Humanistic Inquiry with Large Corpora of Digitized Text and Metadata: Toward New Epistemologies?

Sayan Battacharyya
Jeremy York, Assistant Director, HathiTrust
January 9, 2015

Unless otherwise noted, these slides and their contents are licensed under a Creative Commons Attribution Unported License.
Outline

• Part 1
  – Brief history of digital texts
  – Overview of HathiTrust / Origins of HTRC

• Part 2
  – HathiTrust Research Center Demo
  – HathiTrust Research Center Initiatives

• Part 3
  – Discussion
A Brief History of Digital Texts
Digital text resources

• Project Gutenberg (U of Illinois) – 1971
• Thesaurus Linguae Graecae (UC Irvine) – 1972
• Oxford Text Archive (U of Oxford) – 1976
• ARTFL Project (U of Chicago) – 1982
• Perseus Digital Library (Tufts) – 1985
• Text Encoding Initiative – 1987
• Women Writers Project (Brown U) – 1988
Digital Imaging

- Yale Open Book Project (1991)
- Cornell Demonstration Project (1993)
HathiTrust
HathiTrust Members

Allegheny College
Arizona State University
Baylor University
Boston College
Boston University
Brandeis University
Brown University
California Digital Library
Carnegie Mellon University
Case Western Reserve University
Colby College
Columbia University
Cornell University
Dartmouth College
Duke University
Emory University
Florida State University System
Georgetown University
Getty Research Institute
Harvard University
Indiana University
Iowa State University
Johns Hopkins University
Kansas State University
Lafayette College
Universidad Complutense de Madrid
University of Alabama
University of Alberta
University of Arizona
University of British Columbia
University of Calgary
Syracuse University
Temple University
Texas A&M University
Texas Tech University
Tufts University
University of Notre Dame
University of Oklahoma
Library of Congress
Massachusetts Institute of Technology
McGill University
Michigan State University
Montana State University
Mount Holyoke College
New York Public Library
New York University
North Carolina Central University
North Carolina State University
Northeastern University
Northwestern University
Ohio State University
Oklahoma State University
Pennsylvania State University
Princeton University
Purdue University
Rutgers University
Stanford University
University of California, Berkeley
University of California, Davis
University of California, Irvine
University of California, Los Angeles
University of California, Merced
University of California, Riverside
University of California, San Diego
University of California, San Francisco
University of California, Santa Barbara
University of California, Santa Cruz
University of Chicago
University of Connecticut
University of Delaware
University of Houston
University of Illinois, Chicago
University of Illinois, Urbana Champaign
University of Iowa
University of Kansas
University of Maine
University of Maryland
University of Massachusetts, Amherst
University of Miami
University of Michigan
University of Minnesota
University of Missouri
University of Nebraska-Lincoln
University of New Mexico
University of North Carolina, Chapel Hill
University of Pennsylvania
University of Pittsburgh
University of Queensland
University of Tennessee, Knoxville
University of Texas System
University of Utah
University of Vermont
University of Virginia
University of Washington
University of Wisconsin-Madison
Utah State University
Vanderbilt University
Virginia Tech
Wake Forest University
Washington University, St. Louis
Yale University
The Name

• The meaning behind the name
  – Hathi (hah-tee)—Hindi for elephant
  – Big, strong
  – Never forgets, wise
  – Secure
  – Trustworthy
Growth of the Collection

Year | Collection
-----|-------------
2008 | 2,477,871
2009 | 5,221,092
2010 | 7,836,698
2011 | 9,966,572
2012 | 10,599,355
2013 | 10,878,121
2014 | 13,022,118
Total Numbers

- 13 million total volumes
- 6.6 million book titles
- 340,000 serial titles
- 4.8 million volumes in the public domain (~37%)
- 576,000+ US government publications
The top 10 languages make up ~87% of all content

- English, 49%
- German, 9%
- French, 7%
- Spanish, 5%
- Chinese, 4%
- Russian, 4%
- Italian, 3%
- Japanese, 3%
- Arabic, 2%
- Latin, 1%
- Remaining Languages, 13%
The next 40 languages make up ~12% of total.
Dates

