CS 6140 Project Intermediate Report

Jim Morgan, Foad Namjoo, Ariel Oconitrillo

Introduction

“What's in a name? That which we call a rose by any other name would smell just as sweet.” These words, penned by William Shakespeare, convey how little Juliet’s family name means to Romeo, who would overlook their family feuding for romance’s sake. The quote can also apply to the use of language in Shakespeare’s plays. Foad, Jim, and Ariel have elected to dive deeper into the works of William Shakespeare and perform various data mining techniques on them.

Data Provenance

We plan to use a dataset comprised of lines of dialogue and stage directions in Shakespeare’s plays. The dataset can be found on Kaggle’s website here.

Data Structure

From an initial position, we plan on performing an n-gram analysis of Shakespeare’s various plays, looking for similarities in word usage and sentence structure between plays. Of additional interest are the influences of Shakespeare’s works on later plays and films. It would be interesting to perform a page-rank analysis of Shakespeare’s work with modern media like Star Wars. Additional potential targets for comparison include Moby Dick by Herman Melville and A Christmas Carol by Charles Dickens.

Problem Statement

The works of William Shakespeare are of profound importance to all English speakers and are part of the core curriculum in English-speaking countries. Shakespeare’s word usage, construction of the plot, and usage of performative techniques like the soliloquy to convey a character’s mental and emotional state were all novel devices at the time. However, some questions remain about Shakespeare’s authorship, such as whether they were penned by one man or several. It is hoped that through n-gram analysis, we can better qualify Shakespeare’s works and look for consistencies between plays.

Learning Objectives

While it is likely little room for improving the professor’s understanding of the data mining techniques used, it is hoped that there will be room for learning more of Shakespeare’s literary techniques and how his word choice and sentence structure influence modern language to the current day.

CS 6140 Project Intermediate Report

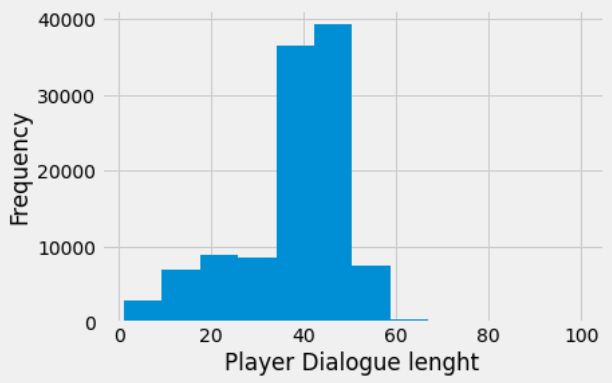
Jim Morgan, Foad Namjoo, Ariel Oconitrillo

