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 distance-read 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 paranovel. In fact, we’ll need to go farther 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. That’s another reason why our current mapping strategy is broken. Right now we expect library catalogers to agree on a single set of categories and by 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.

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 superclass categories 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: 645 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.
• 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.

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 94.8% 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 skinned 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-consumpitively 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 useful 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.

Terrible confusion matrix based on received metadata

| Type of words | Drama (predicted) | Fiction | Nonfiction prose | Poetry | Paratext | Exact
|---------------|------------------|--------|------------------|--------|----------|------
| Drama (exact) | 2,232,961 | 79,141 | 4,666,533 | 7,852 | 2,392 | 32.7%
| Fiction | 0 | 1,830,658 | 3,251,738 | 0 | 5,322 | 37.7%
| Nonfiction prose | 97,291 | 26,318 | 14,034,285 | 147,769 | 19,748 | 85.9%
| Poetry | 1,650 | 100,052 | 1,045,620 | 676,396 | 2,940 | 31.9%
| Paratext | 9,271 | 62,324 | 472,517 | 36,027 | 46,745 | 7.5%
| precision | 95.5% | 87.9% | 99.6% | 78.0% | 58.9% | 43.4%

Confusion matrix for the page-level model we trained

| Type of words | Drama (predicted) | Fiction | Nonfiction prose | Poetry | Paratext | recall
|---------------|------------------|--------|------------------|--------|----------|------
| Drama (actual) | 6,363,354 | 42,328 | 181,247 | 45,378 | 1,963 | 96.3%
| Fiction | 5,372 | 4,660,165 | 217,386 | 6873 | 1103 | 95.6%
| Nonfiction prose | 306,356 | 268,704 | 14,117,791 | 87,241 | 19,906 | 95.2%
| Poetry | 140,049 | 15,153 | 54,937 | 1,915,083 | 1112 | 89.9%
| Paratext | 21,666 | 449 | 144,702 | 61,071 | 334,570 | 62.0%
| precision | 93.5% | 99.2% | 95.9% | 90.7% | 94.8% | 97.3%

Acknowledgments

This project relied on assistance from HathiTrust Research Center, as well as Jonathan Cheng, Tim Cole, Stephen Downie, Colleen Fallaw, Harriet 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


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 94.8% 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 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.