- 1850-1899: 10%
- 1800-1849: 3%
- 2000-2009: 10%
- 1990-1999: 14%
- 1980-1989: 14%
- 1970-1979: 13%
- 1960-1969: 11%
- 1950-1959: 6%
- 1940-1949: 4%
- 1930-1939: 4%
- 1920-1929: 4%
- 1910-1919: 4%
- 1900-1909: 4%
- 1900-1909: 10%
- 0-1500: 0.04%
- 1500-1599: 0.07%
- 1600-1699: 0.01%
- 1700-1799: 0.01%
- 1800-1849: 3%
- 1700-1799: 0.01%
Content Distribution

- In Copyright or Undetermined: 63%
- "Public Domain": 37%
- Open Access: 0.1%
- Creative Commons: 0.1%
- Government Documents: 4%
- Public Domain: 19%
- Public Domain (US): 14%
Core Functionality

• Preservation functions – validation, error-checking
• Discovery
• Reading/Access
• Collections
• APIs
• Data Mining/Analysis
THE WORK OF THE FISCAL YEAR.

The following statement shows the larger items of production for the fiscal years ended June 30, 1910 and 1911:

<table>
<thead>
<tr>
<th>Class of work</th>
<th>1910</th>
<th>1911</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ems of type set</td>
<td>1,801,010,900</td>
<td>1,881,721,000</td>
</tr>
<tr>
<td>Hours of time-work in composing rooms</td>
<td>1,274,913</td>
<td>292,035</td>
</tr>
<tr>
<td>Foundry square-inch production</td>
<td>12,685,357</td>
<td>12,419,416</td>
</tr>
<tr>
<td>Chargeable impressions of presswork, not including postal-card or money-order presses</td>
<td>836,181,344</td>
<td>890,592,006</td>
</tr>
<tr>
<td>Number of forms sent to press</td>
<td>172,089</td>
<td>171,410</td>
</tr>
<tr>
<td>Number of money-order books registered and mailed</td>
<td>396,556</td>
<td>448,139</td>
</tr>
<tr>
<td>Signatures gathered on machine</td>
<td>88,157,456</td>
<td>91,583,833</td>
</tr>
<tr>
<td>Signatures folded on machines</td>
<td>63,812,025</td>
<td>76,496,216</td>
</tr>
<tr>
<td>Number of copies wire-stitched</td>
<td>21,631,448</td>
<td>23,490,410</td>
</tr>
<tr>
<td>Number of sheets passing through ruling machines</td>
<td>22,848,486</td>
<td>25,206,673</td>
</tr>
<tr>
<td>Number of signatures sewed on machines</td>
<td>72,711,576</td>
<td>80,075,988</td>
</tr>
<tr>
<td>Number of tablets made</td>
<td>1,417,521</td>
<td>1,900,421</td>
</tr>
<tr>
<td>Number of cards and sheets punched and drilled</td>
<td>17,128,306</td>
<td>19,265,490</td>
</tr>
<tr>
<td>Number of cases made on machines</td>
<td>1,930,601</td>
<td>2,014,514</td>
</tr>
<tr>
<td>Number of postal cards produced</td>
<td>1,280,895,840</td>
<td></td>
</tr>
</tbody>
</table>

1 Represent 10 months’ actual production and 2 months, approximate or average production.

The expenditure for labor, exclusive of the salaries in the offices of the Public Printer and the superintendent of documents, was $60,790.46 less in 1911 than in the preceding fiscal year.

- The saving for the fiscal year was effected as the result of greater efficiency on the part of the employees and improvement in method and equipment. For example, there was an increase in the chargeable impressions of presswork of 54,410,752 and a decrease of 629 in the number of forms sent to press.
American Guide Series

Owner
soupy

Description
Books and pamphlets about the U.S. published under the auspices of the Federal Writers' Project (FWP)

Status
public

Build and share a collection
APIs

• Bibliographic API
  – Volume and rights information
  – MARC records
  – http://www.hathitrust.org/bib_api

• OAI
  – http://www.hathitrust.org/data

• “Hathifiles”
  – http://www.hathitrust.org/hathifiles

• Data API
  – Volume and rights information
  – Page images
  – OCR
  – http://www.hathitrust.org/data_api
Records loaded into DPLA, local library catalogs, and commercial databases
Data Mining/Analysis

