk HathiTrust data access

### HT and HTRC datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Kind of data</th>
<th>Description</th>
<th>Volumes available</th>
</tr>
</thead>
<tbody>
<tr>
<td>HT Custom data request</td>
<td>Full text</td>
<td>Download page images and plain text OCR</td>
<td>Public domain</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>
</tr>
<tr>
<td>HTRC Data API</td>
<td>Full text</td>
<td>Plain text OCR</td>
<td>All for HT members; public domain for others</td>
</tr>
</tbody>
</table>
Case study: *Inside the Creativity Boom*

**Building a creativity corpus**

- Searched across full text of HTDL for creativ*
- Made initial list of over million volumes
- De-duplicated
  - Kept different editions of same work; discard multiple copies of same edition
- Ended up with refined list (workset) of volumes
Case Study: *Inside the Creativity Boom*

After creating final list of volumes:

- Used rsync to retrieve HTRC Extracted Features for the volumes
- Remember rsync is a command line utility that transfers files between computers
Case Study: *Inside the Creativity Boom*

- What exactly is the HTRC Extracted Features dataset?

```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."
    },
    "features": {
        "schemaVersion": "2.0",
        "dateCreated": "2015-02-20T23:58",
        "pageCount": 616,
        "pages": []
    }
}
```

**Metadata:** bibliographic; inferred

**Data:** words and word counts
Features in the HTRC

- HTRC Extracted Features dataset
- Downloadable
- Structured data consisting of features
- 5 billion pages, in 13.6 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
Per-volume features

- Pulled from bibliographic metadata
- Title
- Author
- Language
- Identifiers
Per-page features

