Digging Deeper into Government Information: An Introduction to Text Mining

2018 Federal Depository Library Conference

Valerie D. Glenn and Eleanor Dickson Koehl

October 22, 2018

Presenter Information

Valerie D. Glenn, Head, Map and Government Information Library, University of Georgia (previously Federal Documents Analyst, HathiTrust)

Eleanor Dickson Koehl, HTRC Digital Humanities Specialist

What is HathiTrust?

- Brief introduction to HathiTrust
- HathiTrust U.S. Federal Documents Program
  - Overall goals
  - Collection priorities
  - More information: Heather Christenson's presentation regarding the UC FedDocArc project: Tuesday, 3-4, Wilson/Harrison room
- HathiTrust Research Center (HTRC)

What is text analysis?

- Using computers to reveal information **in** and **about** text (Hearst, 2003)
  - Algorithms discern patterns in unstructured text
  - More than just search
- Everyday example
  - E-mail spam filter

How does it work?

- Break textual data into smaller pieces
- Abstract (reduce) text so that a computer can crunch it
- Count!
  - Words, phrases, parts of speech, etc.
- Develop hypotheses from relative counts

How does it impact research?

- Shift perspective → shift research questions
- Opens up:
  - Questions not provable by human reading alone
  - Larger corpora for analysis
  - Studies that cover longer time spans
- One step in the research process
  - Can be combined with close reading

Text Analysis Research Questions

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

Example: “Textual Predictors of Bill Survival in Congressional Committees”

- Research question:

  What makes a bill more or less likely to make it out of a Congressional committee?

Yano, et al., “Textual Predictors of Bill Survival in Congressional Committees,” *2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 793–802, Montreal, Canada, June 3-8, 2012, https://dl.acm.org/citation.cfm?id=2382157

Example: "Textual Predictors of Bill Survival in Congressional Committees"

- Modeled likelihood of a bill's success
  - From 1993-2011
- Made use of metadata features
  - E.g. The sponsor, their political affiliation, and their home state; which party has majority; what month it was introduced
- Also used textual features
  - E.g. Unigrams (single words) to show which words are correlated with a bill's success or demise in committee

https://dl.acm.org/citation.cfm?id=2382157

Example: “Textual Predictors of Bill Survival in Congressional Committees”

- Their results
  - Words associated with success: resources, interior
  - Words associated with failure: energy, security
- Their argument
  - “Minor legislation” more likely to survive
  - Some bills are proposed to make a political statement
  - Smaller bills get rolled into larger ones

https://dl.acm.org/citation.cfm?id=2382157

Building Corpora

- Identify texts through:
  - Full text search
  - Metadata (author, date, genre)
- Usually involves deduplication
  - What to keep/discard is project dependent
- Examples of deduplication criteria:
  - OCR quality
  - Earliest edition
  - Editions without forewords or afterwords

Finding Text

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

** issues more pronounced at scale**

Sources of Text Data

Vendor Databases
Library and archives digital collections
Social media
Digitized personal corpus
Government sources

Considerations for choosing:

Comprehensiveness

- "Official" version
- Timeliness
- Authoritative vs. Public voice
- Likelihood for clean data
- Need for unique sources
- Cost

HTRC Services

Offering	Description	Data availability	Account required
HTRC Algorithms	A set of tools for assembling collections of digitized text from the HathiTrust corpus and performing text analysis on them.	Including items in copyright for ALL USERS.	HTRC Analytics
HTRC Worksets	Worksets are sub-collections of HathiTrust volumes that can be analyzed with HTRC algorithms or used to access HTRC Extracted Features.	Including items in copyright for ALL USERS.	HTRC Analytics
Extracted Features Dataset	A dataset allowing non-consumptive analysis on specific features extracted from the full text of the HathiTrust corpus.	Including items in copyright for ALL USERS.	None
HathiTrust+Bookworm	A tool for visualizing and analyzing word usage trends in the HathiTrust corpus.	Including items in copyright for ALL USERS.	None
HTRC Data Capsule	Asecure computing environment for researcher-driven text analysis on the HathiTrust corpus.	All users may access public domain items. Access to items in copyright is available ONLY to member-affiliated researchers.	HTRC Analytics; plus additional restrictions

