HathiTrust Research Center: Challenges and Opportunities in Big Text Data


Presented by Miao Chen
Research Associate, Data To Insight Center, IU

Beth Plale – @bplale
Professor, School of Informatics and Computing
Director, Data To Insight Center
Indiana University
Thanks to sponsors
HathiTrust Digital Library

- HathiTrust is a partnership of academic & research institutions, offering a collection of millions of titles digitized from libraries around the world.
  - Founding members of HathiTrust along with University of Michigan are Indiana University, University of California, and University of Virginia

http://www.hathitrust.org

⇒ Distinguished from

http://www.hathitrust.org/htrc
HathiTrust repository is a latent goldmine for text mining analysis, analysis of large-scale corpi through computational tools, and time-based analysis.

Restricted nature of HT content suggests need for new forms of access that preserve intimate nature of research investigation while honoring restrictions.

Paradigm: computation takes place close to the data.

Currently Digitized (by 2/11/2014)

- 11,014,179 total volumes
- 5,750,943 book titles
- 286,864 serial titles
- 3,854,962,650 pages
- 494 terabytes
- 130 miles
- 8,949 tons
- 3,637,649 volumes (~33% of total) in the public domain

http://www.hathitrust.org/statistics_info
Mission of HT Research Center

• Research arm of HathiTrust
• Goal: enable researchers world-wide to carry out computational investigation of HT repository through
  – Develop model for access: the ‘workset’
  – Develop tools that facilitate research by digital humanities and informatics communities
  – Develop secure cyberinfrastructure that allows computational investigation of entire copyrighted and public domain HathiTrust repository
• Established: July, 2011
• Collaborative effort of Indiana University, University of Illinois, and HathiTrust
Books => HT => HTRC

Source METS
MARC records
OCR text

Digital library, preservation

HathiTrust METS
MARC records
OCR text

Infrastructure, computational access

Books from research and local libraries, Google/Internet Archive scanned books

HTRC portal service
HTRC data API
... 
HTRC search proxy

#HTRC @HathiTrust
Partitioning the HTRC Collections

• Public domain corpus (~3.6M)
  – Non-Google digitized (~250K)
    • Also referred to as the “open-open” corpus
    • On sandbox
  – Google digitized (~2-3M)
    • Production stack

• In-copyright corpus (~7.5M)
  – in progress
The complexity hiding interface

HTRC

Request

Spatial plots

Statistical plots

Tabular info
Complexity hiding interface

HTRC
@IU @UIUC

Extracted feature sets

Other text, e.g., dictionaries, wiki, twitter

Repository

Text mining tools

Blue Waters
HTRC Timeline

- Phase I: development 01 Jul 2011 – 31 Mar 2013
  - HTRC software and services release v1.0
    [http://sourceforge.net/p/htrc/code/](http://sourceforge.net/p/htrc/code/)
- Phase II: outreach, 01 Apr 2013 - present
  - 2nd HTRC UnCamp Sep ‘13

Attendees of UnCamp’13
HTRC architecture

- Philosophy: computation moves to data
- Web services (REST) architecture and protocols
- WSO2 Registry for worksets and results
- Solr Indexes: full text, MARC, and new metadata
- noSQL (Cassandra) store as volume store
- Authentication using WSO2 Identity Server
- Portal front-end, programmatic access
- Mining tools: currently SEASR
HTRC’s guiding principle to computational access

• No computational action or set of actions on part of users, either acting alone or in cooperation with other users over duration of one or multiple sessions can result in sufficient information gathered from the HT repository to reassemble pages from collection for reading

• Definition disallows collusion between users, or accumulation of material over time.

• Defining “sufficient information”: research has shown need to interact directly with select texts. How much of a text to show? Google withholds from showing to reader every 10th page of a book (Int’l NYTimes Nov 16-17, 2013)
HTRC Secure Capsule Architectural Components

- VM Image Store
- VM Image Manager
- VM Image Builder

Researcher

SSH

Secure Capsule cluster

Research results

Registry Services, worksets

WSO2
1) Researcher requests new VM of type X

2) Image instance is created

3) Researcher install tools onto VM through window on her desktop.

4) Upon run, Secure Capsule: controls I/O behind scenes

Registry Services, worksets

Final location of results is registry
HTRC and library

- Out of the 134 UnCamp participants, 42 were librarians (31%)
- Target audience of HTRC news (22% to librarians)

Metadata!
HTRC metadata

• Volume
  – The central entity for metadata description
  – Books, journal, serials, government docs...

• MARC records (xml uploaded from libraries)
  – Specs: http://www.hathitrust.org/bib_specifications
  – Contains OCLC number (master record number)

• METS (metadata encoding and transmission standard)
  – Source METS files, for preservation (from Google, Internet Archive)
  – HathiTrust METS files, for both preservation and access

• Plus HTRC specific metadata fields
HTRC metadata

You searched for:
information representation

Showing item 1 of 869,769 from your search.

Health and Human Services: update on Hispanic representation in HHS Region VIII: fact sheet for the Honorable Timothy E. Wirth, U.S. Senate / United States General Accounting Office.