• Dataset Distribution
  – [http://www.hathitrust.org/datasets](http://www.hathitrust.org/datasets)

• Research Center
Examples of uses

- Oxford English Dictionary research
  @bgzimmer Ben Zimmer 7/4/11
  @armavirumque Problem is "cut the mustard" (OED 1891) predates "muster." Earliest I've seen for "muster" is 1912.http://bit.ly/kOy3aD

- Thesis research
- Islamic Manuscripts
- Local/Family History
Projects (1)

• Burton, Vernon. “The South as ‘Other,’ the Southerner as ‘Stranger.’”
  – Explore how attitudes expressed in print about slavery, southerners, and non-southerners have changed over both time and space.

• Ted Underwood, Associate Professor of English at the University of Illinois, Urbana-Champaign.
  – Using public domain texts received from HathiTrust to explore changing relationships in literary genres from 1700-1899.

• Andrew Piper, Associate professor of German literature at McGill University.
  – Analyzing linguistic patterns in German texts from 1700-1900
Projects (2)

• Amanda Watson, librarian at New York University.
  – Studying How poetry anthologies in selected texts reflect the rise and fall of poets’ reputations over the course of the 19th century.

• Glenn Worthey, Digital Humanities Librarian at Stanford University Libraries.
  – Performing spatio-temporal investigation into the history of Brazilian Portuguese, to be accomplished by text-mining methods (n-gram analysis, etc.).

• Matthew Wilkens, Assistant professor of English, University of Notre Dame.
  – American Council of Learned Societies (ACLS) fellowship for project “Literary Geography at Scale.”
How to find out more

- About: [http://www.hathitrust.org/about](http://www.hathitrust.org/about)
- Resources: [http://www.hathitrust.org/resources](http://www.hathitrust.org/resources)
- Twitter: [http://twitter.com/hathitrust](http://twitter.com/hathitrust)
- Facebook: [http://www.facebook.com/hathitrust](http://www.facebook.com/hathitrust)
- Monthly newsletter:
  - [http://www.hathitrust.org/updates](http://www.hathitrust.org/updates)
  - RSS [http://www.hathitrust.org/updates_rss](http://www.hathitrust.org/updates_rss)
- Contact us: feedback@issues.hathitrust.org
- Blogs: [http://www.hathitrust.org/blogs](http://www.hathitrust.org/blogs)
  - Large-scale Search
  - Perspectives from HathiTrust
Introduction to the HathiTrust Research Center (HTRC)

Sayan Bhattacharyya

With grateful acknowledgements to: Harriett Green, Erica Parker, Loretta Auvil, Boris Capitanu, Ted Underwood, Peter Organisciak, and other members of the Illinois and Indiana HTRC teams
Workshop Outline

• **Overview** of the HathiTrust and the HathiTrust Research Center

• **How to use the HTRC Portal**
  - *Create/manage your own custom set of HathiTrust materials*
    • How to use HTRC Portal Workset Builder
  - *Textual analytics on a workset:*
    • How to use HTRC Portal Algorithms, HTRC Bookworm, Data Capsule

• **Opportunities** to connect you and your research with the HathiTrust Research Center
Content Distribution

- **In-copyright or undetermined**: 70%
- **"Public Domain"**: 30%
- **Public Domain (worldwide)**: 15%
- **Public Domain (US)**: 10%
- **Open Access**: .1%
- **Creative Commons**: .01%
- **U.S. Federal Government Documents (worldwide)**: 4%
HathiTrust Collection Builder
Welcome to the HathiTrust Research Center!

The HathiTrust Research Center (HTRC) provides research access to the public domain text of the HathiTrust Digital Library. The HTRC is a collaborative research center launched jointly by Indiana University and the University of Illinois, along with the HathiTrust Digital Library, to help meet the technical challenges of dealing with massive amounts of digital text that researchers face by developing cutting-edge software tools and cyberinfrastructure to enable advanced computational access to the growing digital record of human knowledge.

The HTRC provides an infrastructure to search, collect, analyze, and visualize the full text of nearly 3 million public domain works and is intended for nonprofit and educational researchers.

