On the Origin of English: Classifying English Genres & Characterizing Language Evolution Over History

Andy Kim, David Whisler, Tim Gianitsos

Stanford
CS 221

Motivation

- Classifying large corpora of documents into coherent groups is an important application of natural language processing
- For instance, finding distinct characteristics of various forms of poetry dates to classical Greece and remains an active area of humanistic research today
- However, classification of texts remains understudied for Modern and Early-Modern English

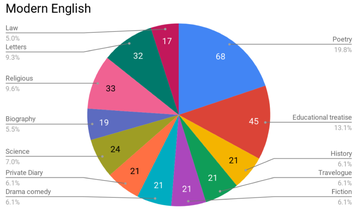
Problem

- Experiment with multiclass classifiers to categorize English texts from the same time period into 12 genres
- Use the best model and apply it to classify English texts from a different time period
- Use these results to investigate how the best features for classifying text change depending on the time period, and thus infer how English itself has changed over time

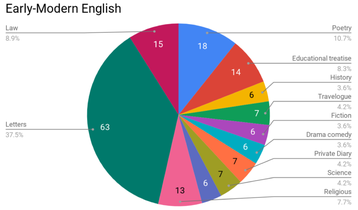
Data

- Modern English: 343 labeled texts from 1707-1914
- Early-Modern English: 168 labeled texts from 1501-1712
- Train-test stratified split of 80-20

Modern English



Early-Modern English



Challenges

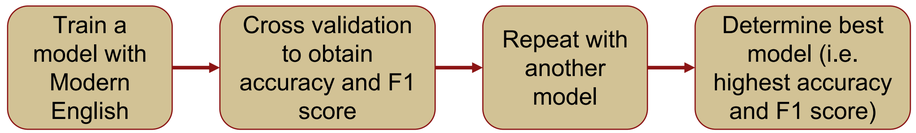
- Collecting sufficient amount of data for each genre and labeling each corpus correctly
- Extracting features from text efficiently (i.e. with good time/space complexity) since each “data point” is a an entire text

Approaches

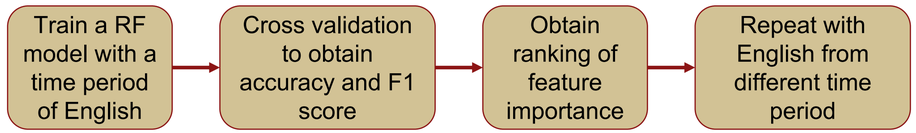
- Feature Set

Average sentence length	Ratio upper to lowercase chars	Ratio lowercase to total chars	Ratio punctuation to spaces	Ratio numeric to alphabetic chars
Average word length	Frequency of the most frequent stop word	Frequency of the most frequent start word	Frequency of most frequent starting letter of stop word	Frequency of most frequent starting letter of start word
Number single occurrence words	Number double occurrence words	Number of words with length 4	Number of words with length 3-5	Number of vowels

- Models: Support Vector, Naive Bayes, K-neighbors, Random Forest

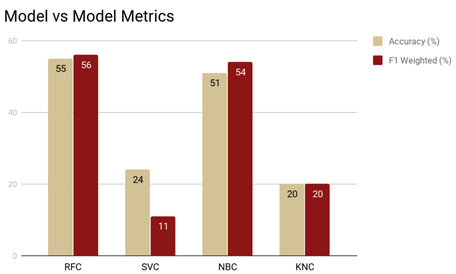


- After the above procedure, determined Random Forest was the best classifier

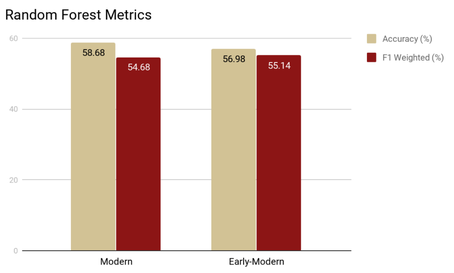


Results

Model vs Model Metrics



Random Forest Metrics



Features with most change in importance (Early-Modern English -> Modern English)	Rank Change (out of 15)	GINI Importance Change
Number of words with length 3-5	+9	+0.049978
Frequency of most frequent starting letter of start word	-8	-0.019756
Frequency of the most frequent start word	-8	-0.015838
Ratio of punctuation to spaces	-5	-0.021948
Frequency of the most frequent stop word	+3	+0.00537

Analysis

- **Classification Accuracy**
  - The trained Random Forest model accuracy was significantly better than the random chance accuracy of 8.33% with 12 categories (54.68% Modern, 55.14% Early-Modern), showing the chosen features are indeed good indications of genre
  - Similar classification accuracy across time periods demonstrates that the chosen features give similar indications of genre and rankings can be compared
- **Feature Rankings**
  - Biggest positive rank change (Early-Modern -> Modern) was the number of words of length 3-5 - one hypothesis is that Modern English incorporates many more long, scientific/domain specific words in different genres of texts (e.g. medical, scientific, educational genres)
  - Biggest negative rank change (Early-Modern -> Modern) was frequency of the most frequent starting word - one hypothesis is that Modern English has changed to allow for more diversity at the beginning of sentences as language has grown in size
- **Error**
  - GINI importance has some variance, feature importances may overlap probabilistically in some cases

Conclusion

- Our analysis shows that English can be classified into genres with relatively high accuracy based on syntactical features in both the Modern and Early-Modern time periods
- The syntactical features chosen changed in relative importance between the Modern and Early-Modern time periods, implying the broader syntax of the language has also changed due to linguistic and historical influences

Future Work

- Collaboration with UT Austin researchers to interpret feature rankings in a linguistic and historical context
- Analysis of grammatical features (ex: parts of speech, clause frequency) extending our work on syntactical features alone
- Extension of the work to Middle English and Old English, which are dramatically different from Modern English and are not even readable by Modern English speakers