The task of cleaning and enriching large collections

what aspects can we share?
Contributing to this work

UIUC English:
    Ted Underwood
    Jordan Sellers
    Mike Black

UIUC Library:
    Harriett Green

I3:
    Loretta Auvil
    Boris Capitanu

Andrew W. Mellon Foundation
“Enrich” as well as “clean.”
Yearly values of a ratio between two wordlists in three different genres. 4,275 volumes. 1700-1899.
Topic 88: like light wild over

Letitia Landon,
The Improvisatrice,
The Troubadour,
and The Vow of the
Peacock

Felicia Hemans,
Records of
Woman and
Songs of the
Affections

P. B. Shelley,
Alastor; or,
The spirit of solitude

short stories by
Nathaniel
Hawthorne

like
light
wild
over
whose
through
dark
spirit
bright
deep
flowers
brow
yet
dream
dim

Blue/fic, purple/poe, green/drama, black/bio, brown/nonfic, triangle/letters or orations.
“representative?”
analyzing the data

cleaning the data
“clean” is relative
different projects will strike a different balance between precision and recall makes it tricky to share resources
Cleaning the data

1. Clean up the OCR / assess error.

2. Identify parts of a volume (e.g., articles in a serial, poetry/prose).

3. Remove library bookplates and running headers — after using them for (3).
The Duel of the Period in France.

duelling, I remember calling at the office of a great Parisian newspaper with a friend who wished to have rectified a statement published in it concerning him. When our business was made known we were ushered into a handsomely furnished room on the first floor. Seated at desks, without a trace of pens, ink, or paper, or of anything in a literary way except some new novels, together with a few packages of cigarettes, were two gentlemen, whose appearance made a considerable impression on me. They were faultlessly dressed in deep black (the duellist's de Dion, the greatest living authority in France on duelling, who has been "out" scores of times, both as principal and second, and whose undisputed loyalty and firmness have made it possible for him to prevent many duels that seemed inevitable. The marquis killed Captain Mayer in a duel with swords at the Ile de la Grande Jatte a few years ago, and in a pistol duel with a deputy, M. Dreyfus, wounded him in the arm. When on his American ranch two cowboys tried to "jump" some of his cattle, he and one of his herdsmen fought them off with "Win-
period-specific lexica incl. foreign lang.

collection-level observations: *proper nouns, words that appear mainly in dirty docs*

context of an individual doc:

It is purely a mortal fin to ...

Correction rules
Cleaning/enriching the metadata

1. “18??”

2. Discard duplicate volumes / select early editions?

3. Add metadata that you need for interpretive purposes, like
   — gender (see Ben Schmidt’s technique),
   — genre.
first stab at genre – naive Bayes

<table>
<thead>
<tr>
<th></th>
<th>really drama</th>
<th>really fiction</th>
<th>really nonfiction</th>
<th>really poetry</th>
</tr>
</thead>
<tbody>
<tr>
<td>as “drama”</td>
<td>85.4%</td>
<td>0%</td>
<td>0%</td>
<td>6.1%</td>
</tr>
<tr>
<td>as “fiction”</td>
<td>0%</td>
<td>95.1%</td>
<td>6.0%</td>
<td>0.8%</td>
</tr>
<tr>
<td>as “nonfiction”</td>
<td>0%</td>
<td>2.8%</td>
<td>94.0%</td>
<td>3.0%</td>
</tr>
<tr>
<td>as “poetry”</td>
<td>14.6%</td>
<td>2.1%</td>
<td>0%</td>
<td>90.1%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>100%</strong></td>
<td><strong>100%</strong></td>
<td><strong>100%</strong></td>
<td><strong>100%</strong></td>
</tr>
</tbody>
</table>
Things we could share

period lexicons / variant spellings
gazetteers of proper nouns
OCR correction rules for a period
document segmentation and/or cleaned
and segmented text
ferberization
cleaned / enriched metadata
code to do all of the above
get clues from metadata
break vols into parts
ensemble / boosting
active learning
active learning: documents classified as “fiction,” plotted by confidence in classification (y axis). Red points are misclassified.