What Can You Do With HTRC Portal?
- Create Workset
- Upload Workset
- Browse Workset
- Execute Algorithms

Sign in to Begin
www.hathitrust.org/htrc
Log in to the HTRC Portal, [https://htrc2.pti.indiana.edu](https://htrc2.pti.indiana.edu)
Create a login id (i.e. username)
How to create a workset

Welcome to the HathiTrust Research Center!

What Can You Do With HTRC Portal?
- Create Workset
- Sign in to Begin
- Browse Workset
- Execute Algorithms
Log In Again to Workset Builder
Workset Builder
(currently works only on non-copyrighted material not digitized by Google)
Why worksets?

• Two reasons:
  
  – Organizational:
    
    • The “virtual study carrel” idea:
      
      – Gather together material of interest to you in one place
      
      – Accomplished by “slicing and dicing” using search and metadata criteria
  
  – Algorithmic:
    
    • Delimitation of the “scope” of the analysis
      
      – You don’t want to run your analysis on the whole of the HathiTrust collection — you want to run it only on material that is interesting/relevant to you
Search (on metadata and on full text) as means for building a workset
Select desired items
Put them in a workset

1. Finances and costs of the present European war. Prepared by the War college division, General staff corps, as a supplement to the Statement of a proper military policy for the United States. Army war college: Washington, November, 1915.

   Title: Finances and costs of the present European war. Prepared by the War college division, General staff corps, as a supplement to the Statement of a proper military policy for the United States. Army war college: Washington, November, 1915.


   Language: English

   Published: 1916

2. Finances and costs of the present European war. Prepared by the War college division, General staff corps, as a supplement to the Statement of a proper military policy for the United States. Army war college: Washington, November, 1915.

   Title: Finances and costs of the present European war. Prepared by the War college division, General staff corps, as a supplement to the Statement of a proper military policy for the United States. Army war college: Washington, November, 1915.
Analysis in the HTRC Portal

Available Algorithms

- Marc_Downloader
- Meandre_Classification_NaiveBayes
- Meandre_Dunning_LogLikelihood_to_Tagcloud
- Meandre_OpenNLP_Date_Entities_To_Smile
- Meandre_OpenNLP_Entities_List
- Meandre_Spellcheck_Report_Per_Volume
- Meandre_Tagcloud
- Meandre_Tagcloud_with_Cleaning
- Meandre_Topic_Modeling
- Simple_Deployable_Word_Count

Algorithm Parameters

Please select an algorithm in the list to display information.
Choose Algorithm

Note: Enter a name of your choosing in the blank field for “Job Name.” This is the same name that will show up later as “Job Title” when looking at the results.
Choose Collection(s) for Analysis

Available Algorithms

- Marc_Downloader
- Meandre_Classification_NaiveBayes
- Meandre_Dunning_LogLikelihood_to_Tagcloud
- Meandre_OpenNLP_Date_Entities_To_Simile
- Meandre_OpenNLP_Entities_List
- Meandre_Spellcheck_Report_Per_VOLUME
- Meandre_Tagcloud
- Meandre_Tagcloud_with_Cleaning
- Meandre_Topic_Modeling
- Simple_Deployable_Word_Count

THATCamptest@harrigreen
1darwin-test@sheilahoover
2darwin-english@sheilahoover
2vesalus@sheilahoover
Agrippa@rklritz
Anarchism@rswarn
AncientGreek@miao
Austen_Dickens_Labels@lauvil
Author_Twain@pjohnone
BestCoreComplexity@karaj
BigLaw@pleale
Bleakhouse@sheilahoover
Cicero_Orationes_Letters@ajl05r
Coffee_Books@rswarn
Cornell_HIS_2293_1@ebl36
Dickens_as_Authors@skowalczyk
Dickens_yo@bmarks
Diderot-test@sheilahoover
Digital_Preservation@skowalczyk
DocSouthMatch@mfal3
THATCamptest@harrigreen

Please provide a comma separated list of entity types to be extracted. Acceptable values are: date, location, money, organization, percentage, person, time. (default: person)

<table>
<thead>
<tr>
<th>Type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>person</td>
<td>person</td>
</tr>
</tbody>
</table>

Submit
Run the Analysis...