- 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"
        }
    ]
}
```
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 experience you are having right here at the present time—that is a running ground for the time when all of you have an opportunity at the local, the State, the Federal level to come from 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 those older fellows can remember it.

But anyway, 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 at some stage found an inspiration to come here and 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—any 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 much.

James M. Wexlerman (senatorial consultant, American Legion). Thank you very much, Mr. President, for those very inspiring words to those 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 second 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 American now and in the future of their precious liberties.

To that end, American Legislators and their Auxiliary throughout the United States are raising funds to insure that the permanent enshrinement of the Freedom Bell is 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 in 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 Legislators and their Auxiliary members everywhere, Mr. President, it is my great pleasure to present you with this replica of our Freedom Bell.

Mr. 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 proudly displayed in the Oval Office and in my private office. Thank you very much.

Congresswoman Vandenbroek. Mr. President, I have a few introductions I would like to make to you, Mr. President, and on your behalf I would like to introduce our guest, the President 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. Straw of Huntington Woods, Michigan, and the President of the Ohio State-Michigan football game.

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 F. Sagerman of Marlborough, Massachusetts, and David T. Hanley of Beloit, Wisconsin.

The gentlemen 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 for Davis, and Vice President John Frank of Idaho and the President of Boys Nation and Boys Nation itself, Mr. President, we present you with an official Boys Nation T-shirt.

Mr. President: Thank you very much.

Mr. Davis. My vice president, Mr. John Frank of Idaho, will come and present...
### List of Items

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of Items</td>
<td>vii</td>
</tr>
<tr>
<td>Cabinet</td>
<td>lxvi</td>
</tr>
<tr>
<td>Public Papers of Gerald R. Ford, July 21—December 31, 1975</td>
<td>1005</td>
</tr>
<tr>
<td>Appendix A—Additional White House Releases</td>
<td>2021</td>
</tr>
<tr>
<td>Appendix B—Presidential Documents Published in the Federal Register</td>
<td>2049</td>
</tr>
<tr>
<td>Appendix C—Presidential Reports to the 94th Congress, 1st Session</td>
<td>2057</td>
</tr>
<tr>
<td>Appendix D—Rules Governing This Publication</td>
<td>2061</td>
</tr>
<tr>
<td>Index</td>
<td>A–1</td>
</tr>
</tbody>
</table>

Public papers of the presidents of the United States (Gerald R. Ford, book 2)
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

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

- “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.”

- Principle 2 for collections as data: “Collections as data development aims to encourage computational use of digitized and born digital collections.”

--- 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?
Questions?
3. Working with Textual Data
In this module we’ll…

- Think about what happens when text is data
  - Understand best practice in the field
- Consider common steps to cleaning and preparing text data
  - Make recommendations to researchers
- Learn how Sam prepared his *Creativity Corpus* for analysis
  - See how one scholar data prepared data
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
  - HathiTrust OCR is dirty (uncorrected)

- Analyzed by corpus/corpora
  - Text corpus: a digital collection OR an individual’s research text dataset
  - Text corpora: “bodies” of text

- Text decomposition/recomposition (Rockwell, 2003)
  - Cleaning data involves discarding data
  - Prepared text may be illegible to the human reader
Preparing data

A researcher may:

- 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

Tokenization

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

[four], [score], [and], [seven], [years], [ago], [our], [fathers], [brought], [forth], [on], [this], [continent], [a], [new], [nation], [conceived], [in], [liberty], [and], [dedicated], [to], [the], [proposition], [that], [all], [men], [are], [created], [equal]
Preparing data

- Preparation affects results
  - Amount of text and size of chunks
  - Which stop words removed; which characters are included
  - Whether to lowercase and normalize words

- Preparation for analysis takes time, effort
  - This is where scripting becomes useful!
