Opportunities and Challenges of Text Mining HathiTrust Digital Library

Koninklijke Bibliotheek | 15.Nov.13

Beth Plale – @bplale
Professor, School of Informatics and Computing
Director, Data To Insight Center
Indiana University

Tweet us - @HathiTrust #HTRC
Thanks to sponsors

ALFRED P. SLOAN FOUNDATION

INDIANA UNIVERSITY

ILLINOIS

The Andrew W. Mellon Foundation
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

http://www.hathitrust.org/htrc
HathiTrust repository is a latent goldmine for text mining analysis, analysis of large-scale corpora 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.
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 and University of Illinois
HTRC

All the complexity

Complexity hiding interface

Request

Spatial plots

Statistical plots

Tabular info
Complexity hiding interface

HTRC
@IU @UIUC

TEXT MINING TOOLS
mahout
R
seasr

HATHI TRUST
REPOSITORY

EXTRACTED FEATURE SETS

OTHER TEXT, E.G., DICTIONARIES, WIKI, TWITTER

BLUE WATERS
1. "All's well!" : some helpful verse for these dark days of war / by John Oxenham.

   Title:  "All's well!" : some helpful verse for these dark days of war / by John Oxenham.
   Author:  Oxenham, John.
   Language:  English
   Published:  1915
HTRC Timeline

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

- Philosophy: computation moves to data
- Web services (REST) architecture and protocols
- WS02 Registry for worksets and results
- Solr Indexes: full text, MARC, and new metadata
- noSQL (Cassandra) store as volume store
- Authentication using WS02 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 withholding from showing to reader every 10th page of a book (Int’l NYTimes Nov 16-17, 2013)
Researcher requests new VM

Researcher

SSH
Research install tools onto VM through window on her desktop

Secure Capsule: controls I/O behind scenes

Final location of results is registry
Registry Services, worksets

HTRC Secure Capsule Architecture

VM Image Store
VM Image Manager
VM Image Builder
VM Manager

Research results
Sampling of use

1. Metadata enhancement
   Miao Chen, i-school, Indiana Univ

2. Large scale data analytics
   Guangchen Ruan, computer science, Indiana Univ

3. Gender author identification
   Stacy Kowalczyk, library science, Dominican Univ

4. Topic modeling to identify philosophical arguments in scientific texts
   Colin Allen et al., cognitive science, Indiana Univ
METADATA ENHANCEMENT:
PRELIMINARY STUDY

Miao Chen, PhD, Indiana University
Metadata Enhancement

• 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

Thanks to Miao Chen
Top Metadata Enhancement Items

• 1st round user survey, top requested items
  – Word frequency count and document length. At volume level ✔
  – Author gender identification ✔
  – Metadata de-duplication
  – Word frequency count at page level
  – Word frequency count for full 10.8 M volume repository
Other Metadata Enhancement Items

- Stats analysis: TF-IDF
- Readability score
- Language identification
- Topic modeling (e.g. LDA probability)
- Genre
- Era of compilation
- Book length (e.g. short or long)
- Concordance index (indexing with context)
Guangchen Ruan, PhD candidate, IU

LARGE SCALE DATA ANALYSIS ON XSEDE
Experimental Environment and Results

- Dataset
  2,592,210 volumes, in total 2.1 TB, divided into 1024 partitions of 2GB each

- Computation platform
  XSEDE Blacklight, 1024-core, each 2.27 GHz, 8.2 TB memory. Each core processes one partition

- Results
  Whole corpus word count completed in 1,454 seconds or 24.23 minutes
Computation Time Distribution

Computation time distribution of 1024-core (100-bin)
Word Frequency Distribution

Word frequency distribution

Frequency

# of words
Stacy Kowalczyk, Asst. Professor, Dominican University
Zong Peng, HTRC, Indiana University

GENDER IDENTIFICATION OF HTRC AUTHORS BY NAMES

Gender Identification of Text

• Can we use author names in bibliographic records to identify gender?
• 2.6 million bibliographic records
  – Extracted personal author data
  – Marc 100 abcd and 700 abcd
• 606,437 unique personal author strings
• Bibliographic data is not fielded like patent names
• Relying on Standard cataloging practice
  – Last name, first name middle name, titles/honorifics, dates
Authors vs Names

