PREDICTING DATE OF ORIGIN FOR LONG FORM TEXT

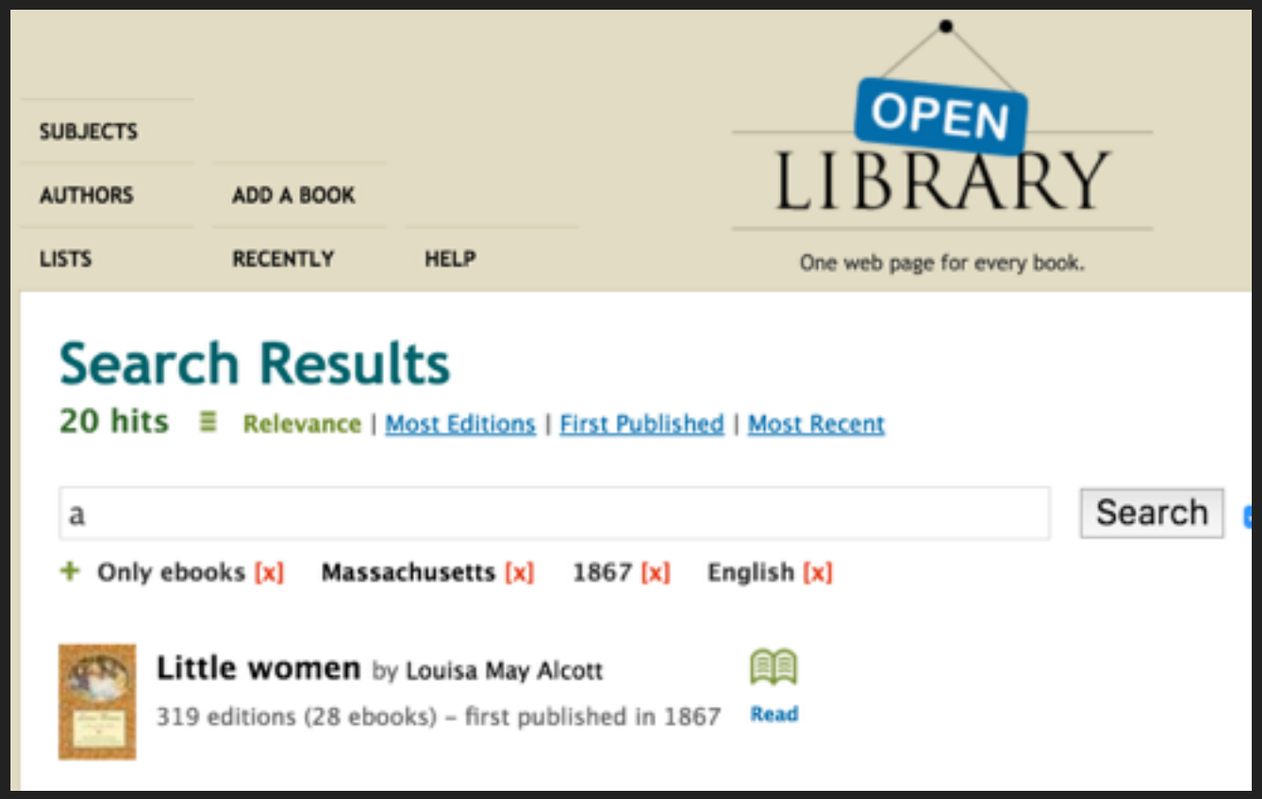
PROJECT FLETCHER

GOAL: ESTIMATE DATE OF ORIGIN OF ENGLISH LANGUAGE TEXT

- Input block of text, output an estimate of when it's from
- Need *large* body of text, has to be labeled with time, curated for English
- Ideally would have genres and regions tagged for future work

DATA

- Scraped openlibrary.org… a lot
- >25,000 books
- Most between 1700-2000; everything from textbooks to Congressional proceedings.