Intermediate Report

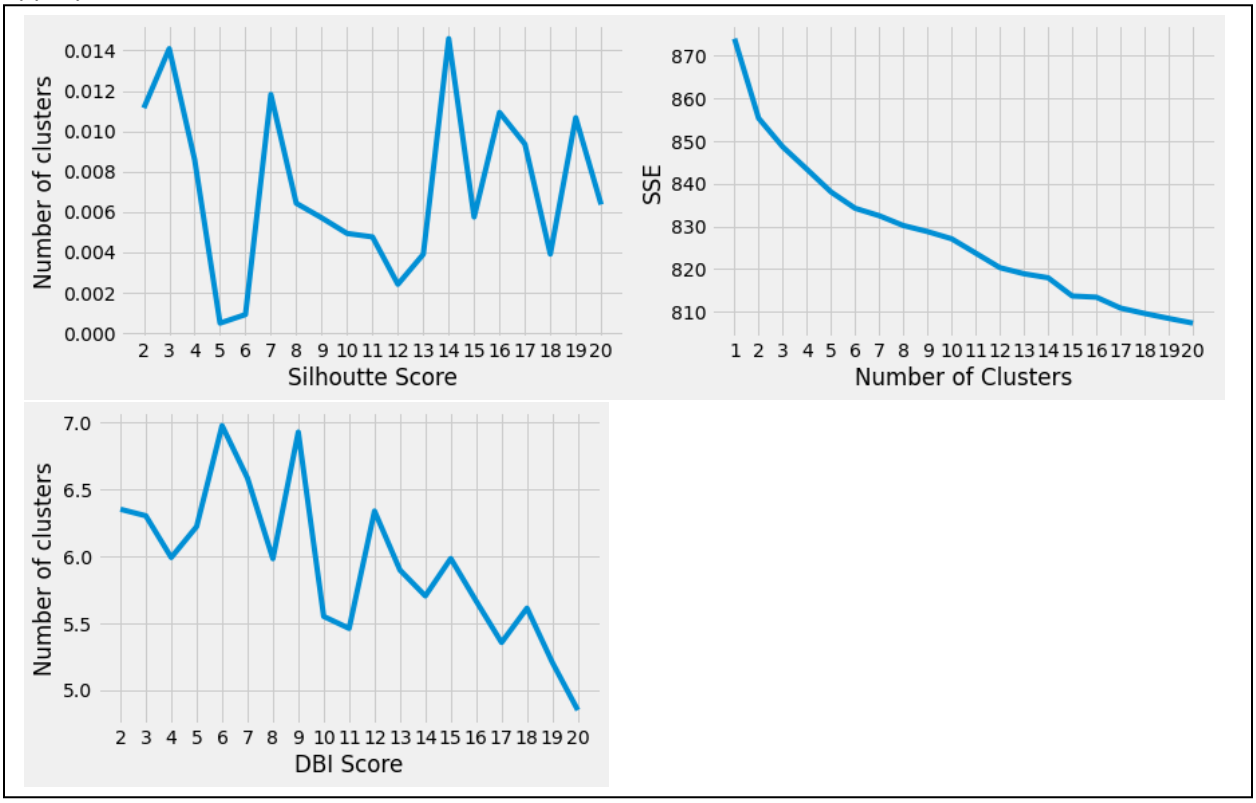
Our first objective was to modify the Shakespearean dataset to consolidate all character dialogue into individual lines.

Accounting for stop words reduces total word count to 22,309 unique words across 934 players, tokenized using TF-IDF.

Word associations are formed using k-means clustering, analyzed using Davies Bouldin Score (DBI), Silhouette Score (SS), and the sum of the squared Euclidean distances of each point to its closest centroid (SSE).