- Methuen, Algernon Methuen Marshall, Sir bart., 1856-1924
- Methuem, Algernon
- Methuen Algernon
- Methuen Marshall, Sir, bart., 1856-
- Methuen, A. Sir, 1856-1924
- Methuen, A. Sir, bart., 1856-1924
- Methuen Marshall, Sir bart 1856-1924
- Methuen, Algernon Methuen Marshall, Sir, 1856-1924
- Methuen, Algernon Methuen Marshall, Sir, bart., 1856-1924
- Methuen, Algernon, 1856-1924
Sources of Data

• The Virtual International Authority File
  – Hosted by OCLC

• Harvested names from multiple data sources
  – Census bureau
  – Baby name sites

• EU Patent Research names list (Frietsch et al, 2009; Naldi et al. 2005)
  – Developed an extensive list of European names

• Titles and honorifics
  – Multiple web resources
  – Sir, Baron, Count, Duke, Father, Cardinal, etc
  – Lady, Mrs. Miss, Countess, Duchess, Sister, etc
Initial Gender Results

- Approximately 80% of name strings have initial gender identification
  - Female
    - 59,365
    - 10%
  - Male
    - 425,994
    - 70%
  - Unknown
    - 114,204
    - 19%
  - Ambiguous
    - 5,965
    - Less than 1%
Results by Data Source

Against the whole set of name strings

- VIAF
  - 19% hit rate

- Web Names
  - 54% hit rate

- Patents Names
  - 8%
Colin Allen, Jamie Murdock  
Cognitive Science, Indiana University
Digging into philosophy of science

- Establish points of contact between philosophy and science: where philosophical arguments on anthropomorphism appear in science texts
- Use topic modeling to identify the volumes and pages within these volumes that are “rich” in a chosen topic
- Use semi-formal discourse analysis technique to identify key arguments in selected pages to incrementally expose and represent argument structures
The How

• 1315 volumes from HTRC selected using keyword search for ‘darwin’, ‘romanes’, ‘anthropomorphism’, and ‘comparative psychology’
• Set contains lots of uninteresting books: e.g., college course catalogs
• Apply LDA on 86 volume subset
• Using iPy Notebook
LDA topic modeling

- LDA (Latent Dirichlet Analysis) uses a Bayesian updating method to generate a set of “topics” – probability distributions over set of terms in a corpus
- Number of topics is a parameter in the modeling technique
- Method finds set of topics that is best able to reproduce the term distributions in documents belonging to the corpus
- Documents may be whole volumes, chapters, articles, single pages, even individual sentences – modeler’s choice
Volume level topic modeling on ‘anthropomorphism’ yields set of topics

In [34]: # finding related topics using a single term
   v1.sim_word_top('anthropomorphism')