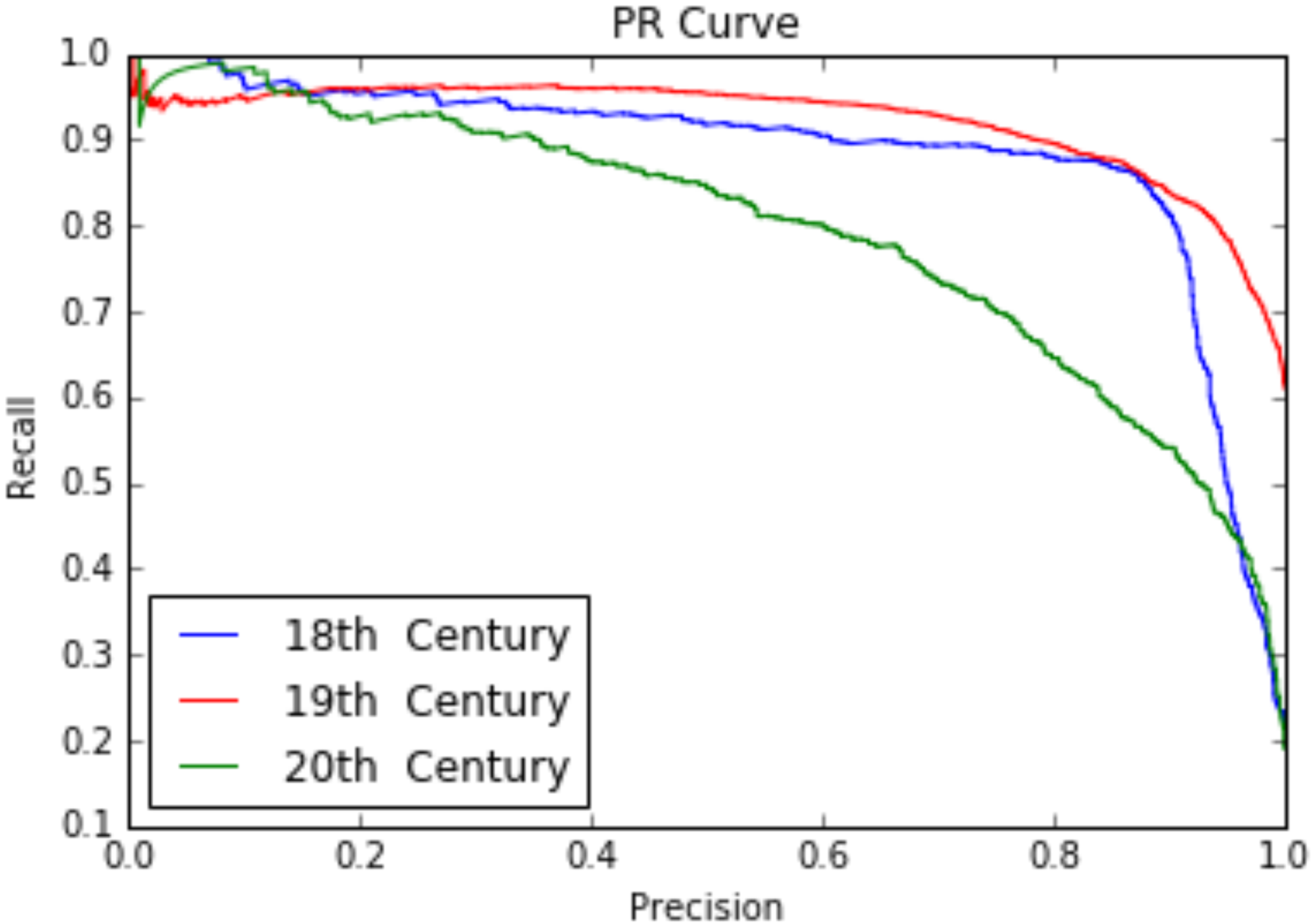

```
System information as of Fri Aug 26 11:48:14 UTC 2016

System load:  0.0                 Processes:           101
Usage of /:   63.3% of 29.39GB    Users logged in:     0
```

ANALYSIS

- Used subset (75 MB, representative) for computational and cleanliness reasons
- Goal was to classify tweets by by century
- Pre-1700 too sparse, so just 18th-20th