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 Study: *Inside the Creativity Boom*

After downloading the Extracted Features data for the relevant volumes, used scripting to:

- Narrow corpus to individual pages that contained creativ*
  - Discarded all other pages

- Discard certain tokens such as pronouns and conjunctions
  - To keep only to most "meaningful" terms
Read and Reflect…

- Passage from “Against Cleaning” by Katie Rawson and Trevor Muñoz
- They suggest a strategy for dealing with humanities data:
  - Shared authority control across data sets
  - Indexes for nuance
  - Tidy, not clean data
“When humanities scholars recoil at data-driven research, they are often responding to the reductiveness inherent in this form of scholarship. This reductiveness can feel intellectually impoverishing to scholars who have spent their careers working through particular kinds of historical and cultural complexity… From within this worldview, data cleaning is then maligned because it is understood as a step that inscribes a normative order by wiping away what is different. The term “cleaning” implies that a data set is ‘messy.’ “Messy” suggests an underlying order. It supposes things already have a rightful place, but they’re not in it—like socks on the bedroom floor rather than in the wardrobe or the laundry hamper.”

- “Against Cleaning” (Rawson and Muñoz, 2016)
Discussion

- What does this excerpt suggest about the nuances of data cleaning?
- What does “clean” imply?
- How might you talk to researchers on your campus who would be uncomfortable with the idea of clean v. messy data?
Questions?
4. Analyzing Textual Data

*Using Off-the-Shelf Tools*
In this module we’ll…

- Weigh the benefits and drawbacks of pre-built tools for text analysis
  - Evaluate researcher questions and requests, and match tool to request

- Learn how a web-based topic modeling algorithm works
  - Gain experience with off-the-shelf solutions text mining

- Run the HTRC Topic Modeling algorithm and analyze the results
  - Build confidence with the outcomes of data-intensive research

- See how Sam explored HTRC Algorithms for his research
  - Understand how a researcher evaluated an off-the-shelf tool
Where we’ll end up

Bubble visualization of topics created with HTRC algorithm
Pre-built tools

- **Benefits**
  - Easy to use, good for teaching

- **Drawbacks**
  - Less control, limited capabilities

- **Examples:**
  - Voyant, Lexos
  - HTRC algorithms: e.g. Topic Modeling algorithm
Choosing a pre-built tool

- Quick analysis and visualizations:
  - Voyant
  - Lexos

- Concordances:
  - AntConc
  - Voyant

- Machine learning
  - WEKA Workbench aids machine learning
Do-it-yourself tools

- Alternative to pre-built, off-the-shelf tools
- Involve programming

Benefits:
  - Run on your own, allow for more parameterization and control

Drawback:
  - Require technical knowledge
HTRC algorithms

- Plug-and-play text analysis
- Built into the HTRC interface
  - Mostly “as-is”
  - Limited parameterization
  - Analyze HTRC worksets
- Good when you want to use HT text specifically
Choosing an HTRC algorithm

- **Task-oriented algorithms:**
  - Produce list of named entities
  - Visualize most frequently used words
  - Generate script for downloading Extracted Features files

- **Analytic algorithms:**
  - Generate topic models
Key terms in text analysis

Bag-of-words

Concept where grammar and word order of the original text are disregarded and frequency is maintained.

created the four in new are ago Liberty fathers that forth continent a nation seven and conceived equal score dedicated on to years this all our men brought and proposition
Key terms in context

**Topic Modeling**

- **Chunk** text into documents
- Documents = bags of words
- **Stop words** are removed
- Each word in each document is compared
- Words that tend to occur together in documents are likely to be about the same thing
- Topics are predictions of words co-occurrence
Tips for topic modeling

- Treat topic modeling as step in analysis
- Input affects output
  - Number of texts analyzed, number of topics generated
  - Be familiar with your input data
  - Know that stop words can shape results
- Examine results to see if they make sense
- Understand the tool
HTRC topic modeling description

InPhO Topic Model Explorer

Volumes in a workset may be inaccessible to HTRC Analytics algorithms. You can validate your workset before submitting your job using the Workset Validation page to check which volumes are accessible.

Description

The InPho Topic Explorer trains multiple LDA topic models and allows you to export files containing the word-topic and topic-document distributions, along with an interactive visualization. For full detailed description, please review the documentation.

How it works:

• Downloads each HathiTrust volume from the Data API.
