**Final Project**

**“Is any merit to conspiracy theory that Shakespeare was multiple authors”**

**Rebecca Karunakaran Warren Brewster Louis Haskins**

**Introduction**

|  |  |
| --- | --- |
| William Shakespeare was an English poet and playwright who i is considered one of the greatest writers in English Literature.  His works span tragedy, comedy, and historical works, both i in poetry and prose. He is also the most famous playwright in t the world, with his plays being translated in over 50  languages and performed across the globe for audiences of all  ages. Shakespeare was also an actor and the creator of the  Globe Theatre, a historical theatre company. |  |

Between about 1590 and 1613, Shakespeare wrote at least 37 plays and [collaborated](https://www.bl.uk/collection-items/shakespeares-handwriting-in-the-book-of-sir-thomas-more) on several more. These are typically divided into four categories: histories, comedies, tragedies, and romances. His earliest plays were primarily comedies and histories such as *Henry VI* and *The Comedy of Errors*, but in 1596, Shakespeare wrote *Romeo and Juliet*, his second tragedy, and over the next dozen years he would return to the form, writing the plays for which he is now best known: *Julius Caesar*, *Hamlet*, *Othello*, *King Lear*, *Macbeth*, and *Antony and Cleopatra*. In his final years, Shakespeare turned to the romantic with *Cymbeline*, *A Winter’s Tale*, and *The Tempest*.

In his poems and plays, Shakespeare invented thousands of words, often combining or contorting Latin, French, and native roots. His impressive expansion of the English language, according to the Oxford English Dictionary, includes such words as: arch-villain, birthplace, bloodsucking, courtship etc.

Although he is the most-recognized playwright in the world, very little of his life is known. No known autobiographical letters or diaries have survived to modern day, and with no surviving descendants, Shakespeare is a figure both of magnificent genius and mystery. One scholarly explanation for this lack of historical verification is that "William Shakespeare" was the pen name of some more illustrious, well-educated figure of the Elizabethan era.

**Analysis**

**About the Data**

A few plays from each of the Categories “Comedy”, History” and “Tragedy” were chosen to form a collection of **15 plays** that were used for the analysis. Word Cloud indicates that the word “**love**” featured as prominent topic in many of his plays. Other words like “**sir**” and “**thou**” were used commonly in his plays.

|  |  |
| --- | --- |
|  |  |

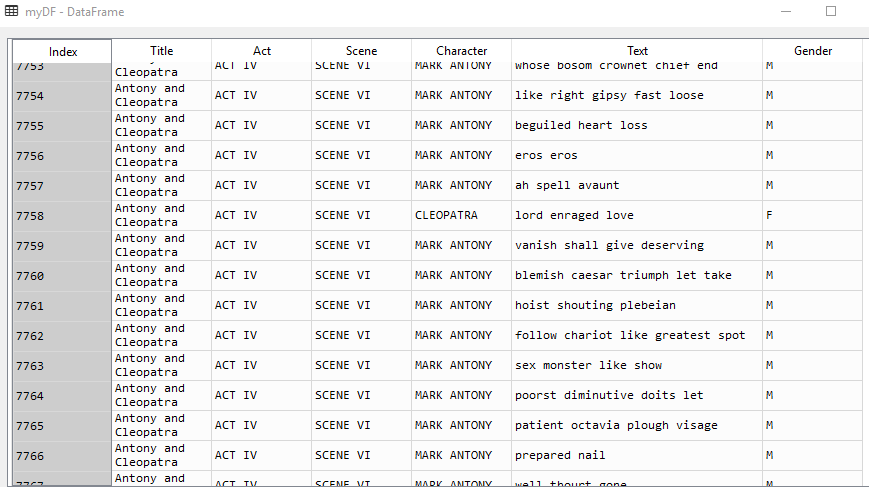
**Preprocessing** of the play data was a very time-consuming process. In order to produce the analysis wanted, the data was segmented into different lists -

* Play Character: This is a list of characters for each play.
* Play: This is the entire text of each play with character reference removed.
* Play Act: This is the entire text of each play segmented by Acts.
* Play Act Scene: This is the entire text of each play segmented by Acts and by Scene.
* Play Act Scene Character: This is the entire text of each play segmented by Acts, by Scene, and finally by Character.