ANALYSIS

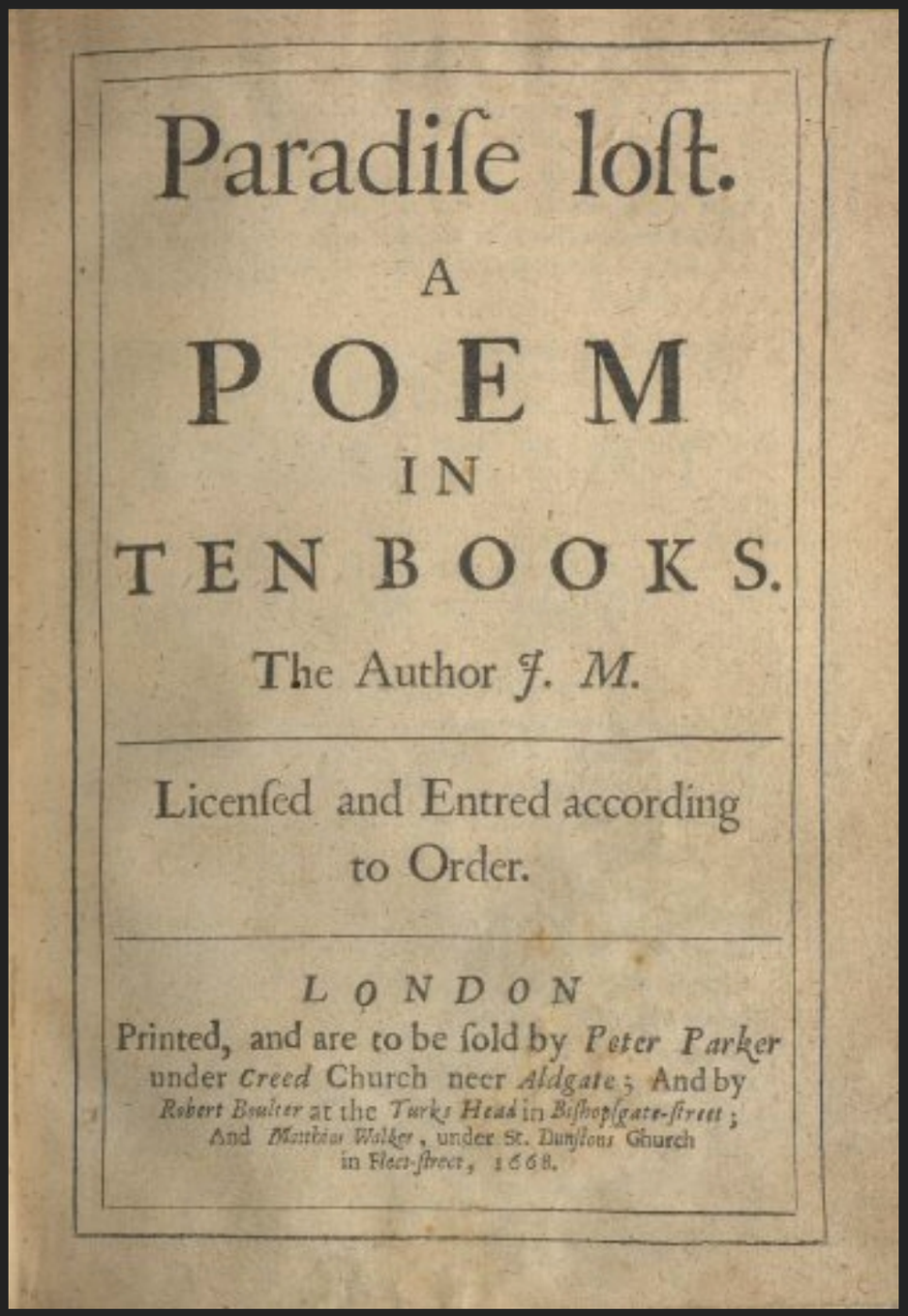
- Ran all sorts of pipeline combinations
- Three-class classification moderately accurate; breaking into binary classifiers did better
- Best results with pipeline TF-IDF, SVD, Logistic Regression
  - Scores: .94 for 18th century, .84 for 19th, .88 for 20th


PR Curve
Recall
Precision
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.2
0.4
0.6
0.8
1.0
18th Century
19th Century
20th Century

ANALYSIS

- Got help from (arguable) leakage from what looks like transcription issues

```
[['fo', 0.56949332406688014],
 ['moft', 0.36027059301172443],
 ['fuch', 0.35395030555736778],
 ['thofe', 0.34285897488338185],
 ['firft', 0.33597087168740863],
 ['fome', 0.30659836345205282],
 ['fame', 0.27789193210410701],
 ['feveral', 0.27094545833955558],
 ['fhall', 0.26773871467581922],
 ['himfelf', 0.26448368416610507]]
```



Paradiſe loſt.

A

POEM

IN

TEN BOOKS.

The Author *J. M.*

Licenſed and Entred according to Order.

LONDON

Printed, and are to be ſold by *Peter Parker* under *Creed* Church neer *Aldgate*; And by *Robert Boulter* at the *Turks Head* in *Biſhopſgate-ſtreet*; And *Matthias Walker*, under St. *Dunſtons* Church in *Fleet-ſtreet*, 1668.

FUTURE WORK

- More granular predictions
- Host on flask
- Include more data in model
  - Kept getting noticeably better
- Separate by topic and region
  - Options with both topic modeling and scraping