• Tokenizes each volume using the topicexplorer init command.
• Apply stoplists based on the frequency of terms in the corpus, removing the most frequent words accounting for 50% of the collection and the least frequent words accounting for 10% of the collection.
• Create a new topic model for each number of topics specified. For example, "20 40 60 80" would train separate models with 20 topics, 40 topics, 60 topics and 80 topics.
• Display a visualization of how topics across models cluster together. This enables a user to see the granularity of the different models and how terms may be grouped together into "larger" topics.

More documentation of the Topic Explorer is available at https://inpho.github.io/topic-explorer/.

https://analytics.hathitrust.org/algorithms/InPhO_Topic_Model_Explorer
Sample Reference Question

I’m a student in history who would like to incorporate digital methods into my research. I study American politics, and in particular I’d like to examine how concepts such as liberty change over time.

Approach: Run topic modeling algorithm to get a feel for the topics present in your workset.
Activity

In this activity you will run the topic modeling algorithm in HTRC Analytics to explore the most prevalent topics in our president public papers workset.

What You Need:

Website: https://analytics.hathitrust.org

Workset: poli_science_DDRF
About the political science workset

- Government-published series: *Public papers of the presidents of the United States*
  - “Public Messages, Speeches, and Statements of the President”
- 16 volumes from U.S. presidents during the 1970s:
  - Jimmy Carter
  - Gerald Ford
  - Richard Nixon
- We’ll use the same workset (‘poli_science_DDRF@eleanordickson’) so that we can all examine the same results!
Using the HTRC Algorithms

Handout p. 7

https://analytics.hathitrust.org
Analysis in the HTRC

**Algorithms**

**Extracted Features Download Helper (v3.0.2)**

Generate a script that allows you to download extracted features data for your workset of choice. The script is a file containing a list of the rsync commands to access the volumes of the workset. After you download the script from HTRC Analytics, it can be run locally (from your computer), which will then download the extracted features data to your computer via rsync. For more information on the extracted features data see the documentation.

[Execute]

**InPhO Topic Model Explorer (v1.0)**

The InPho Topic Explorer trains multiple LDA topic models and allows you to export files containing the word-topic and topic-document distributions, along with an interactive visualization. For full detailed description, please review the documentation.

[Execute]
Prepare to run an algorithm

Form:

Author(s)
Jaimie Murdock

Job Name (required)

Collection (required)
Select workset
The workset you would like to analyze.
This collection has a size limit of 3000, hence the above workset selector shows the worksets which has less than 3000 volumes.

Number of iterations (required)
200
A lower number of iterations will process faster. A higher number will yield higher quality results.

Number of topics (required)
20 40 60 80
The number of topics (k) to train the model on. Accepts multiple values, separated by spaces, e.g., "20 40 60 80". You will be able to toggle between the models in your results.

Submit
Prepare to run an algorithm

Job Name (required)

Collection (required)
- Select workset

The workset you would like to analyze. This collection has a size limit of 3000, hence the above workset selector shows the worksets which has less than 3000 volumes.

Number of iterations (required)
- 200

A lower number of iterations will process faster. A higher number will yield higher quality results.
Choose workset(s) for analysis

Check box to include public worksets first
Prepare to run an algorithm

<table>
<thead>
<tr>
<th>Job Name (required)</th>
<th>TestJobName</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collection (required)</td>
<td>poll_science_DDRF@eleanordickson</td>
</tr>
<tr>
<td></td>
<td>Include public worksets</td>
</tr>
</tbody>
</table>
| The workset you would like to analyze.
This collection has a size limit of 3000, hence the above workset selector shows the worksets which has less than 3000 volumes. |
| Number of iterations (required) | 200 |
| A lower number of iterations will process faster. A higher number will yield higher quality results. |
| Number of topics (required) | 20 40 60 80 |
| The number of topics (k) to train the model on. Accepts multiple values, separated by spaces, e.g., "20 40 60 80". You will be able to toggle between the models in your results. |
| Submit |
Set the number of topics

**Number of topics (required)**

- 20 60

The number of topics (k) to train the model on. Accepts multiple values, separated by spaces, e.g., “20 40 60 80”. You will be able to toggle between the models in your results.
Run the analysis