<table>
<thead>
<tr>
<th>Active Jobs</th>
<th></th>
<th></th>
<th>Cancel?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Job Title</td>
<td>Last Updated</td>
<td>Status</td>
<td></td>
</tr>
<tr>
<td>Harriettgtest</td>
<td>2014-04-22 13:05:58</td>
<td>Staging</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Completed Jobs</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>There are no completed jobs..</td>
<td></td>
</tr>
</tbody>
</table>
### Results!

#### Active Jobs

<table>
<thead>
<tr>
<th>Job Title</th>
<th>Last Updated</th>
<th>Status</th>
<th>Cancel?</th>
</tr>
</thead>
<tbody>
<tr>
<td>harrietgt2</td>
<td>2014-04-22 14:01:02</td>
<td>Staging</td>
<td></td>
</tr>
</tbody>
</table>

#### Completed Jobs

<table>
<thead>
<tr>
<th>Job Title</th>
<th>Last Updated</th>
<th>Status</th>
<th>Delete/Save?</th>
<th>Saved?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harrietgt</td>
<td>2014-04-22 13:11:22</td>
<td>Finished</td>
<td></td>
<td>unsaved</td>
</tr>
</tbody>
</table>
View Results

<table>
<thead>
<tr>
<th>Job Title</th>
<th>Harrietttest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithm Name</td>
<td>Meandre_Tagcloud</td>
</tr>
<tr>
<td>Last Updated</td>
<td>2014-04-22 13:11:22</td>
</tr>
<tr>
<td>Results:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>stderr.txt</td>
</tr>
<tr>
<td></td>
<td>stdout.txt</td>
</tr>
<tr>
<td></td>
<td>tagcloud2tokencounts.html</td>
</tr>
<tr>
<td></td>
<td>tagcloud2tokencounts.csv.txt</td>
</tr>
</tbody>
</table>

Job Parameters:

<table>
<thead>
<tr>
<th>Name</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>input_collection</td>
<td>Monster@claireystew</td>
</tr>
</tbody>
</table>

Job Id:  
c815b209-aed2-49f-2eb-2952be44579d

Status: Finished
Topic modeling vs. Dunning log-likelihood

• Topic modeling is useful when you want to get a sense of the contents of your workset.

• Dunning log-likelihood algorithm is useful when you want to do a focused comparison/contrast between two worksets.
  – [If interested in the gory details of how the Dunning log-likelihood algorithm works, see: this blog post by the researcher Ben Schmidt of Northeastern University.  
    (We won’t cover the details today.)]
Example: Comparing (contrasting) two novels by Charles Dickens:

*Little Dorrit* and *Bleak House*

---

*Little Dorrit* ("LittleDorrit1") as the “analysis” workset and *Bleak House* ("BleakHouse1") as the reference workset.

( Words that are **more** represented in *Little Dorrit* than in *Bleak House* are in the file dunning_tagcloud_over.html )
An example: Comparing (contrasting) two novels by Charles Dickens [contd.]

Switch the “analysis” and “reference” worksets

_Bleak House (“BleakHouse1”) as the “analysis” workset and Little Dorrit (“LittleDorrit1”) as the reference workset. (This time, words that are more represented in *Bleak House* than in *Little Dorrit* are in file dunning_tagcloud_over.html, because we switched the “analysis” and “reference” from before.)_
Making sense of the results

Based on what we know of the plots of these two novels, do the generated results make sense?

Plot summaries (abbreviated, from the Dickens Fellowship website):

**Bleak House**: A prolonged law case concerning the distribution of an estate, which brings misery and ruin to the suitors but great profit to the lawyers, is the foundation for this story. Bleak House is the home of John Jarndyce, principal member of the family involved in the law case.

**Little Dorrit**: Here Dickens plays on the theme of imprisonment, drawing on his own experience as a boy of visiting his father in a debtors' prison. William Dorrit is locked up for years in that prison, attended daily by his daughter, Little Dorrit. Her unappreciated self-sacrifice comes to the attention of Arthur Clennam, recently returned from China, who helps bring about her father's release but is himself incarcerated for a time.
Topic modeling *Bleak House*

- For reference, here is a partial snapshot of the results of topic modeling *Bleak House* (number of tokens/topic = 20):
Using the Portal for Research/Teaching

• Two common use cases:
  – Topic modeling
    • If you have a set of books and want to see what common themes run through them, run topic modeling on them
  – Dunning log-likelihood
    • If you have two sets of books and want to compare and contrast them

• Both these algorithms are available on the portal for use with worksets.
Are these algorithms positivist?

