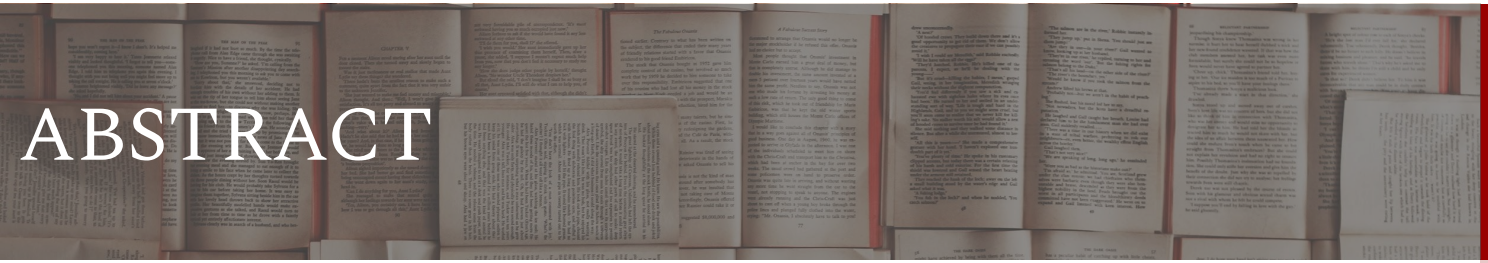

ABSTRACT

This project uses an early modern English corpus to explore authorship attribution (specifically, the Shakespearean authorship question) using Natural Language Processing, and to perform supervised classification determining whether an Elizabethan-era play was written by Shakespeare. The total dataset from Project Gutenberg is composed of 39 Shakespeare plays and 50 non-Shakespeare plays. To clean the text data I created functions to remove the editors notes, footnotes, and other legal text, and used Regex and NLTK for splitting text into sentences, tagging parts-of-speech tagging, and lemmatization to create a clean set of tokens.

A TF-IDF of the resulting tokens by unigrams, bigrams, and unigrams with dimensionality reduction were the primary data used with Scikit-learn, XGBoost and Keras to create classification models. Additional vectors were generated using GloVe weights. The most successful models were XGBoost and a LSTM Recurrent Neural Network which both achieved an accuracy of 100%, respectively, classifying all of our test data correctly. Additionally, I used Latent Dirichlet Allocation and Non-Negative Matrix Factorization to generate topics for each entire dataset, resulting in a low cohesion score with 7 topics of .298. The cohesion score did not improve for k topics between 3 and 30. In conclusion, we are able to identify 100% of Elizabethan-era plays correctly. Further work includes an expansion of our dataset, continued sourcing of pre-trained weights, and creation of a recommendation system.

NLP Techniques with Shakespeare's Plays

Kyle Aguilar
Full Time Online Data Science
June 2020
Flatiron School