DEREK UPDEGRAFF

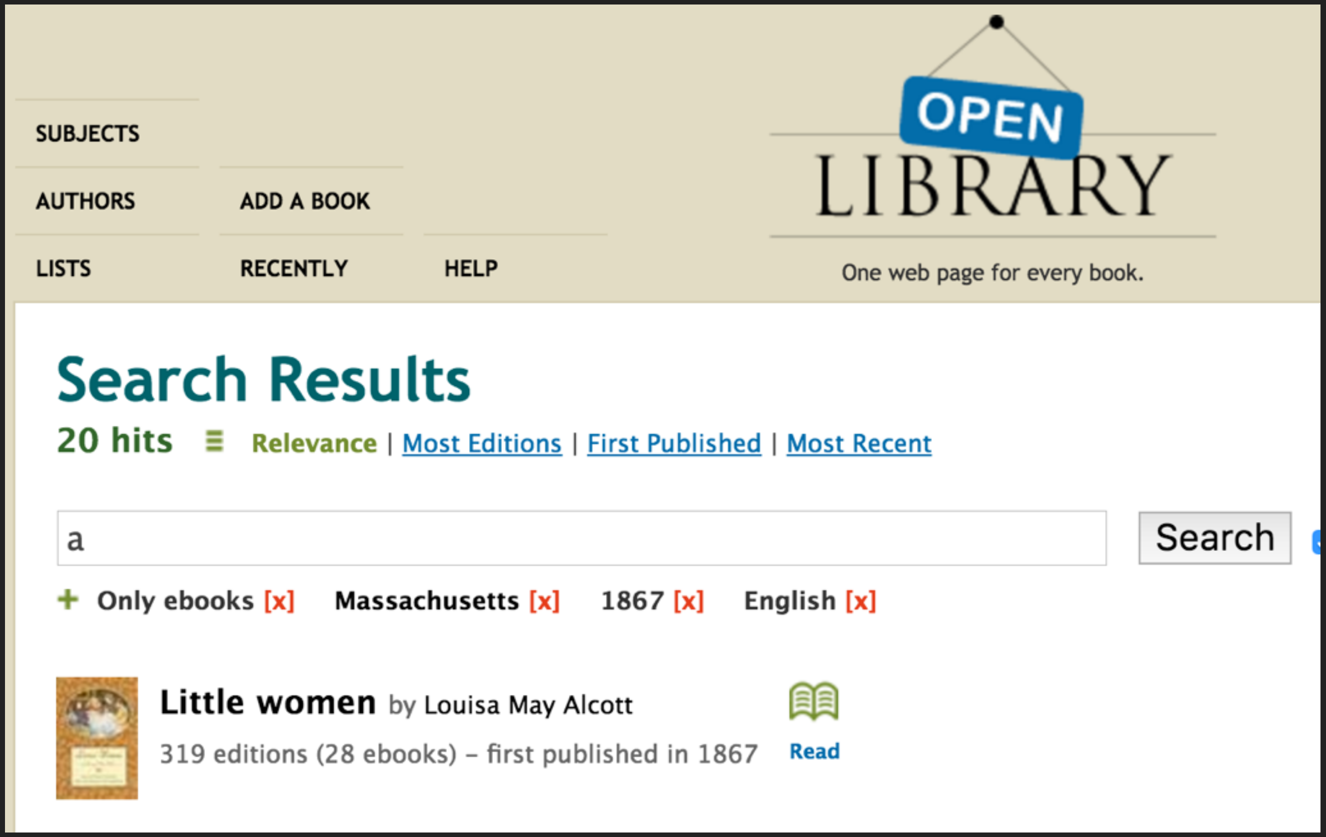
PREDICTING DATE OF ORIGIN FOR LONG FORM TEXT

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.



