Web based computer-assisted text analysis

Christopher Ohge, Martin Steer
Riga Technical University, September 2019
Hathi Trust Research Centure (HTRC)

https://pro.europeana.eu/post/text-mining-2-how-scholars-can-support-digital-libraries
Hathi Trust Research Centure (HTRC)

- Promotes scholarly use of the HathiTrust Digital Library
- Custodian of Google books
  - OCR corpus – many character errors!
  - Biased – e.g. C19th lacks genre metadata
- Provides
  - Data analysis infrastructure
  - “Non-consumptive” use
  - Extracted Features Dataset
  - Visualisation tools
  - Lots more…

https://pro.europeana.eu/post/text-mining-2-how-scholars-can-support-digital-libraries
HTRC Extracted Features Dataset

• Pre-extracted features
• Python/R libraries, API access
• Volume-level
  • Bibliographic metadata
    • title, pubDate, language,
    • genre, typeOfResource
    • names, imprint, rightsAttributes
  • pageCount, size

Data Stats

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td># of volumes</td>
<td>15,722,079</td>
</tr>
<tr>
<td># of pages</td>
<td>5,787,519,444</td>
</tr>
<tr>
<td># of tokens</td>
<td>2,449,739,213,773</td>
</tr>
<tr>
<td># of IC volumes</td>
<td>9,914,509</td>
</tr>
<tr>
<td># of IC pages</td>
<td>3,005,448,348</td>
</tr>
<tr>
<td># of IC tokens</td>
<td>1,777,793,828,310</td>
</tr>
<tr>
<td># of PD volumes</td>
<td>5,807,570</td>
</tr>
<tr>
<td># pd pages</td>
<td>2,602,212,586</td>
</tr>
<tr>
<td># pd tokens</td>
<td>1,197,838,539,662</td>
</tr>
</tbody>
</table>

https://wiki.htrc.illinois.edu/display/COM/ExtractedFeaturesDataset
HTRC Extracted Features Dataset

• Page level
  • Header, Body, Footer info
  • Page sequence numbers
  • tokenCount, lineCount, sentenceCount
  • Languages (inferred per page)
  • tokenPosCount (POS tags)
  • beginCharCounts, endCharCounts

```json
{
  "id": "uc1.b2617728",
  "metadata": {
    "schemaVersion": "1.3",
    "dateCreated": "2016-06-25T03:45:05.180536Z",
    "volumeIdentifier": "uc1.b2617728",
    "accessProfile": "google",
    "rightsAttributes": "ic",
    "hathitrustRecordNumber": "10281140",
    "enumerationChronology": "",
    "sourceInstitution": "UC",
    "sourceInstitutionRecordNumber": ".b165",
    "oclc": [
      "70545428"
    ],
    "isbn": []
  }
}```
HTRC Extracted Features Dataset

• Page level
  • Header, Body, Footer info
  • Page sequence numbers
  • tokenCount, lineCount, sentenceCount
  • Languages (inferred per page)
  • tokenPosCount (POS tags)
  • beginCharCounts, endCharCounts