HTRC Analytics

www.analytics.hathitrust.org

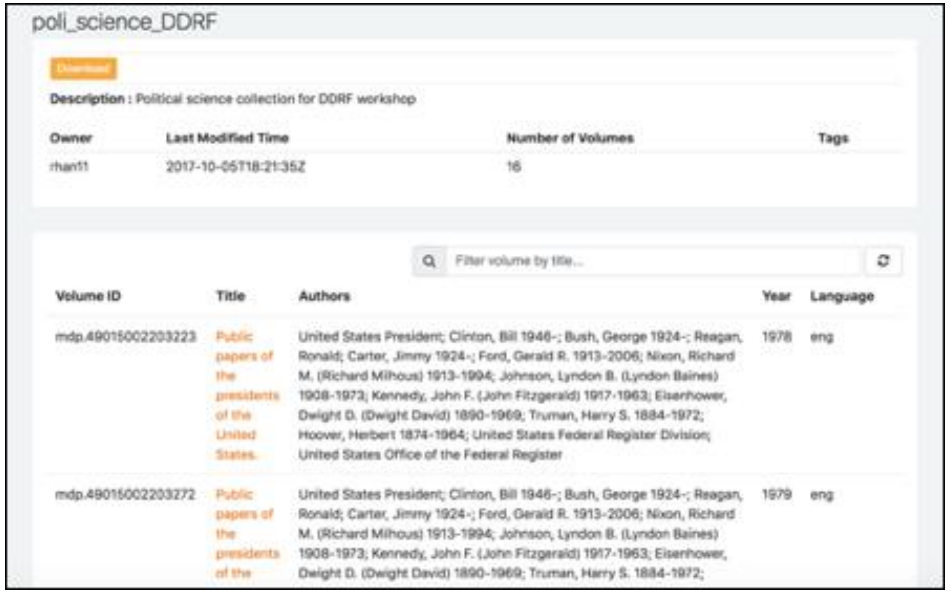
- Accounts for those from institutions of non-profit research or higher education
- Don’t need to be an HT member

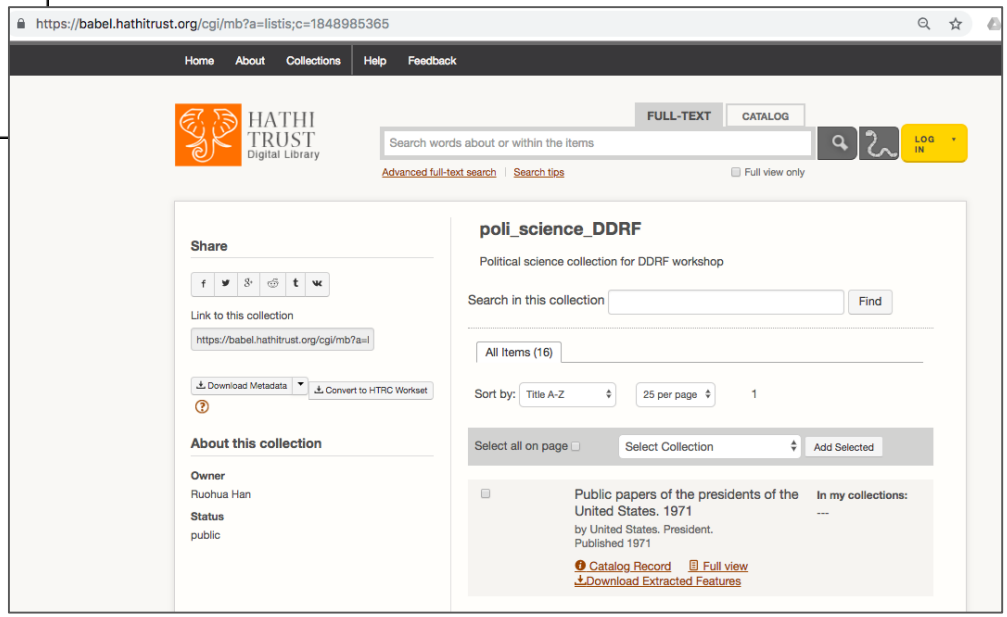
HTRC Worksets

- 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

HT volume IDs

HathiTrust collection





HTRC workset

HTRC U.S. Federal Documents Worksets

- HathiTrust-created collections of text from the HathiTrust Digital Library
  - U.S. Environmental Protection Agency publications
  - U.S. Bureau of Indian Affairs publications
  - U.S. Congressional Serial Set
  - Foreign Relations of the United States

Building Worksets

- Ways to build:
  - Search and create in HT Collection Builder (https://babel.hathitrust.org/cgi/mb)
  - Compile volume IDs making use of HT metadata services (https://www.hathitrust.org/data)
- Import to HTRC Analytics

HTRC Algorithms

- For analyzing HT data only
- Run against a workset
- Kind of analysis:
  - Token count and word cloud
  - Visualize themes (topic modeling)
  - List people, places, organizations, dates, etc. (named entity recognition)

Output

entities.csv | stdout.txt | stderr.txt

Click here to download entities.csv

vol_id	page_seq	entity	type
mdp.39015037380378	00000007	December 1, 1966	DATE
mdp.39015037380378	00000007	1967	DATE
mdp.39015037380378	00000007	Regd	MISC
mdp.39015037380378	00000007	AFRICA	ORGANIZATION
mdp.39015037380378	00000008	Algeria	LOCATION
mdp.39015037380378	00000008	MPLA	ORGANIZATION
mdp.39015037380378	00000008	Sudanese	MISC
mdp.39015037380378	00000008	1966	DATE
mdp.39015037380378	00000008	Independence four years ago	DATE
mdp.39015037380378	00000008	Ghana	LOCATION
mdp.39015037380378	00000008	Central Committee	ORGANIZATION
mdp.39015037380378	00000008	Amin Shaker	PERSON
mdp.39015037380378	00000008	November 29, 1966	DATE
mdp.39015037380378	00000008	Egyptian Gazette	MISC
mdp.39015037380378	00000008	1966	DATE
mdp.39015037380378	00000008	November	DATE