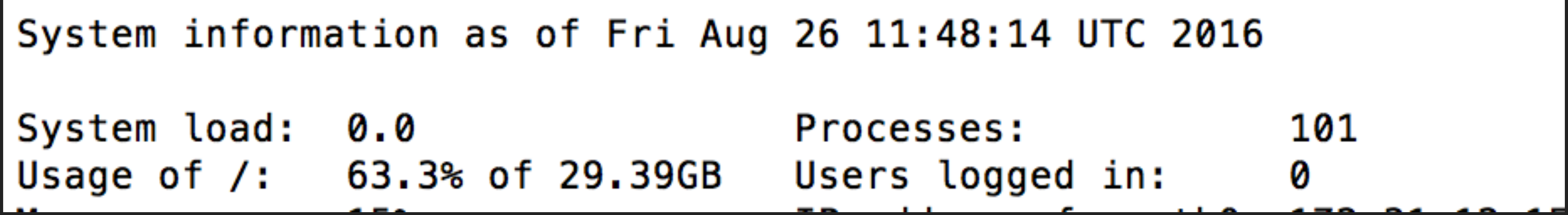
Search Results

20 hits ≡ Relevance | Most Editions | First Published | Most Recent

a

+ Only ebooks [x] Massachusetts [x] 1867 [x] English [x]

Little women by Louisa May Alcott
319 editions (28 ebooks) – first published in 1867 Read



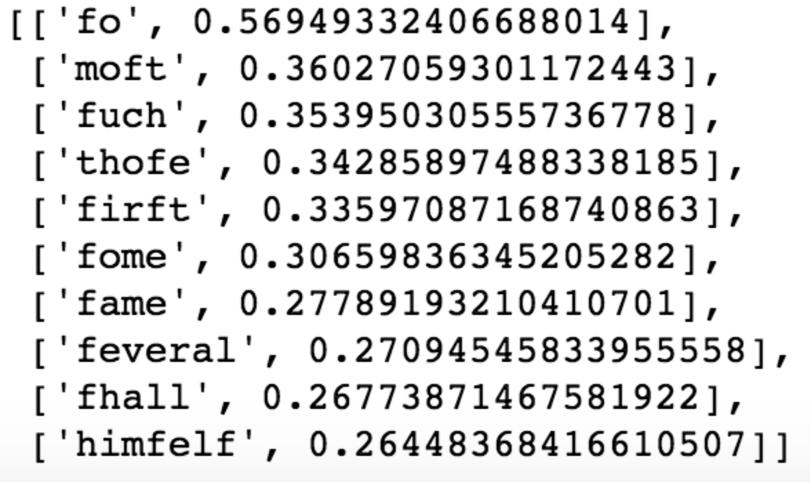
```
System information as of Fri Aug 26 11:48:14 UTC 2016

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

DATA CLEANING/SUBSETTING

- Used segment of each document for faster analysis
- Removed documents with f/s transcription error

Word-level feature Importances for earlier 18th century model



