HTRC Use Cases
HathiTrust Corpus Usage Patterns

HathiTrust Corpus

The image cannot be displayed. Your computer may not support the format of the image.
HathiTrust Corpus Usage Patterns (cont’d)
Word Counts from HTRC Sample*

- **Top 10 words**
  - the (1,092,274,158)
  - of (729,347,125)
  - and (515,034,460)
  - to (429,304,807)
  - in (337,513,888)
  - a (315,487,516)
  - that (167,847,940)
  - is (163,694,582)
  - was (138,907,857)
  - I (123,743,522)

- **Bottom 10 tokens**

<table>
<thead>
<tr>
<th>Occurrence</th>
<th>Num of unique tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>109</td>
</tr>
<tr>
<td>2</td>
<td>217</td>
</tr>
<tr>
<td>3</td>
<td>360</td>
</tr>
<tr>
<td>4</td>
<td>526</td>
</tr>
<tr>
<td>5</td>
<td>583</td>
</tr>
<tr>
<td>6</td>
<td>551</td>
</tr>
<tr>
<td>7</td>
<td>541</td>
</tr>
<tr>
<td>8</td>
<td>515</td>
</tr>
<tr>
<td>9</td>
<td>416</td>
</tr>
<tr>
<td>10</td>
<td>356</td>
</tr>
</tbody>
</table>

*Public Domain non-Google digitized HT materials, 250,000 volumes*
## OCR Corrections on HTRC Sample

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of N-grams</td>
<td>20,173,974,251</td>
</tr>
<tr>
<td>Total number of N-grams (minus numbers only and other easy-to-spot noises)</td>
<td>19,282,108,416</td>
</tr>
<tr>
<td>Number of corrections made</td>
<td>131,571,046</td>
</tr>
<tr>
<td>Number of valid correction rules</td>
<td>99,455</td>
</tr>
</tbody>
</table>
HTRC Online Tools for Simple Analysis

Author and Title Search

Simple examples of web access to the Hathitrust Research Center collections

ARTICLE | JUNE 9, 2011 - 8:56AM

This tool provides several examples of text analysis and visualization which are created by searching bibliographic information stored in the Hathitrust Research Center collections and calling SEASR web services for analytical processing.

Type the name of an author, or choose "Titles" and type a few letters. After a few seconds, scroll down to see results.

Twain, Mark

Suggest: ○ Authors ○ Titles
Meandre Flow: ○ Tag Cloud Viewer ○ OpenNLP Protovis ○ Readability Test

About this volume:
Author(s): Twain, Mark;
Title: Life on the Mississippi
Catalog URL: http://catalog.hathitrust.org/Record/005994569
Tag Cloud Viewer
Topic Modeling

• Uses MALLET Topic Modeling to cluster
• Top 8 topics showing at most 200 keywords for that topic
Concept Mapping

- Sentiment Analysis
  
  - six core emotions (Love, Joy, Surprise, Anger, Sadness, Fear)
Visualization for Extracted Entities
Mayor Rex Luthor announced today the establishment of a new research facility in Alderwood. It will be known as Boynton Laboratory.
Metadata Enrichment

• Gender
• Genre
• Structural
  – Chapters
  – Front matter
  – Indexes
  – Bibliographies
• Part-of-Speech (POS) tagging

Example source: http://www.stanford.edu/~mjockers/cgi-bin/drupal/node/17