# 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](https://www.kaggle.com/datasets/kingburrito666/shakespeare-plays).

# 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

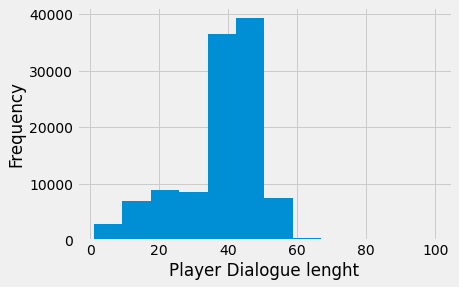
While it is likely little room for improving the professor’s understanding of the data mining techniques used, it is hoped that there will be room for learning more of Shakespeare’s literary techniques and how his word choice and sentence structure influence modern language to the current day.

# Intermediate Report

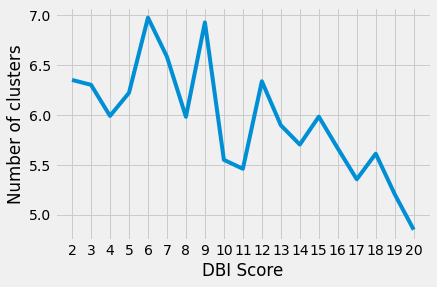
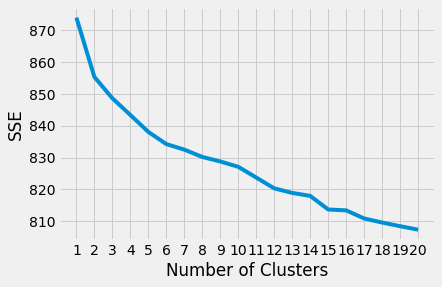
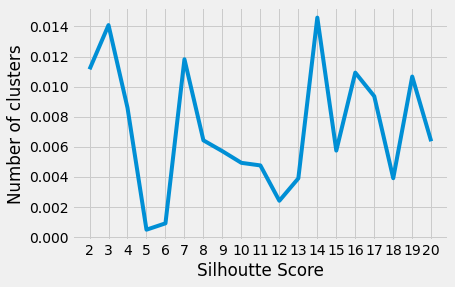
Our first objective was to modify the Shakespearean dataset to consolidate all character dialogue into individual lines.

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

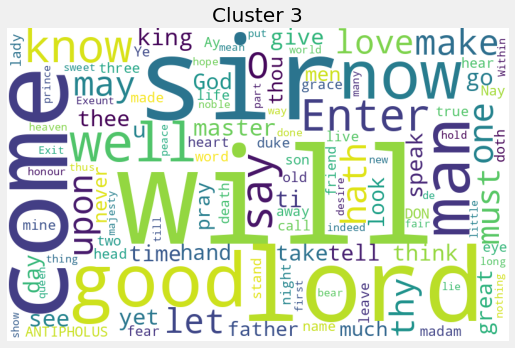
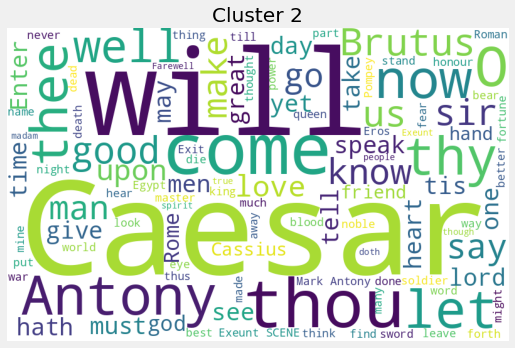
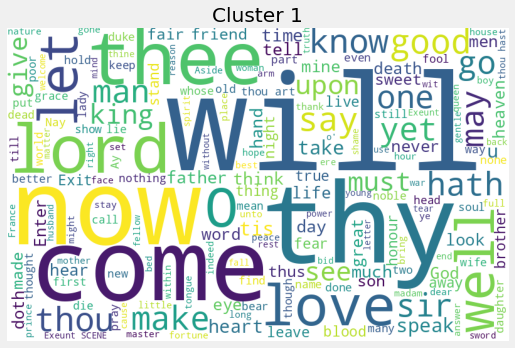
Word associations are formed using k-means clustering, analyzed using Davies Bouldin Score (DBI), Silhouette Score (SS), and the sum of the squared Euclidean distances of each point to its closest centroid (SSE).