```table
<table>
<thead>
<tr>
<th>Job Name</th>
<th>Algorithm</th>
<th>Last Updated</th>
<th>Status</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>TestJobName</td>
<td>InPhO_Topic_Model_Explorer</td>
<td>2018-08-06 16:51:59</td>
<td>Staging</td>
<td></td>
</tr>
</tbody>
</table>
```

Showing 1 to 10 of 1 entries
View results

InPhO Topic Explorer

- Hover over bubble for top terms in topic
- Check box to minimize overlap among nodes
Topics visualized

InPhO Topic Explorer
Results files

```
Output

/topics.html  cluster.csv  workers.tez  topics.json  stdout.txt  stderr.txt

Click here to open topics.json in a new tab

{
   "0": { 
      "color": "#1b9e77",
      "words": { 
         "peace": 0.008993002586066723,
         "tion": 0.00785261131966114,
         "con": 0.0076391128823161125,
         "made": 0.00725898239761591,
         "first": 0.0071079717017171021,
         "because": 0.0070194480940699958,
         "economic": 0.006956961005926132,
         "two": 0.0068840594030916699,
         "ing": 0.00661848857998848,
         "programs": 0.006519550457596779
      }
   },
   "1": { 
      "color": "#d76003",
      "words": { 
         "going": 0.018236661329865456,
         "tax": 0.014271358959376812,
         "model": 0.01355786650937035
      }
   }
}
```
Topics listed

Examples from 20-topics cluster:

**Topic 1**
- nation, because, problems,
- under, america, security,
- nations, programs, con, much

**Topic 2**
- may, such, peace, war,
- between, america, last, must,
- after, soviet

Examples from 60-topics cluster:

**Topic 3**
- like, department, percent, said, things, office,
- get, assistance, programs, every

**Topic 4**
- oil, programs, presidents, nations,
- cooperation, york, billion, council, kind, visit

**Topic 5**
- problems, much, system, economy, proposed,
- must, each, end, case, effective
Analyzing results

- What would you name these topics?
- Are you skeptical of any of the results?
- Did you learn anything new from the topics produced?
Key approaches to text analysis

Broad Area: Natural Language Processing (NLP)

Using computers to understand the meaning, relationships, and semantics within human-language text

• Specific Methods:
  • **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 approaches to text analysis

Broad Area: Machine Learning

Training computers to recognize patterns.

• **Specific Methods**
  • **Topic modeling** – What thematic topics are present in the text?
  • **Naïve Bayes classification** – Which of the categories that I have named does the text belong to?
Activity: Identify the method

<table>
<thead>
<tr>
<th>Broad area</th>
<th>Specific method</th>
</tr>
</thead>
<tbody>
<tr>
<td>'Rowling and &quot;Galbraith&quot;: an authorial analysis</td>
<td></td>
</tr>
<tr>
<td>Significant Themes in 19th Century Literature</td>
<td></td>
</tr>
<tr>
<td>The Emergence of Literary Diction</td>
<td></td>
</tr>
</tbody>
</table>

Note:
Broad areas/specific methods are those defined in the previous two slides

Link to project summaries: [http://go.illinois.edu/ddrf-research-examples](http://go.illinois.edu/ddrf-research-examples)
Case Study: *Inside the Creativity Boom*

- Before making his Creativity Corpus, Sam experimented with an older version of the HTRC topic modeling algorithm.
- His practice HTRC workset included public domain texts from 1950 to present:
  - Creativ* in the title
Case Study: Inside the Creativity Boom

Are these good topics?
Tips for topic modeling

- Treat topic modeling as step in analysis
- Be familiar with input text
- Examine results to see if they make sense
- Know that stop words can shape results
- Understand the tool
Case Study: *Inside the Creativity Boom*

- Sam then used HTRC Extracted Features to get the data needed to analyze contemporary material.
- The fits and starts of his project are a great real-world example!
Case Study: *Inside the Creativity Boom*

After reducing Creativity Corpus to pages containing forms of creativ*:

- Performed topic modeling on those pages
- Ended up with topics that reflect what themes are prevalent around concept of “creativity” in the 20th century
- Graphed the topics over time to see how their usage changed
Case Study: Inside the Creativity Boom

- Topics that decreased in usage over time
  - god, christ, jesus, creation, word
  - species, animals, natural, plants, soil
  - nature, mind, creative, world, human
  - invention, power, creative, own, ideas
Case Study: *Inside the Creativity Boom*

- Topics that increased in usage over time
  - advertising, media, marketing, sales, television
  - economic, development, capital, economy, production
  - poetry, language, poet, poets, poems
  - social, creative, study, development, behavior
Discussion

- To what kinds of researchers on your campus would you recommend pre-built text analysis tools?

- What additional skills do you feel you would need to develop in order to support advanced researchers?
5. Visualizing Textual Data
In this module we’ll…

- Introduce common visualization strategies for text data
  - Communicate with researchers about their options
- Use a web-based visualization tool, HathiTrust+Bookworm
  - Gain experience creating and reading data visualizations
- See how Sam used HathiTrust+Bookworm for his project
  - Learn how HT+BW was utilized in research
Where we’ll end up

Create visualization of word usage trends across the HathiTrust corpus.
