Mapping Mutable Genres in Structurally Complex Volumes

Ted Underwood (@Ted_Underwood),
Michael L. Black, Univ. of Illinois,
Dept. of English

Loretta Auvil, Boris Capitanu
Illinois Informatics Institute

The best version is on arXiv; parts of the argument are blogged at The Stone and the Shell.
Overview

• 469,200 nonserial vols from HathiTrust, 1700–1899.
• Goal: map genre at the page level or below, algorithmically.
• This paper mainly demonstrates methods.
Classify literary genres algorithmically?

- Allison, Heuser, Jockers, Moretti, and Witmore, 2011. Clustering, PCA.
- Yu, 2008. Naïve Bayes, SVM.
- In this paper: naïve Bayes and regularized logistic regression. A variety of feature-selection strategies.
Naive Bayes at the volume level, calibrated with logistic regression.

(432 documents held out from corpus of 1356 19c vols.)

<table>
<thead>
<tr>
<th>predicted ➔ actual ➣</th>
<th>prose nonfiction</th>
<th>prose fiction</th>
<th>verse and drama</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>prose nonfiction</td>
<td>118</td>
<td>5</td>
<td>0</td>
<td>0.959</td>
</tr>
<tr>
<td>prose fiction</td>
<td>1</td>
<td>143</td>
<td>1</td>
<td>0.986</td>
</tr>
<tr>
<td>verse and drama</td>
<td>0</td>
<td>7</td>
<td>157</td>
<td>0.957</td>
</tr>
<tr>
<td>Precision</td>
<td>0.992</td>
<td>0.923</td>
<td>0.994</td>
<td></td>
</tr>
</tbody>
</table>
“Genre” can mean a wide range of different things.

form-like; easier to define
mode-like; harder to define

verse  table of contents  Robinsonade  sensation novel  elegy  gothic
Some genres can be treated as mutually exclusive and used to divide volumes.

prose fiction, nonfiction, drama, nondramatic verse, title page, table of contents, bookplate, etc.
Once divided, volume-parts have to be described by multiple overlapping ontologies.
Nondramatic poetry: .99
Narrative poem: .85
Progress poem: .65
Footnoted poem: .93

Multiple overlapping ontologies.
Mutable genres

Probability of fiction predicted by 18c classifier vs Probability of fiction predicted by 19c classifier
Page-level classifiers + hidden Markov model of volume structure
Prevalence of first-person narration in two different collections

32,209 algorithmically selected volumes.

774 hand-selected volumes.
Gender and point of view

![Bar chart showing the proportion of works in first person by gender from 1700-1749 to 1850-1899. The chart shows a significant increase in works by men from 1750-1799 compared to other time periods.](image)