Named entity extractor results

HTRC Extracted Features

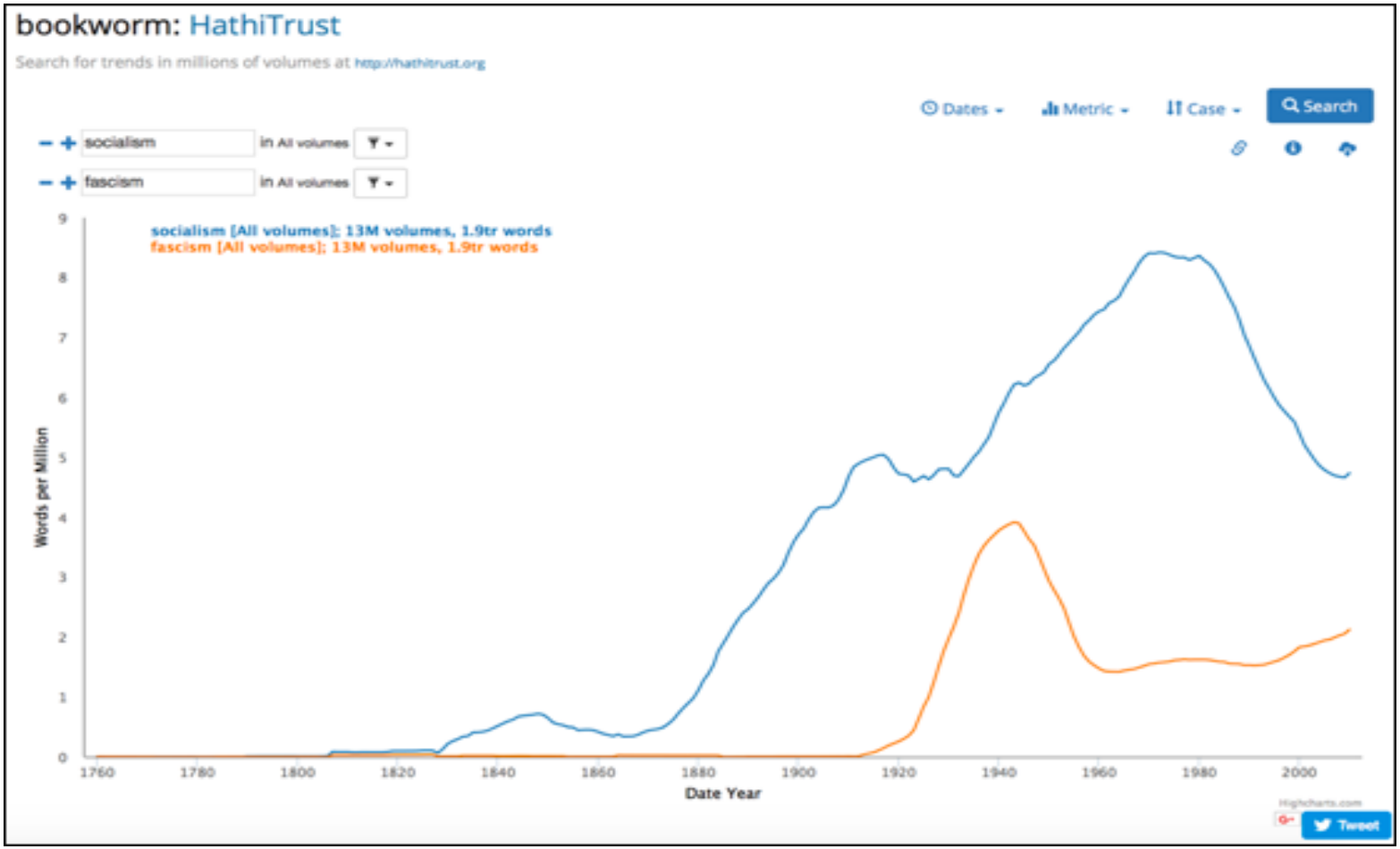
- Selected data and metadata
  - Extracted from raw text
- Includes words and word counts per page


```
{
   "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":[
```

HTRC Data Capsules Environment

- Secure computing desktop environments
- Researcher-driven analysis
- In-copyright data limited to approved researchers from HathiTrust member institutions

HathiTrust + Bookworm



Some non-HTRC Tools

- **Voyant:** easy to use, good for instruction
- **Lexos**: text cleaning, visualizations
- **AntConc**: good for concordances
- **WEKA Workbench:** aids machine learning
- **HTRC algorithms:** good for using HT text specifically

Voyant dashboard

Getting Started

- Create an account for HTRC Analytics
  - HT-member institution affiliation NOT required
- Identify HathiTrust volumes to analyze
  - Search & create collection
  - Make use of metadata services
  - Ask for help!

Getting Started

- Download Extracted Features for the volumes and analyze locally

or

- Import the collection or volume IDs as a workset and run an algorithm

or

- Import data into a Data Capsule and analyze in-Capsule

Thank you! Questions?

Or contact htrc-help@hathitrust.org for more assistance!