BitCurator NLP: Natural Language Processing for Archivists

SAA Research Forum
August 14, 2018
Washington, DC

Cal Lee
University of North Carolina at Chapel Hill
School of Information and Library Science
BitCurator

- Funded by Andrew W. Mellon Foundation
  - Phase 1: October 1, 2011 – September 30, 2013
  - Phase 2 – October 1, 2013 – September 30, 2014
- Partners: School of Information and Library Science (SILS) at UNC and Maryland Institute for Technology in the Humanities (MITH)
October 1, 2014 – September 30, 2016 at School of Information and Library Science, University of North Carolina at Chapel Hill

Funded by Andrew W. Mellon Foundation

Developed open-source software to support access to disk images. Core areas of focus:

- Web access services for disk images
- Redacting complex born-digital objects (disk images)
- Emulated access to data from disk images
BitCurator Access Redaction Tools

• Software to redact strings and byte sequences identified in disk images

• Three types of redaction actions:
  • SCRUB (overwrite the bytes in the target with zeroes),
  • FILL (overwrite by filling with a given character),
  • FUZZ (altering the content of a binary, so it can no longer run).

• Best used through a command-line interface but also include a graphic user interface (GUI) that supports the same functions

• Python API allowing institutions to develop custom redaction facilities using open-source tools including lightgrep

https://github.com/bitcurator/bitcurator-access-redaction
• Funded by Andrew W. Mellon Foundation: October 1, 2016 – September 30, 2018
• Use natural language processing (NLP) to identify and report on items relevant to preservation, information organization, and access
  • Entities (e.g. persons, places, and organizations)
  • Relationships among entities (e.g. appear together within documents or set of documents)
  • Topic models to provide insight into how concepts are clustered
The Challenge of Diverse Archival Collections

• Most application of NLP is based on relatively consistent, clean collections, but...

• Archival collections can contain:
  • Raw or forensically packaged disk images
  • Many different file types
  • Limited or inconsistent metadata
## Directory Listing

<table>
<thead>
<tr>
<th>Type</th>
<th>Filename</th>
<th>Bytes</th>
<th>Created</th>
<th>Modified</th>
<th>Download</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TESTFAT (Volume Label Entry)</td>
<td>0.0B</td>
<td>N/A</td>
<td>2013-05-02T16:01:26</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>.</em>.Trashes</td>
<td>4.0KB</td>
<td>N/A</td>
<td>2013-05-02T14:11:00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>_RASHE~1.YEN</td>
<td>0.0B</td>
<td>N/A</td>
<td>2013-05-02T14:11:00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.Trashes</td>
<td>4.0KB</td>
<td>N/A</td>
<td>2013-05-02T14:11:00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.Spotlight-V100</td>
<td>4.0KB</td>
<td>N/A</td>
<td>2013-05-02T14:11:00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>06311397.pdf</td>
<td>4.0KB</td>
<td>N/A</td>
<td>2013-05-02T14:11:00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2013-02-20_AAFS.pdf</td>
<td>6.2MB</td>
<td>N/A</td>
<td>2013-05-02T15:56:36</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.fsevents</td>
<td>4.0KB</td>
<td>N/A</td>
<td>2013-05-02T14:11:00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$MBR</td>
<td>512.0B</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$FAT1</td>
<td>953.0KB</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$FAT2</td>
<td>953.0KB</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$OrphanFiles</td>
<td>0.0B</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
</tr>
</tbody>
</table>

Showing 1 to 12 of 12 entries
Tolerate some noise / errors in access copies
Intertopic Distance Map (via multidimensional scaling)

Top-30 Most Relevant Terms for Topic 6 (9.9% of tokens):
- state
- law
- rights
- international
- human
- made
- court
- states
- council
- board
- president
- parliament
- high
- office
- college
- house
- web
- service
- space
- general
- case
- respect
- person
- members
- power
- legislature
- governor
- district
- legal

Marginal topic distribution

Overall term frequency
Estimated term frequency within the selected topic
Thank you!

https://bitcurator.net/

https://bitcurator.github.io/