• No.
  – Depending on what parameters are selected when running the algorithms, you will get very different results.
  – There is no assumption that there is knowledge “out there” that is being “discovered”
    • Knowledge is *constructed* out of text by the algorithms, which are a form of subjectivity.
• Currently you cannot change/tune parameters on the portal, but in the future, you may be able to.
The Future: HTRC Bookworm

http://sandbox.htrc.illinois.edu/bookworm/

Use for trend analysis: like Google N-gram Viewer, but using metadata to allow analysis on focused worksets.
Looking into the future

- Tools for non-consumptive text analysis on copyrighted texts:
  - HTRC Data Capsule: [https://wiki.htrc.illinois.edu/display/COM/HTRC+Data+Capsule](https://wiki.htrc.illinois.edu/display/COM/HTRC+Data+Capsule)
  - Extracted Features: [https://sandbox.htrc.illinois.edu/HTRC-UI-Portal2/FeatureAction](https://sandbox.htrc.illinois.edu/HTRC-UI-Portal2/FeatureAction)

- HathiTrust + Bookworm Project: [https://htrcbookworm.wordpress.com/](https://htrcbookworm.wordpress.com/)

- Workset Creation for Scholarly Analysis (WCSA) study: [http://worksets.htrc.illinois.edu/worksets/](http://worksets.htrc.illinois.edu/worksets/)

- User guides developed at [http://uiuc.libguides.com/htrcguide](http://uiuc.libguides.com/htrcguide)
Non-consumptive reading via Feature Extraction

What can you do with (1)1 million books...
...if they are under copyright?

You can read book fragments at the HathiTrust Research Center via Feature Extraction...
The Hathi Trust’s (1)1 million books

- abstracted, bird’s-eye view of many texts in aggregated form
- offers possibility for making sense of mankind’s cultural legacy
  - in the form of digitized text
    - available through the world’s great research libraries

But...

- constrained by legal issues related to intellectual property:
  - not all material is in the public domain
  - direct consumption of non-public-domain material is prohibited
Solution: Textual processing \textit{without} downloading of text

• How can this be accomplished?
  – “Non-consumptive” reading
    • Instead of bringing the data to the algorithm:
      – “Data capsule” approach
        » bring the algorithm to the text
      and/or
      – “Features” approach
        » bring certain relevant features of the text to the algorithm
Different modes of distant reading

- Usual mode of text analytics
  - Algorithm
  - Data (Text)

- "non-consumptive" text analytics (e.g. "data capsule")
  - Algorithm
  - Data (Text)
Different modes of distant reading (contd.)

Textual processing using “features”:

Text data → “features” → Analytic algorithm
How to talk about books you haven’t read!
Motivation:

- Almost of all post-1923 publications under copyright
- No bulk downloads of non-copyrighted material permitted

Need for allowing textual analytics without necessitating downloading of full text
Features: What they are

- New HTRC service in alpha release
  - (pilot: 250,000 volumes)

- A dataset of "features"

- Features are:
  - notable or informative characteristics of the text
  - Features are (mostly) fragments of text
Features currently provided:

- **Format:** Per-page features, packaged (as JSONs) in one file per volume, with page section (header, footer and body) identification

  - counts of part-of-speech-tagged words  
    (bag of words, per-page, with frequencies)
  
  - various line-level information:
    - number of lines containing characters of any kind in a page section
    - counts of the initial character and final character of each line in a page section
     
  - etc.
Feature-extraction

- Features are a “translation” of text
  - From language that humans understand
  - To machine-readable fragments

Text as data as text:

raw text → intermediate representation → results of analysis

Tiger, tiger, burning bright...