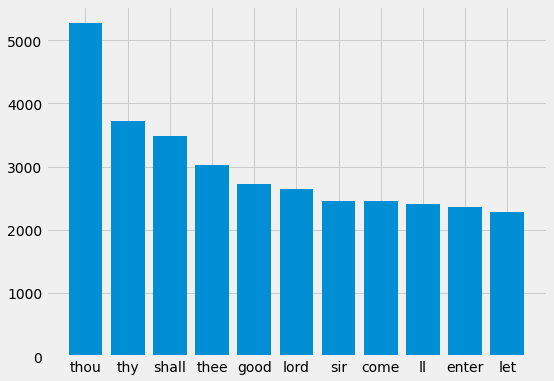
Using the ‘elbow method’, we determined that a k-gram consisting of four words would be most appropriate.



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).

**Characters Similarity Detection in Shakespeare’s Plays**

**ABSTRACT**

The English language is home to some of the greatest playwrights in history, including William Shakespeare. There are 38 plays, 154 sonnets, and other works attributed to him. In high schools and colleges across the US and abroad, his writings are still studied 400 years later. In just 24 years, could a single person write such a diverse set of masterpieces? Perhaps some of the acts or scenes were written by other notables of the day or perhaps a group of his students. In this project after cleaning dataset we uses data mining techniques like n-grams and k-means on the language of Shakespeare’s works.

1. **INTRODUCTION**

The motivation for this research is to explore the similarities between Shakespeare’s players due to plays by multiple characters. We started by collecting the works of Shakespeare, including plays and sonnets.

1. **DATASET DESCRIPTION**

We downloaded all shakespeares plays Our dataset contains 934 unique players and 6 columns. The colomns are Dataline a identical number, Play for identify plays, PlayerLinenumber, ActSceneLine for scene number, Player, and PlayerLine for dialogue.

Graphical user interface, application

Description automatically generated

After reading data, For knowing about lengh of dialogue of each player, we add one column to it as “PlayerLineLength.”

Graphical user interface, application

Description automatically generated

To better illustration we need to derive dialogue length distribution from our dataset. In the histogram below we have shown the length of dialogues frequency. For example, the dialouges length between 40 to 50 had most frequency about 40,000.

Chart, histogram

Description automatically generated

In addition, we concatenate all dialogues of a player in one row using pandas data frame. This resulted in a table with 934 rows for players and 2 columns for player and player dialogues (figure below).

Table

Description automatically generated

Now, we import TfidfVectorizer from sklearn.feature\_extraction.text; and RegexpTokenizer from nltk.tokenize. we convert all uppercase to lower case and remove all stop word. So we reach 22,309 unique words in 934 players. Thus we have a matrix with 22,309 columns and 934 rows. This matrix is calculated by the TF-IDF tokenizing method. Now, we use a 1-word gram on the new dataset without stop words and create a data frame.

Table

Description automatically generated

1. **CLUSTERING**

We performed an additional KMeans clustering using a slightly different approach. We applied the kmeans algorithm to cluster the Shakespeare plays. Now we can do clustering in this matrix by the Kmeans clustering algorithm. Kmeans put the users with similar dialogue in a cluster. The Kmeans put the dataset in the k cluster. To know what is the best number for k we run Kmeans for k from 2 to 20. In each run, we calculated the clustering quality using Davies Bouldin Score (DBI), Silhouette Score (SS), and the sum of the squared Euclidean distances of each point to its closest centroid (SSE). In the figure below we have shown the diagrams.

1. Chart, line chart

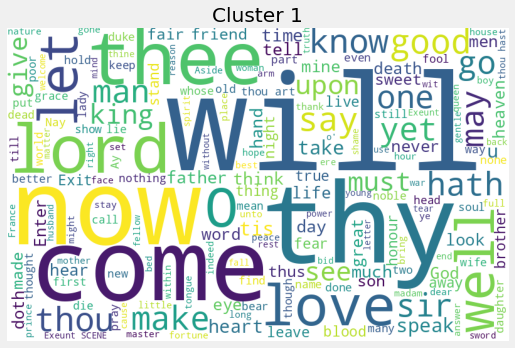
   Description automatically generated
2. Chart, line chart

   Description automatically generatedChart, line chart

   Description automatically generated

Based on DBI, SSE, and Silhoutte 20 cluster is more performance but here we chose 4 clusters for clustering to show. In addition, there are word cloud presentations for each cluster. The words that are more significant in this figure are the most common words in clusters.





A picture containing text

Description automatically generated

A picture containing text

Description automatically generated

For example cluster 0 characters uses Lord, noble, … that may related to people in low rank situation.

.In the following, Frequent words in plays has calculated. We use CountVectorizer from Sklearn for contructing the following matrix. The columns are for unique words and the rows are players. For example row 2 and column “zounds”, is 5, means that this player 5 times use “zounds” in his/her dialouges.



The following histogram shows the most common frequency words in plays after remove stop word.

Chart, bar chart

Description automatically generated