```json
{
  "seq": "00000033",
  "tokenCount": 273,
  "lineCount": 36,
  "emptyLineCount": 0,
  "sentenceCount": 11,
  "header": {
    "tokenCount": 0,
    "lineCount": 0,
    "emptyLineCount": 0,
    "sentenceCount": 0,
    "capAlphaSeq": 0,
    "beginCharCounts": {},
    "endCharCount": {},
    "tokenPosCount": {}
  },
  "body": {
    "tokenCount": 273,
    "lineCount": 36,
    "emptyLineCount": 0,
  }
}
```
<table>
<thead>
<tr>
<th>page</th>
<th>section</th>
<th>token</th>
<th>pos</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>27</td>
<td>body</td>
<td>those</td>
<td>DT</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>within</td>
<td>IN</td>
<td>1</td>
</tr>
<tr>
<td>28</td>
<td>body</td>
<td>a</td>
<td>DT</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>be</td>
<td>VB</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>deserted</td>
<td>VBN</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>faintly</td>
<td>RB</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>important</td>
<td>JJ</td>
<td>1</td>
</tr>
<tr>
<td>section</td>
<td>token</td>
<td>count</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
<td>-------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>body</td>
<td>,</td>
<td>3258</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&quot;</td>
<td></td>
<td>1670</td>
<td></td>
<td></td>
</tr>
<tr>
<td>the</td>
<td></td>
<td>1565</td>
<td></td>
<td></td>
</tr>
<tr>
<td>.</td>
<td></td>
<td>1532</td>
<td></td>
<td></td>
</tr>
<tr>
<td>and</td>
<td></td>
<td>1252</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
HTRC Extracted Features - Plotting

# HTRC Extracted Features - Grouping

<table>
<thead>
<tr>
<th>page</th>
<th>section</th>
<th>token</th>
<th>pos</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>header</td>
<td>CHAPTER</td>
<td>NNP</td>
<td>1</td>
</tr>
<tr>
<td>35</td>
<td>header</td>
<td>CHAPTER</td>
<td>NNP</td>
<td>1</td>
</tr>
<tr>
<td>56</td>
<td>header</td>
<td>CHAPTER</td>
<td>NNP</td>
<td>1</td>
</tr>
<tr>
<td>73</td>
<td>header</td>
<td>CHAPTER</td>
<td>NNP</td>
<td>1</td>
</tr>
<tr>
<td>91</td>
<td>header</td>
<td>CHAPTER</td>
<td>NNP</td>
<td>1</td>
</tr>
<tr>
<td>115</td>
<td>header</td>
<td>CHAPTER</td>
<td>NNP</td>
<td>1</td>
</tr>
<tr>
<td>141</td>
<td>header</td>
<td>CHAPTER</td>
<td>NNP</td>
<td>1</td>
</tr>
<tr>
<td>158</td>
<td>header</td>
<td>CHAPTER</td>
<td>NNP</td>
<td>1</td>
</tr>
<tr>
<td>174</td>
<td>header</td>
<td>CHAPTER</td>
<td>NNP</td>
<td>1</td>
</tr>
<tr>
<td>193</td>
<td>header</td>
<td>CHAPTER</td>
<td>NNP</td>
<td>1</td>
</tr>
</tbody>
</table>
HTRC Extracted Features - Plotting

Split-apply-combine

SPLIT based on a common criteria

APPLY a change to each group

COMBINE the newly processed groups
Split-apply-combine

**Movie** | **Director** | **Gross**
---|---|---
Annie Hall | W. Allen | $38m
Transformers | M. Bay | $319m
Manhattan | W. Allen | $39m
The Rock | M. Bay | $134m

**Director** | **Avg. Gross**
---|---
W. Allen | $38.5m
M. Bay | $226.5m

**SPLIT** by director

**APPLY** averaging

**COMBINE** to new data
Text analysis use cases

• Distant reading
• Word similarity
• Topic models
• Plot analysis
• Emotion analysis
• Visual structure

Text analysis use cases

- Distant reading
- Word similarity
- Topic models
- Plot analysis
- Emotion analysis
- Visual structure

https://wiki.htrc.illinois.edu/display/COM/Extracted+Features+in+the+Wild
Text analysis use cases

- Distant reading
- Word similarity
- Topic models
- Plot analysis
- Emotion analysis
- Visual structure

Text analysis use cases

- Distant reading
- Word similarity
- Topic models
- Plot analysis
- Emotion analysis
- Visual structure

Within-book topic modelling, Peter Organisciak

Topics in The Scarlet Letter

Topics pertaining to: 'customhouse office surveyor official general salem public years pue port make sketch long inspect...'

Topics pertaining to: 'table actual floor corner wall room library scene sombre chairs traffic literature watched path a...'

Topics pertaining to: 'town street soil wooden front portion scenes present moral edifice entered dust claim houses plac...

Topics pertaining to: 'nature life age character mind change state person men kind indian order day long position effect...

Topics pertaining to: 'marketplace prison people embroidered crowd public needle rude beauty prison door sentence rigid i...'

Interpretation of mediums

5 copies of Origin of Species in HTC corpora

Two trends:
• 1929 and after are on average about one quarter of an inch taller and one eighth of an inch wider than books published before

• the font size (as determined by the median line height) also appears to be increasing (by a quarter of a point)

Figure 1: Five versions of Darwin’s Origin of Species, illustrating different design choices in page layout.

Insight driven investigation

Distant Horizons, Ted Underwood (2019)

Frequency of colour terms in random fiction increases post 1800 – why?

• Does it represent the decline in 3\textsuperscript{rd} person narration? A shift to description?
• We need to look for correlations

\begin{figure}
\centering
\includegraphics[width=\textwidth]{figure1.png}
\caption{Frequency of color terms in a random sample of fiction, 1700-1922. Points represent 347 individual volumes. After 1800, volumes reviewed in a sample of prestigious magazines are distinguished as black triangles.}
\end{figure}

Distant Horizons, Ted Underwood (2019), p10
Frames of context

Standord 2012 Heuser & Le-Khac:

- Rising frequency of concrete adjectives in 19th century novels (colours, names, body parts, action verbs – used for physical descriptions)

Concluded fictional discourse was transitioning “from telling to showing”

Practical – Finding Trends

bookworm: HathiTrust
Search for trends in millions of volumes at http://hathitrust.org

Find 3-5 queries/query words which relate to your interests:
- Use the filters (e.g categories/languages)
- Refer to the source texts

Discuss the trends you find.

https://bookworm.htrc.illinois.edu/develop/

https://books.google.com/ngrams
Examples

Literary trends
- Red
- White
  (Narrow class filter: fiction/biography)

Meaning change
- Hysteria
- Gay
- legacy/heritage

Parts of speech
- Book_VERB
- Book_NOUN

Geopolitical
- Persia
- Iran
  - ایران
- Beijing
- China
Pick 3 topics in each group:
- Use the GUI to investigate
- Why does the GUI help?
- Refer to the source texts

Discuss your findings.

French literature cluster?
Label = French Women or French Gender?
Contains = Fiction with a lot of gender descriptions?
A lot of polite conversation with women characters?
Computer-assisted tools and Hermeneutics

Descartes, *Discourse on Method*: “[T]hings made up of different elements and produced by the hands of several master craftsmen are often less perfect than those on which only one person has worked.”

Individual versus communal-participatory inquiry; dialogic relationship between human and human and machine, criticism and code; augmentation of criticism.

Teamwork changes the single-interpreter model that Descartes describes: “analytical tools are instantiations of interpretive methods that can be woven closely into other hermeneutical things” (Rockwell and Sinclair).