2020

Building Collections for Analysis with HathiTrust

Matthew Steele

Follow this and additional works at: https://researchrepository.wvu.edu/grad-immersion

Part of the Library and Information Science Commons
Building Collections for Analysis with HathiTrust

MATTHEW STEELE, GOVERNMENT INFORMATION LIBRARIAN
DOWNTOWN CAMPUS LIBRARY
MATTHEW.STEELE1@MAIL.WVU.EDU
HathiTrust

Membership organization (140+ partners)

“Collecting, organizing, preserving, communicating, and sharing the record of human knowledge”

Large-scale digitization initiative at research libraries

- Members contribute >1 million new volumes/year
HathiTrust Data

16,744,704 volumes (books and serials)

Representing content in wide variety of disciplines

50% in English; though 450+ languages, ancient and modern

2/3 of the collection is protected by copyright or access restriction

Stats updated daily: https://www.hathitrust.org/about
What is Text Analysis/Mining

Using computers to reveal information in and about text (Hearst, 2003)

Algorithms discern patterns

Text may be "unstructured"

More than just search
HathiTrust Research Center (HTRC)

Provides access for computational text analysis to HathiTrust

Collaborative effort at Indiana University and the University of Illinois, with support from the University of Michigan through the HathiTrust
HTRC Mission

To provide infrastructure and tools enabling and supporting computational research on the more than 16 million volumes of the HathiTrust collection. HTRC enable scholars to fully utilize content of HathiTrust under Fair Use, while preventing violations of U.S. copyright law.
Three Approaches

1. **HTRC Analytics** (for web-based tools, including **Bookworm**)  

2. **Feature Extraction Services** (including downloadable data sets)  

3. **Secure Data Capsule** access

[https://analytics.hathitrust.org/](https://analytics.hathitrust.org/)
HTRC Overview

HTRC Analytics

- Algorithms
- Worksets
- Data Capsules
- Bookworm
- Extracted Features Dataset

HathiTrust Data
Non-Consumptive Research Policy

**Not** research in which a researcher reads or displays substantial portions of an in-copyright or rights-restricted volume to understand the expressive content presented within that volume.

More here: [https://www.hathitrust.org/htrc_ncup](https://www.hathitrust.org/htrc_ncup)
HTRC Algorithms

**Input Parameters**

<table>
<thead>
<tr>
<th>Name</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>iter</td>
<td>200</td>
</tr>
<tr>
<td>Input_collection</td>
<td>pol.science/DDRF@leidenuni.org/set1</td>
</tr>
<tr>
<td>k</td>
<td>20 60</td>
</tr>
</tbody>
</table>

**Output**

- **InPhO Topic Explorer**

**Output**

<table>
<thead>
<tr>
<th>vo_id</th>
<th>page_seq</th>
<th>entity</th>
<th>type</th>
</tr>
</thead>
<tbody>
<tr>
<td>mdp.39015037380379</td>
<td>00000007</td>
<td>December 1, 1966</td>
<td>DATE</td>
</tr>
<tr>
<td>mdp.39015037380379</td>
<td>00000007</td>
<td>1967</td>
<td>DATE</td>
</tr>
<tr>
<td>mdp.39015037380379</td>
<td>00000007</td>
<td>Regal</td>
<td>MISC</td>
</tr>
<tr>
<td>mdp.39015037380379</td>
<td>00000007</td>
<td>AFRICA</td>
<td>ORGANIZATION</td>
</tr>
<tr>
<td>mdp.39015037380379</td>
<td>00000008</td>
<td>Algeria</td>
<td>LOCATION</td>
</tr>
<tr>
<td>mdp.39015037380379</td>
<td>00000008</td>
<td>MPLA</td>
<td>ORGANIZATION</td>
</tr>
<tr>
<td>mdp.39015037380379</td>
<td>00000008</td>
<td>Sudanese</td>
<td>MISC</td>
</tr>
<tr>
<td>mdp.39015037380379</td>
<td>00000008</td>
<td>1966</td>
<td>DATE</td>
</tr>
<tr>
<td>mdp.39015037380379</td>
<td>00000008</td>
<td>Independence four years ago</td>
<td>DATE</td>
</tr>
<tr>
<td>mdp.39015037380379</td>
<td>00000008</td>
<td>Ghana</td>
<td>LOCATION</td>
</tr>
<tr>
<td>mdp.39015037380379</td>
<td>00000008</td>
<td>Central Committee</td>
<td>ORGANIZATION</td>
</tr>
<tr>
<td>mdp.39015037380379</td>
<td>00000008</td>
<td>Amin Shaker</td>
<td>PERSON</td>
</tr>
<tr>
<td>mdp.39015037380379</td>
<td>00000008</td>
<td>November 29, 1966</td>
<td>DATE</td>
</tr>
<tr>
<td>mdp.39015037380379</td>
<td>00000008</td>
<td>Egyptian Gazette</td>
<td>MISC</td>
</tr>
<tr>
<td>mdp.39015037380379</td>
<td>00000008</td>
<td>1966</td>
<td>DATE</td>
</tr>
<tr>
<td>mdp.39015037380379</td>
<td>00000008</td>
<td>November</td>
<td>DATE</td>
</tr>
</tbody>
</table>
HTRC Extracted Features Dataset

Volume and page metadata

Feature data, including word & word counts

https://analytics.hathitrust.org/e xplore
HTRC Extracted Features Updates

Feature data for 15.7 million volumes now available

Pre-processed data derived from the raw text

- Positions researcher to begin work, without exposing full text
HathiTrust+Bookworm

https://bookworm.htrc.illinois.edu/
How does HTRC support researchers?

Research Support

Dataset and workset creation
  ◦ For those who need specialized support

Monthly office hours
  ◦ Second Wednesday of every month at 3 p.m. ET
  ◦ go.illinois.edu/htrchelp-live

Email: htrc-help@hathitrust.org
Thank you!