Mapping Genre at the Page Level

In English-Language Volumes from HathiTrust, 1700-1899

Ted Underwood, Shawn Ballard, Michael L. Black, and Boris Capitanu, University of Illinois, Urbana-Champaign

So, you want to do distant reading. How will you actually find thousands of texts in a given genre?

Genre metadata for digital volumes is spotty; even with broad categories like "poetry" and "drama," we're able to deduce genre from volume-level metadata only about a third of the time. Moreover, volumes are divided internally. A volume of poetry may include plays, begin with a life of the author, and end with twenty pages of publisher's ads, followed by a "date due" slip.

If we want to distant-read public digital collections, we need to develop a map that identifies (at a minimum) the specific pages we expect to be fiction, or poetry, or nonfiction prose, or paratext. In fact, we'll need to go further than that; we'll want to map narrower categories like "the epistolary novel," and divide genres below the page level. In this poster we're only demonstrating the first phase of this process.

"Wait. Aren't genres blurry social categories, defined differently by different readers?"

They are.[1] That's another reason why our current mapping strategy is broken. Right now we expect library catalogers to a) agree on a single set of categories and b) decide whether each volume does or does not deserve a given genre tag. But all genre categories have blurry edges. Even the broad divide between "fiction" and "nonfiction" is troubled by almanacs of marvels, lightly fictionalized biographies, and so on. A probabilistic approach to classification can acknowledge these regions of dissensus, and even identify texts that are likely to trouble a given boundary. Moreover, algorithmic mapping is fast enough that we can map a large collection iteratively, trying out many different ontologies. That would be hard with crowdsourcing.

Terrible confusion matrix based on received metadata

# of words	Drama (predicted)	Fiction	Nonfiction prose	Poetry	Paratext	recall
Drama (actual)	2,320,961	79,141	4,690,533	7,852	3,292	32.7%
Fiction	0	1,953,638	3,221,738	0	5,322	37.7%
Nonfiction prose	97,201	26,318	14,534,285	147,780	19,748	98.0%
Poetry	1,650	100,852	1,345,420	678,399	2,940	31.9%
Paratext	9,271	62,224	472,517	36,007	46,745	7.5%
precision	95.5%	87.9%	59.9%	78.0%	59.9%	Microavg F1: 65.4%

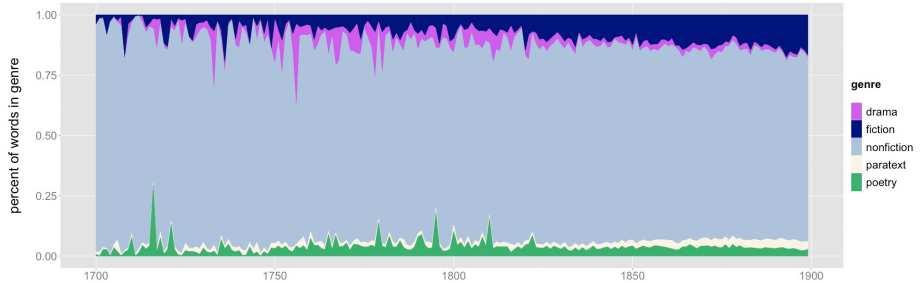

Fig 1. The proportion of the collection devoted to different genres varies over time. 394,827 English-language nonserial volumes from HathiTrust. Although we mapped 469,000 volumes, this visualization is based only on volumes with dates that were easy to parse.

Methods

Supervised learning requires an initial source of training data. We tagged 324 volumes manually, at the page level, with detailed genre information. Although our goal, at the moment, is to map pages onto the five broad supercategories plotted above, we do that by training classifiers for a larger number of specific subclasses — for instance we look for "front matter," "back matter," or "advertisements," but count all three as "paratext."

- The learning algorithm: regularized logistic regression (Weka).
- Features: 654 words or word groups, but also, for instance, information about line length and the first characters of lines.