Data visualization

- Data visualization is the process of converting data sources into a visual representation.
- Visualizations present particular ways of interpreting data.
- Data visualization is an entire field of study; we’re barely scratching the surface.
Why visualize text data?

- Understand broader themes of a dataset
- Explore patterns in the data
- Cluster texts for overview or classification
- Compare data to other data (e.g., correlating with social networks)

Adapted from Jason Chuang’s Text Visualization course at Stanford University
Place in research process

- In the earlier exploration stage of a project:
  - Explore full range of data
  - Discover characteristics and themes in data

- In the later explanation stage of a project:
  - Communicate findings to others in a clearer and more efficient way
Common text data visualizations

Word cloud

- Relatively unsophisticated, but effective
- Size of word relates to prominence or salience

Topic models from HTRC Algorithms
Common text data visualizations

Trees or hierarchies

- Word trees

Occurrences of “I have a dream” in Martin Luther King’s historical speech. (Wattenberg and Viégas, 2008)
Common text data visualizations

Networks

- Node-link diagrams
- Good for representing topic models
- Visualize connections between named entities

Topic model of English books, 1850-1899
(Underwood, 2012)
Common text data visualizations

Temporal- or spatial-based visualizations

- Temporal visualizations

Percent representation of female characters in English literature (Underwood and Bamman, 2016)
Common text data visualizations

Temporal or spatial visualizations

- Maps

Percent of newspaper pages containing the term “hoosier” (Palmer, Polley, & Pollock, n.d.)
http://centerfordigschol.github.io/chroniclinghoosier/map1.html
Common text data visualizations

Other “multi-dimensional” visualizations

- Bubble charts
- Heat maps

Common text data visualizations

Other “multi-dimensional” visualizations

- Heat maps

Heatmap of MARC cataloging at the Library of Congress by book year and cataloging year (Schmidt, 2017)
http://sappingattention.blogspot.com/2017/05/a-brief-visual-history-of-marc.html
**Activity**

Match type of use to the type of visualization:

<table>
<thead>
<tr>
<th>Visualization</th>
<th>What would it be good for?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word cloud</td>
<td></td>
</tr>
<tr>
<td>Trees or hierarchies</td>
<td></td>
</tr>
<tr>
<td>Networks</td>
<td></td>
</tr>
<tr>
<td>Timeline</td>
<td></td>
</tr>
<tr>
<td>Map</td>
<td></td>
</tr>
<tr>
<td>Bubble chart</td>
<td></td>
</tr>
<tr>
<td>Heatmap</td>
<td></td>
</tr>
</tbody>
</table>

**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?**
Common visualization tools

- **Word clouds**
  - Voyant
  - Wordle

- **Word use trends**
  - Google Books Ngram Viewer
  - HathiTrust+Bookworm

- **Tabular data visualization**
  - Tableau

- **Mapping**
  - ArcGIS Online with StoryMaps
  - Tableau

- **Network graphs**
  - Gephi
  - NodeXL
  - DH Press
Common visualization libraries

- **Python**
  - matplotlib, pyplot
  - ggplot library

- **R**
  - ggplot2

- **D3.js**
  - Javascript library for visualizations
Review: 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.

four score, score and, and seven, seven years, years ago, ago our, our fathers, fathers brought, brought forth, forth on, on this, this continent, continent a, a new, new nation, nation conceived, conceived in, in liberty, liberty and…
N-gram visualization: HathiTrust + Bookworm

Brings together:

- Text data (unigrams)
- Bibliographic metadata
- Visualization tool
- Track trends in a repository
Bookworm 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”
Example HT+Bookworm view

Track social change: lady vs. woman over time
Reading an HT+BW graph

- Let’s look at how verbs change over time
  - Eg. Burned vs. burnt

Do you see any trends?
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
Sample Reference Question

I’m a student in history who would like to incorporate digital methods into my research. I study American politics, and in particular I’d like to examine how concepts such as liberty change over time.

Approach:

Explore word usage trends of political concepts within the HathiTrust using HT+BW
Activity

- In this activity, you will use HT+BW to explore lexical trends

Website: https://bookworm.htrc.illinois.edu/develop
Examples
Examples
Bookworm review

- What trends did you discover?
Case Study: *Inside the Creativity Boom*

- Sam used HT+Bookworm to visualize the use of “creative” in the HTDL over time.
Case Study: *Inside the Creativity Boom*

- Sam also used an experimental HT+BW interface to create different kinds of visualizations…
Case Study: Inside the Creativity Boom

“Creative” by language and year
Case Study: *Inside the Creativity Boom*

- “Creativity” by library classification and year
Discussion

- Where does visual literacy fit into data literacy overall?
- What would it mean to be visually literate, particularly with regard to text analysis?
Questions?
Need more help?

Contact: htrc-help@hathitrust.org

Materials developed as part of project funded by award #RE-00-15-0112-15

https://teach.htrc.illinois.edu
1. References

  http://people.ischool.berkeley.edu/~hearst/text-mining.html

  http://digitalcommons.unl.edu/englishfacpubs/105/.


2. References


3. References


4. References

5. References


5. References