Out[34]:
```
<table>
<thead>
<tr>
<th>Topic</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>38</td>
<td>god, religion, life, man, religious, spirit, world, nature, spiritual, divine</td>
</tr>
<tr>
<td>16</td>
<td>animals, evolution, life, animal, development, man, species, cells, living, theory</td>
</tr>
<tr>
<td>51</td>
<td>philosophy, nature, knowledge, world, thought, idea, things, reason, truth, science</td>
</tr>
<tr>
<td>58</td>
<td>man, among, tribes, primitive, men, people, also, races, women, race</td>
</tr>
<tr>
<td>21</td>
<td>social, life, new, mind, upon, individual, human, mental, world, subfield</td>
</tr>
<tr>
<td>12</td>
<td>child, children, first, development, movements, play, life, little, mental, mother</td>
</tr>
<tr>
<td>11</td>
<td>motion, force, must, forces, matter, changes, us, parts, like, evolution</td>
</tr>
<tr>
<td>31</td>
<td>gods, religion, p, name, see, god, india, ancient, one, worship</td>
</tr>
<tr>
<td>1</td>
<td>pp, der, vol, die, de, des, und, ibid, university, la</td>
</tr>
</tbody>
</table>
```
.. Of set of topics, choose ‘16’ as best
Volumes most similar to topic 16

```
In [35]: # display the documents most similar to (best predicted) by one topic
   : v1.sim_top_doc(16, print_len=20, label_fn=htrc_label_link_fn_1315)

Out[35]:

<table>
<thead>
<tr>
<th>Document</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>The riddle of the universe at the close of the nineteenth ce,</td>
<td>0.86750</td>
</tr>
<tr>
<td><a href="http://hdl.handle.net/2027/locark:/13960/t2s47h57b">http://hdl.handle.net/2027/locark:/13960/t2s47h57b</a></td>
<td></td>
</tr>
<tr>
<td>The riddle of the universe at the close of the nineteenth ce,</td>
<td>0.86704</td>
</tr>
<tr>
<td><a href="http://hdl.handle.net/2027/uc2ark:/13960/t5v69b880">http://hdl.handle.net/2027/uc2ark:/13960/t5v69b880</a></td>
<td></td>
</tr>
<tr>
<td>The riddle of the universe at the close of the nineteenth ce,</td>
<td>0.86254</td>
</tr>
<tr>
<td><a href="http://hdl.handle.net/2027/locark:/13960/t2s47h57b">http://hdl.handle.net/2027/locark:/13960/t2s47h57b</a></td>
<td></td>
</tr>
<tr>
<td>The germ-plasm : a theory of heredity /,</td>
<td>0.84662</td>
</tr>
<tr>
<td><a href="http://hdl.handle.net/2027/uc2ark:/13960/t0qr4pc9z">http://hdl.handle.net/2027/uc2ark:/13960/t0qr4pc9z</a></td>
<td></td>
</tr>
<tr>
<td>Last words on evolution; a popular retrospect and summary,,</td>
<td>0.83998</td>
</tr>
<tr>
<td><a href="http://hdl.handle.net/2027/nyp.33433081629184">http://hdl.handle.net/2027/nyp.33433081629184</a></td>
<td></td>
</tr>
<tr>
<td>Biology and its makers, with portraits and other illustratio,</td>
<td>0.81621</td>
</tr>
<tr>
<td><a href="http://hdl.handle.net/2027/uc2ark:/13960/t1td9pw8v">http://hdl.handle.net/2027/uc2ark:/13960/t1td9pw8v</a></td>
<td></td>
</tr>
<tr>
<td>The psychic life of micro-organisms. A study in experimental,</td>
<td>0.74668</td>
</tr>
<tr>
<td><a href="http://hdl.handle.net/2027/uc2ark:/13960/t73t9h556">http://hdl.handle.net/2027/uc2ark:/13960/t73t9h556</a></td>
<td></td>
</tr>
<tr>
<td>The psychic life of micro-organisms : a study in experimenta,</td>
<td>0.73637</td>
</tr>
<tr>
<td><a href="http://hdl.handle.net/2027/uc2ark:/13960/t75t3j390">http://hdl.handle.net/2027/uc2ark:/13960/t75t3j390</a></td>
<td></td>
</tr>
</tbody>
</table>
```
Repeat LDA at page level
Topic model at page level for topics anthropomorphism, animal, and psychology

```
In [36]: # finding related topics using multiple terms
v1.sim_word_top(['anthropomorphism', 'animal', 'psychology'])

Out[36]:
<table>
<thead>
<tr>
<th>Topic</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>26</td>
<td>consciousness, experience, p, psychology, process, individual, object, activity, relation, feeling</td>
</tr>
<tr>
<td>16</td>
<td>animals, evolution, life, animal, development, man, species, cells, living, theory</td>
</tr>
<tr>
<td>10</td>
<td>animals, water, animal, food, birds, one, leaves, insects, species, many</td>
</tr>
<tr>
<td>1</td>
<td>pp, der, vol, die, de, des, und, ibid, university, la</td>
</tr>
<tr>
<td>58</td>
<td>man, among, tribes, primitive, men, people, also, races, women, race</td>
</tr>
<tr>
<td>47</td>
<td>college, university, professor, school, law, work, students, degree, education, new</td>
</tr>
<tr>
<td>25</td>
<td>nature, ii, us, mr, without, life, human, natural, language, every</td>
</tr>
<tr>
<td>29</td>
<td>fig, two, body, form, cells, animals, first, ii, side, organs</td>
</tr>
<tr>
<td>12</td>
<td>child, children, first, development, movements, play, life, little, mental, mother</td>
</tr>
<tr>
<td>21</td>
<td>social, life, new, mind, upon, individual, human, mental, world, subfield</td>
</tr>
<tr>
<td>4</td>
<td>acid, water, body, action, blood, food, alcohol, air, substances, work</td>
</tr>
</tbody>
</table>
```
# Words sorted by similarity