```
[['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]]
```

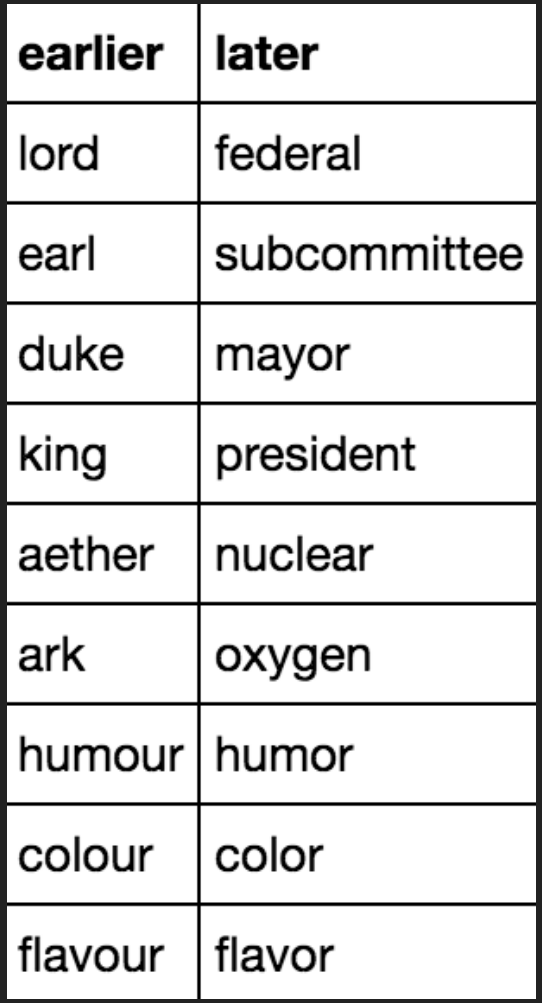
- Removed documents with numbered years
- Balanced Classes to avoid bias and speed up analysis
- Focused from period 1800-1979 because had sufficient documents
- Final set had 257 documents from each of 18 decades

MODEL/RESULTS

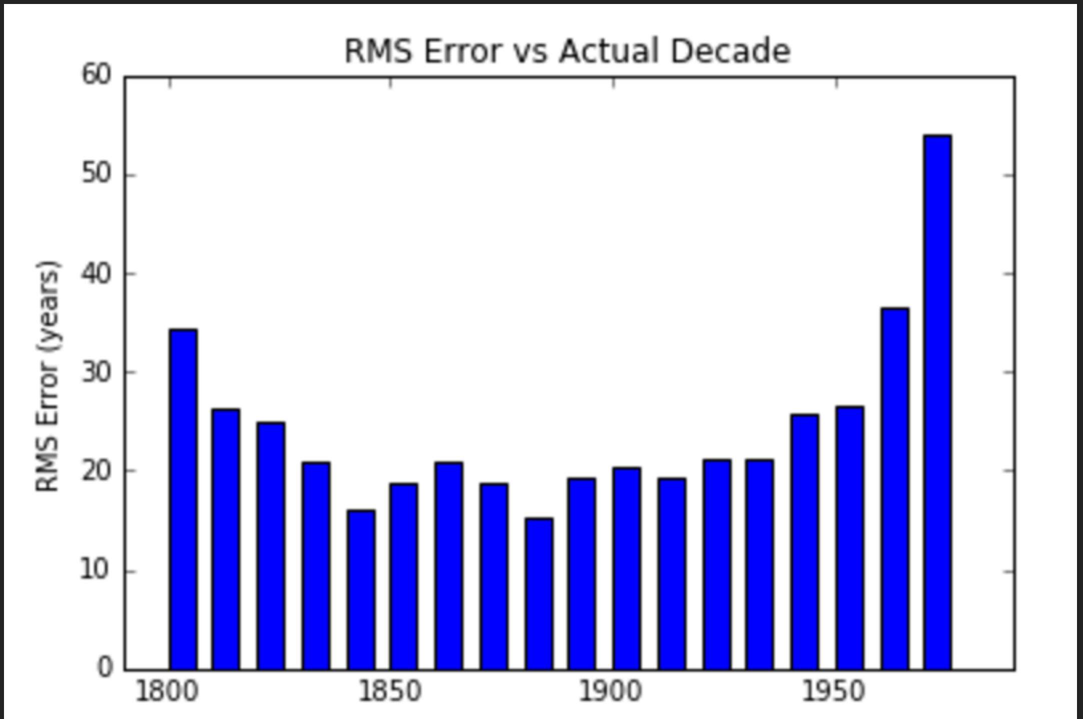
- Ran numerous models
  - 5-fold cross validation on train set to select model type and NLP/model parameters
  - Most effective approach linear regression, unigrams bigrams, tfidf
  - R2: 0.75, exact decade 20% of the time

MODEL/RESULTS

- 5-fold cross validation on train set to select model type and NLP/model parameters
- Most effective approach linear regression, unigrams bigrams, tfidf
- R2: 0.75, exact decade 20% of the time
- Most accurate in middle period
- Strong correlations with political, science words; US/UK English



earlier	later
lord	federal
earl	subcommittee
duke	mayor
king	president
aether	nuclear
ark	oxygen
humour	humor
colour	color
flavour	flavor



FUTURE WORK

- Host on flask
- Include more data in model
  - Vary document lengths
- Separate by topic and region
  - Options with both topic modeling and scraping