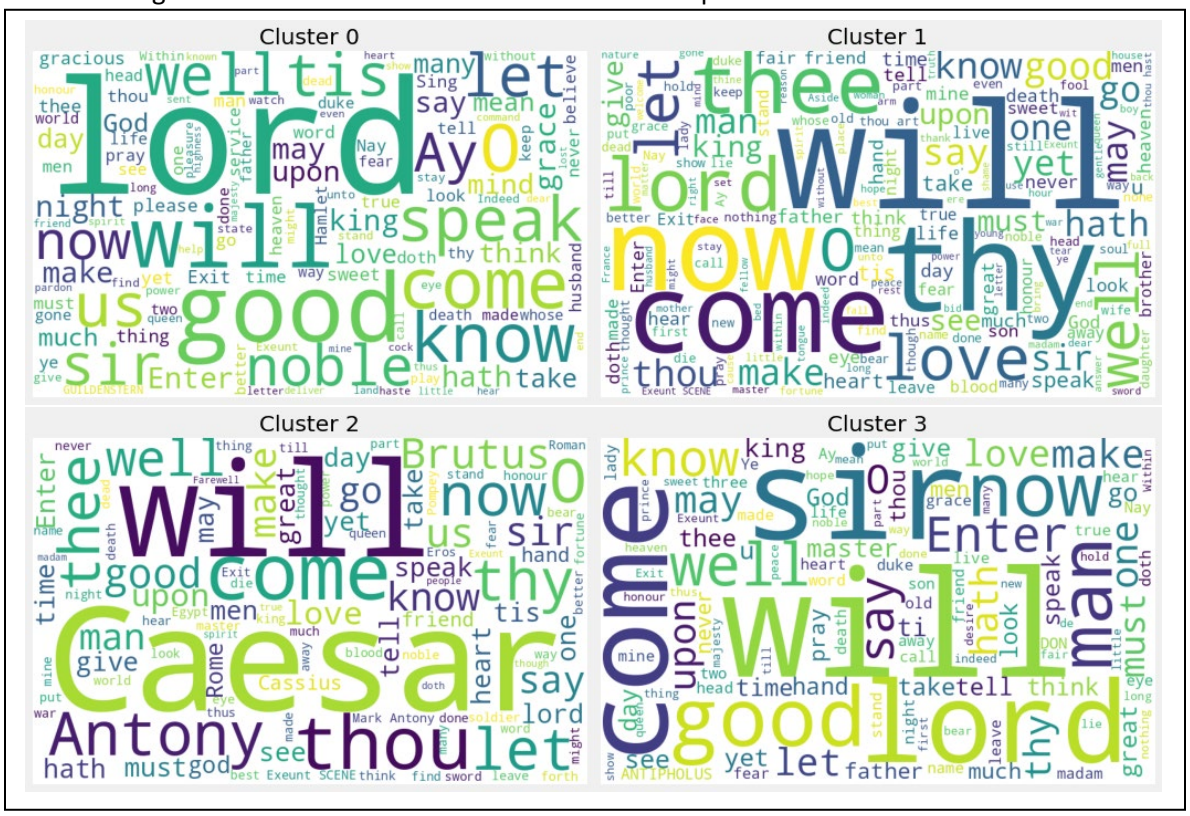
Using the 'elbow method', we determined that a k-gram consisting of four words would be most appropriate.



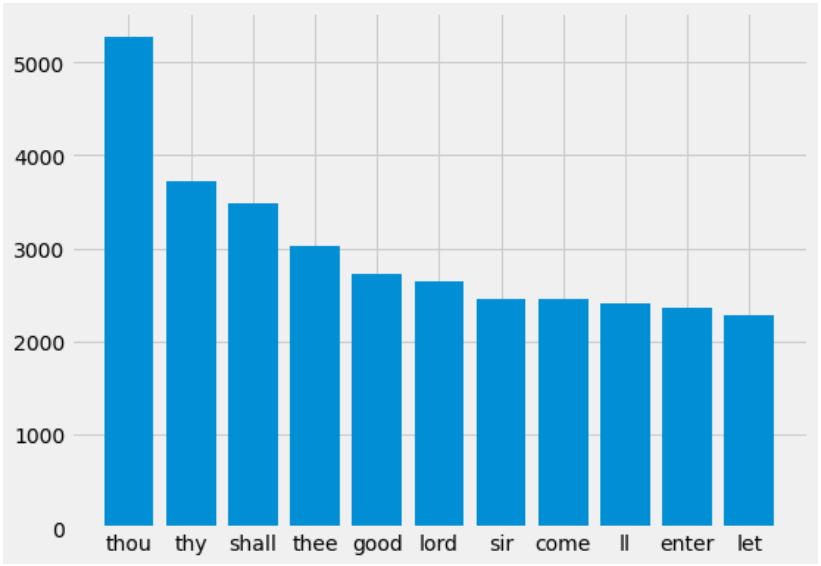
CS 6140 Project Intermediate Report

Jim Morgan, Foad Namjoo, Ariel Oconitrillo

The following word clouds show the most common words per cluster.



The overall most common words are indicated in the histogram below.



Based on the results, there appear to be additional stop words that we would like to add to the library of standard stop words that apply in particular to the Shakespearean dataset. Additionally, more work needs to be done to view the similarities between the words presented in the Shakespearean dataset and those found in film datasets (TBD).