<table>
<thead>
<tr>
<th>Topic</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>animals, evolution, life, animal, development, man, species, cells, living, theory</td>
</tr>
<tr>
<td>10</td>
<td>animals, water, animal, food, birds, one, leaves, insects, species, many</td>
</tr>
<tr>
<td>26</td>
<td>consciousness, experience, p, psychology, process, individual, object, activity, relation, feeling</td>
</tr>
<tr>
<td>58</td>
<td>man, among, tribes, primitive, men, people, also, races, women, race</td>
</tr>
<tr>
<td>25</td>
<td>nature, ii, us, mr, without, life, human, natural, language, every</td>
</tr>
<tr>
<td>29</td>
<td>fig, two, body, form, cells, animals, first, ii, side, organs</td>
</tr>
<tr>
<td>4</td>
<td>acid, water, body, action, blood, food, alcohol, air, substances, work</td>
</tr>
<tr>
<td>1</td>
<td>pp, der, vol, die, de, des, und, ibid, university, la</td>
</tr>
<tr>
<td>45</td>
<td>brain, nerve, fibres, nervous, motor, nerves, sensory, cord, cells, spinal</td>
</tr>
<tr>
<td>12</td>
<td>child, children, first, development, movements, play, life, little, mental, mother</td>
</tr>
<tr>
<td>20</td>
<td>man, moral, men, good, law, society, social, conduct, action, pleasure</td>
</tr>
<tr>
<td>31</td>
<td>gods, religion, p, name, see, god, india, ancient, one, worship</td>
</tr>
<tr>
<td>47</td>
<td>college, university, professor, school, law, work, students, degree, education, new</td>
</tr>
<tr>
<td>11</td>
<td>motion, force, must, forces, matter, changes, us, parts, like, evolution</td>
</tr>
</tbody>
</table>
Pick top 3: topics 16, 10, 26

<table>
<thead>
<tr>
<th>Topic</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>animals, evolution, life, animal, development, man, species, cells, living, theory</td>
</tr>
<tr>
<td>10</td>
<td>animals, water, animal, food, birds, one, leaves, insects, species, many</td>
</tr>
<tr>
<td>26</td>
<td>consciousness, experience, p, psychology, process, individual, object, activity, relation, feeling</td>
</tr>
<tr>
<td>58</td>
<td>man, among, tribes, primitive, men, people, also, races, women, race</td>
</tr>
<tr>
<td>25</td>
<td>nature, ii, us, mr, without, life, human, natural, language, every</td>
</tr>
<tr>
<td>29</td>
<td>fig, two, body, form, cells, animals, first, ii, side, organs</td>
</tr>
<tr>
<td>4</td>
<td>acid, water, body, action, blood, food, alcohol, air, substances, work</td>
</tr>
<tr>
<td>1</td>
<td>pp, der, vol, die, de, des, und, ibid, university, la</td>
</tr>
<tr>
<td>45</td>
<td>brain, nerve, fibres, nervous, motor, nerves, sensory, cord, cells, spinal</td>
</tr>
<tr>
<td>12</td>
<td>child, children, first, development, movements, play, life, little, mental, mother</td>
</tr>
<tr>
<td>20</td>
<td>man, moral, men, good, law, society, social, conduct, action, pleasure</td>
</tr>
<tr>
<td>31</td>
<td>gods, religion, p, name, see, god, india, ancient, one, worship</td>
</tr>
<tr>
<td>47</td>
<td>college, university, professor, school, law, work, students, degree, education, new</td>
</tr>
<tr>
<td>11</td>
<td>motion, force, must, forces, matter, changes, us, parts, like, evolution</td>
</tr>
</tbody>
</table>
Show documents of topics 10, 16, 26