Tiger (noun): 2
Burning (verb): 1
Bright (adj): 1

features

E.g. topic model (features are reconstellated into larger units)
Features


A library for help with reading with features:

github.com/organisciak/htrc-feature-reader
Ted Underwood in ‘Theorizing Research Practices We Forgot to Theorize Twenty Years Ago’, *Representations*, Summer 2014:

- “Humanists are gearing up to have a conversation about digital research methods; a new kind of interdisciplinary conversation [between humanists and computer scientists] is about to begin, one in which] a rare opportunity is emerging for a genuinely productive exchange between scientific methodology and humanistic theory.”
(At least) two opportunities here for “interdisciplinary conversation” about digital research methods

1) A conversation about how the hermeneutical cycle can work at the level of a collection (rather than a single text) too large to be surveyed by a single reader

2) “Fragments” as data structures and as anxiety-producing artifacts in the history and future of the humanities
Hermeneutic cycle for text collections

In *supervised* topic models (e.g. Blei and McAuliffe, 2007), which uses a supervised learning technique, topics generated are shaped by prior assumptions of the modeler (as communicated in the supervised training phase).
Humanism and (reading) fragments

Fragments of text and reading them have fascinated humanists (and caused them anxiety) at least since the early modern period (when, after the Renaissance, “humanism” in its current form first appeared).
Humanist anxiety produced by text fragments (early modern Europe)

(John Donne, from the poem ‘An Anatomy of the World’, written in 1611)

“But this were light, did our less volume hold All the old text; or had we changed to gold Their silver; or disposed into less glass Spirits of virtue, which then scattered was…”
“Shall I at least set my lands in order?
London Bridge is falling down falling down falling down...
These fragments I have shored against my ruins
Why then Ile fit you. Hieronymo’s mad againe
Datta. Dayadhvam. Damyata…”

Anxiety for a lost whole, the classical past in text fragments

“Grenfell gets so anxious to recover even scraps
It’s brought the poor chap almost close to a collapse...
He heard Apollo yammering for scraps and tatters
Of some lost Sophoclean play called *The Tracking Satyrs.*”

(Tony Harrison, *The Trackers of Oxyrhynchus: The Delphi Text*, written in 1990; about the lost papyri fragments recovered in Oxyrhynchus, Egypt, in 1912)
Humanist anxiety produced by (text and other) fragments (Augustan England)

“The ruins of Rome provided the humanists with a powerful image of the kind of desolation inevitably wrought by innovation, novelty and wilful change.”

Paul Fussell, The Rhetorical World of Augustan Humanism, 1965
Posthumanism and reading via fragments

- Today, technological change and innovation appear to many humanists as a similar threat: “desolation... wrought by innovation.”

  - Anxiety about being rendered marginalized
  - Anxiety about creeping “scientism” and utilitarianism
  - Anxiety/fear of the digital
  - Anxiety about a post-humanist/ anti-humanist) future
Text fragments (as a metaphor) are paradigmatic of much in the digital humanities.

Can computationalism be (neo)humanist?
“Ever since Leibniz, the deployment of one strand of computationalism has been to excise the element of ambiguity which is part of the human experience.”


This need not always be so.
(1) Fragments can be recombined into ludic, ambiguous, open-ended wholes

Computationalism as neo-humanism

By unlocking knowledge resources and accommodating ambiguity, play, and open-ended interpretation, computationalism can also be a neo-humanism.

(2) Non-consumptive reading is post-humanist:

Interfaces can have agency

Re-imagining discursive practices

“Discursive practices are not human-based activities but rather specific material (re)configurings of the world through which local determinations of boundaries, properties and meanings are differentially enacted.”

How you can get involved

HathiTrust Research Center Announcements:

htrc-announce-l @ list.indiana.edu

HathiTrust Research Center User Group:

htrc-usergroup-l @ list.indiana.edu
Resources

Guide to the HTRC Portal:
http://uiuc.libguides.com/htrcguide

HathiTrust + Bookworm:
https://htrcbookworm.wordpress.com/

Getting started with the HTRC Data Capsule:
http://bit.ly/1rWHPfH

Detailed Data Capsule guide: http://bit.ly/1BzP9O1

Feature Extraction:
https://sandbox.htrc.illinois.edu/HTRC-UI-Portal2/Features
More resources:
The HathiTrust Research Center Publications and Presentations page:
https://wiki.htrc.illinois.edu/display/OUT/HTRC+Publications,+Presentations

Questions?
Contact the HathiTrust Research Center Support Team at
htrc-support-l@list.indiana.edu

We are happy to answer your questions and to help you use our resources!