Classifying pages as independent texts, our predictions were 87% accurate (tenfold cross-validated). To improve that, we used learning strategies custom-designed for the problem:

- A hidden Markov model trained on page *sequence*.[2]
- Taking library metadata in effect as hierarchical priors, we supplemented our main model with specialized models trained on subsets of the collection. A paper from Google proved useful.[3]

Those strategies brought accuracy up to 94.5%. But what does "accuracy" mean here? Since human readers disagree about genre, what's an appropriate benchmark for comparison?

Evaluating results

Most of the volumes in our training set were tagged by multiple human readers; we reached a provisional consensus about genre both by voting and by preferring experienced judgment. We used this provisional consensus as "ground truth" for the experiment (the confusion matrices at lower left and upper right are based on it). But we can also compare this standard to the judgments of individual readers in order to expose the human disagreement that created it.

For instance, we found that individual human readers matched the consensus genre for only 94.4% of words in the collection. So algorithmic classification matched the human consensus almost exactly as often as *individual* human readers did. In other words, our algorithmic map should be about as reliable as a system where each volume gets skimmed once by an English major with brief training for the task.

Our model is presumably not as reliable as a scheme where each volume would get multiple human readings. But a system like that would be hard to scale even to thousands of volumes, whereas we expect to expand this solution to cover millions of twentieth-century volumes, using features extracted non-consumptively by the HathiTrust Research Center.

Mapping ambiguity

One of the advantages of a probabilistic approach is that uncertainty is built into the method. The logistic models we train report a real-valued probability between 0 and 1 for each genre on each page. We use that information (along with metadata) to train a meta-model that characterizes our overall confidence about predictions for each volume. This model of confidence correlates strongly with actual out-of-sample accuracy ($r > 0.40$). We have found that sorting the collection by algorithmically-predicted confidence is in practice a good way to identify puzzling boundary cases.

Next steps

- Expand this basic page map to 1923, and beyond; share results.
- Begin to divide broad categories into subgenres. This will produce arguments of a more provisional kind, no longer resting on ~94% human consensus.
- Divisions below the page level. Serials.

Confusion matrix for the page-level model we trained

# of words	Drama (predicted)	Fiction	Nonfiction prose	Poetry	Paratext	recall
Drama (actual)	6,836,264	42,528	181,247	40,278	1463	96.3%
Fiction	5,372	4,950,165	217,386	6673	1103	95.6%
Nonfiction prose	304,590	296,704	14,117,791	87,241	19,006	95.2%
Poetry	143,049	15,150	54,867	1,915,083	1112	89.9%
Paratext	21,866	4490	144,762	61,071	394,575	63.0%
precision	93.5%	93.2%	95.9%	90.7%	94.6%	Microavg F1: 94.5%

Acknowledgments

This project relied on assistance from HathiTrust Research Center, as well as Jonathan Cheng, Tim Cole, Stephen Downie, Colleen Fallaw, Harriett Green, Nicole Moore, Clara Mount, Lea Potter, and Jeremy York. The project is supported by an NEH Digital Humanities Start-Up Grant and an ACLS Digital Innovation Fellowship.

Find out more

On *The Stone and the Shell*, I've posted a more detailed discussion of methods, and a link to an interactive visualization.

Key citations

1. P. DiMaggio. "Classification in Art." *American Sociological Review* 52 (1987): 440-455.
2. T. Underwood, et. al. "Mapping Mutable Genres." *IEEE BigData '13.*
3. D. Sculley, et. al. "Detecting Adversarial Advertisements in the Wild." *KDD 2011.*

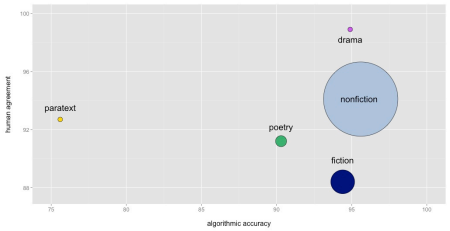

Fig 2. Although individual human readers had nearly the same overall accuracy as an algorithmic model, and were commonly ($r = 0.10$) confused by the same specific volumes, they did have different strengths. Notably, human readers struggled to distinguish fiction from nonfiction prose, while machine learning struggled to distinguish text from paratext. Circles here correspond to the sizes of genres in the whole collection.