<table>
<thead>
<tr>
<th>Document</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Secrets of animal life, <a href="http://hdl.handle.net/2027/uc2.ark:/13960/t7wm15g73">URL</a></td>
<td>0.63954</td>
</tr>
<tr>
<td>Comparative studies in the psychology of ants and of higher, <a href="http://hdl.handle.net/2027/uc2.ark:/13960/t6057f659">URL</a></td>
<td>0.63085</td>
</tr>
<tr>
<td>The colours of animals, their meaning and use, especially co, <a href="http://hdl.handle.net/2027/uc2.ark:/13960/t9t14w82w">URL</a></td>
<td>0.55333</td>
</tr>
<tr>
<td>The foundations of normal and abnormal psychology, <a href="http://hdl.handle.net/2027/loc.ark:/13960/t9m33nm99">URL</a></td>
<td>0.54171</td>
</tr>
<tr>
<td>The bird rookeries of the Tortugas, <a href="http://hdl.handle.net/2027/uc2.ark:/13960/t3pv6cc9j">URL</a></td>
<td>0.53789</td>
</tr>
<tr>
<td>Mind in animals, <a href="http://hdl.handle.net/2027/mdp.39015005169357">URL</a></td>
<td>0.53783</td>
</tr>
<tr>
<td>Ants and some other insects; an inquiry into the psychic pow, <a href="http://hdl.handle.net/2027/wu.89095158218">URL</a></td>
<td>0.53606</td>
</tr>
<tr>
<td>Systematic science teaching ; a manual of inductive elementa, <a href="http://hdl.handle.net/2027/uc2.ark:/13960/t11n8195t">URL</a></td>
<td>0.53152</td>
</tr>
<tr>
<td>The riddle of the universe at the close of the nineteenth ce, <a href="http://hdl.handle.net/2027/uc2.ark:/13960/t5v69b880">URL</a></td>
<td>0.52804</td>
</tr>
</tbody>
</table>
Drop to sentence level

• Select three books with highest aggregate of 20-40 topic-relevant pages for more precise analysis
• Manually augment argument analysis
  – Remodeling of three volumes at sentence level
  – Training other methods using human analysis plus sentence similarity
Remodeling of 3 volumes at sentence level
Promising early results ...

What did we get? -- tokenized sentences (word lists) followed by original text...

```
In [48]: #tok_sents

In [49]: orig_sents

Out[49]: ['Every statement that another being possesses \npsychic qualities is a conclusion from analogy, not a certainty; it is a matter of faith.',
        'If any consciousness \naccompanies it, then the nearest human analogy to such \nconsciousness is to be found in organic sensations, and these, \nas has just been said, must necessarily be in the human mind \nwholly different in quality from anything to be found in an \nanimal whose structure is as simple as the Amoeba\'s."
        '; learning, 208, \n\n214.',
        'On the other \n\n1).',
        'Dytiscus, 86.',
        'Burnett, 126, 170.',
        'Willem, 130, 192.',
        'Caterpillars, 192, 196.',
        'Murbach, 107, 158.',
        'Fancy, for example, one of us entering a \nroom in the dark and groping about among the furniture.','
        'This, of \ncourse, does not refer to the power to judge distance.','','
• For details [http://www.hathitrust.org/htrc/faq](http://www.hathitrust.org/htrc/faq)

• General contact info
  – Beth Plale, Director HTRC, [plale@indiana.edu](mailto:plale@indiana.edu)

• Requests for capability, interest
  – Miao Chen, Asst. Director HTRC
    [miaochen@indiana.edu](mailto:miaochen@indiana.edu)