Full View

Title: Health and Human Services: update on Hispanic representation in HHS Region VIII: fact sheet for the Honorable Timothy E. Wirth, U.S. Senate / United States General Accounting Office.
Subtitle: Update on Hispanic representation in HHS Region VIII.
Author: United States. General Accounting Office.
Language: English
Published: 1992
Country: United States
Call number: HD4903.5.U6 A3823
OCLC: (OCColC)ocm27370022
Source: University of Michigan
Volume ID: mdp.39015048657141

Character count: 21,008
Page count: 24
Volume size by page: Small
Volume size by word: Small
Word count: 3,261

Search result page
HTRC specific metadata

Work set display page

Available Worksets

AncientGreek
biglow_sequpload
nlp
nlpontology_classlabel
darwin-test
2darwin-english
2evelalus
Anarchism
Austen_Dickens_Labels
Author_Twain
BestCoreComplex
BigLaw
Bleakhouse
Cicero_Orationes_Letters
Coffee_Books
Dickens_as_Authors
Dickens_yo
Diderot-test
DocSouthMatch
EPAPoeAsAuthor
ECCOMatch
EBBOmatch
Edgar_Allan_Poe
Elizabeth_Gaskell_Works

Workset Details

Name: AncientGreek
Description: query=ancient greek
Author: miao
Last Modified By: miao
Last Modified Time: 2013-09-03T15:37:29-04:00
Number of Volumes: 20

Books which we recommend to our members who intend to visit Greece.

Volume Id: yale:390002089540237
Authors:
Page Count[HTRC]: 4
Word Count[HTRC]: 806

The claim of antiquity, with an annotated list of books for those who know neither Latin nor Greek, issued by the councils of the Societies for the promotion of Hellenic and Roman studies and of the Classical association.

Volume Id: mdp:39015013434559
Authors:
Page Count[HTRC]: 36
Word Count[HTRC]: 7999

#HTRC @HathiTrust
Metadata Enhancement

• Big data needs good metadata
• Current metadata fields are MARC-based
  – E.g. publication date, authors, title, subject
• MARC fields are fundamental
• Needed more fields of users’ interest for granular analytics (Metadata Enhancement)
• Solicit user requirements and prioritize for implementation
  – Mainly digital humanities uses now
Top Metadata Enhancement Items

• 1st round user requirement collection, top 3 items were metadata related:
  – Word frequency count and document length for a volume
  – Metadata de-duplication
  – Author Gender Analysis

• We have added word count and gender fields to HTRC metadata, and more are being planned and investigated.
HTRC specific metadata: Gender

<table>
<thead>
<tr>
<th>Name: Austen_Dickens_Labels</th>
<th>Author: Austen, Jane</th>
<th>Description:</th>
<th>Last Modified By: erhtec</th>
<th>Last Modified Time: 2013-09-03T14:46:32-04:00</th>
<th>Number of Volumes: 12</th>
</tr>
</thead>
</table>

1. Emma / by Jane Austen.
   - Volume Id: tnp.3343074943568
   - Author: Austen, Jane, 1775-1817
   - Page Count [HTRC]: 130
   - Word Count [HTRC]: 82,356

   - Volume Id: tnp.3343074943568
   - Author: Austen, Jane, 1775-1817
   - Page Count [HTRC]: 128
   - Word Count [HTRC]: 82,658

3. David Copperfield ...
   - Volume Id: tnp.3343074954560
   - Author: Dickens, Charles, 1812-1870
   - Page Count [HTRC]: 132
   - Word Count [HTRC]: 30,656

   - Volume Id: tnp.3343080377706
   - Author: Austen, Jane, 1775-1817
   - Page Count [HTRC]: 222
   - Word Count [HTRC]:

Gender field
Other Metadata Enhancement Items

- Stats analysis: tf-idf
- Readability score
- Language
- Topic modeling (e.g. LDA probability)
- Genre
- Era of compilation
- Book length (e.g. short or long)
- Concordance index (indexing with context)
Portal and Work Set Builder

- https://htrc2.pti.indiana.edu/HTRC-UI-Portal2/
Data API

• Data retrieval
  – [http://wiki.htrc.illinois.edu/display/COM/Python+client+for+accessing+volumes+in+bulk+through+HTRC+Data+API](http://wiki.htrc.illinois.edu/display/COM/Python+client+for+accessing+volumes+in+bulk+through+HTRC+Data+API)
  – Demo code available in Python and Java
  – Page-level and volume-level word count
  – Option: Concatenate pages
  – Option: return METS metadata, of which MARC record is a part
Some Challenges

• Internationalization
  – Non-western text, especially for text processing
• Users contributed code
  – How to create a mechanism to describe and archive such contributions from the community?
• OCR errors
  – An experiment shows avg number of errors per page is 0.57
  – Crowdsourcing? Automatic correction?
• Implications
  – HTRC as a service or resource to library?
Going beyond the volume level?

• Work set level
  – Should be richer than a list of volume ids
  – The resources that scholars work with
  – Can be within HTRC corpus, or not
  – How to formalize it?
  – What’s the metadata for work set?
Going beyond the volume level?

• Page level
  – Emerges as an important unit of analysis in the recent UnCamp in early Sep
  – Important to scholars in some circumstances
  – What are the important metadata fields?
  – E.g. Word frequency count, image/illustration
Recent Updates

• “The HTRC drafted documents covering system architecture, workflows, security measures, and data use cases in preparation for offering “non-consumptive” access to in-copyright volumes in the HathiTrust repository. “

Cited from http://www.hathitrust.org/updates_january2014