The plays were laid out in a format where the character descriptions would come first and then the play content would follow. Within the character description section, there were several times when a character would be referenced by a different name. For example, in play Julius Caesar, the character Julius Caesar is referenced as **Caesar** and **Octavius Caesar** is referenced as **Octavius.** Special logic was used to handle these situations. There were also rare cases when a character was referenced by two names such as **Mark Antony** in **Cleopatra**. Special logic was also needed for supporting characters such as First Officer, Second Senator, etc.

Following the character description section of each play, is the main body of the play. Logic was created to identify when Acts and Scenes change and this information was stored in the lists - Play Act, Play Act Scene. The beginning of each dialog or speaking line starts with the name of the character which was also stored, sometimes needing special logic as indicated before, to build the lists - Play Act Scene Character.

**For the Gender Analysis,** a corpus with 9000 speaking lines was created using Python Code with about 4500 for Female Character and 4500 for Male Characters so that the corpus was equally balanced between the genders. Screen shot below show the dataframe that was created for Gender Analysis -



Standard pre-preprocessing steps like converting the words to lower case and removing punctuations were applied. When reviewing the dataset, it was clear that NLTK’s English Stop words will not be sufficient since the dialogues are written in what is called “**Early Modern English**” and would require some additional stop words to remove words like “thou”, “yonder”, “art” etc.



A **custom list of stop words** for “Early Modern English” was created and appended to NLTK’s English Stop words which resolved the issue



In order to analyze the **conspiracy theory** that Shakespeare’s works may have be written by other people, a selection of poems attributed to William Shakespeare was compared against poems attributed to both Christopher Marlowe and Edward de Vere. In total there were **52 poems** were analyzed using clustering techniques, with 19 from Shakespeare, 20 from Edward de Vere, and 13 from Marlowe. The reason for using only poems is to compare similarly styled writings, as plays and poems could lead to potential clustering issues between the authors. Furthermore, they were the only style of writings found for the possible alternate authors. Finally, due to the highly artistic style of poetry very little preprocessing was done. Only punctuation marks were removed as they were mostly artifacts from the formatting of the poetry.

**About Tools Analyzed**

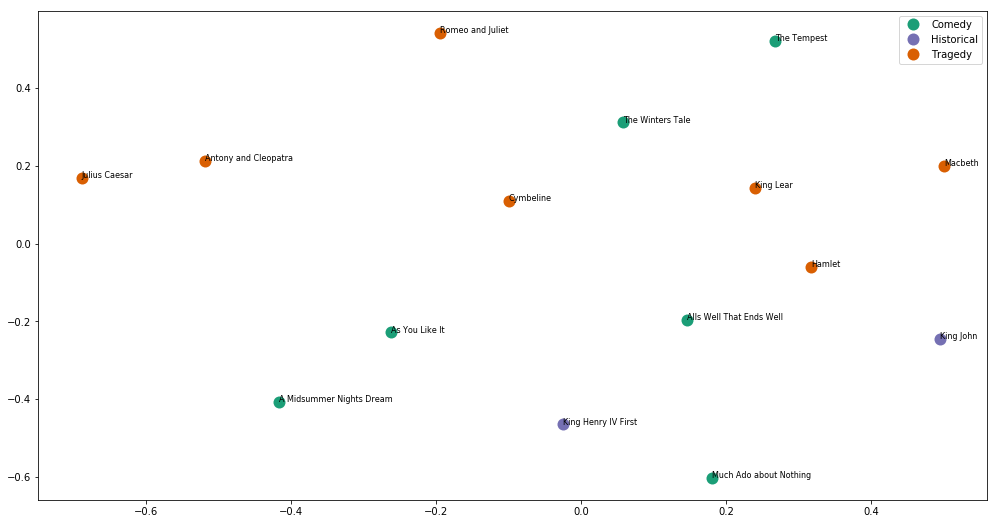
**Naive Bayes** is among one of the most simple and powerful algorithms for classification based on Bayes’ Theorem with an assumption of independence among predictors. It is commonly used for Text Classification, Email Spam Filtration and Sentiment Analysis.

**Multinomial Naive Bayes** is a specialized version of Naive Bayes that is designed more for text documents and explicitly models the word counts. It estimates the probability of the event that one of the N unique words occurs in a position. It is used for larger text. The **Sklearn’s multinomial Naive** **Bayes** classifier is suitable for classification with discrete features like word counts for text classification.

**Matplotlib** is a python library used to create 2D graphs and plots by using python scripts. It has a module named **pyplot** which makes things easy for plotting by providing feature to control line styles, font properties, formatting axes etc. It supports a very wide variety of graphs and plots namely - histogram, bar charts, power spectra, error charts etc.

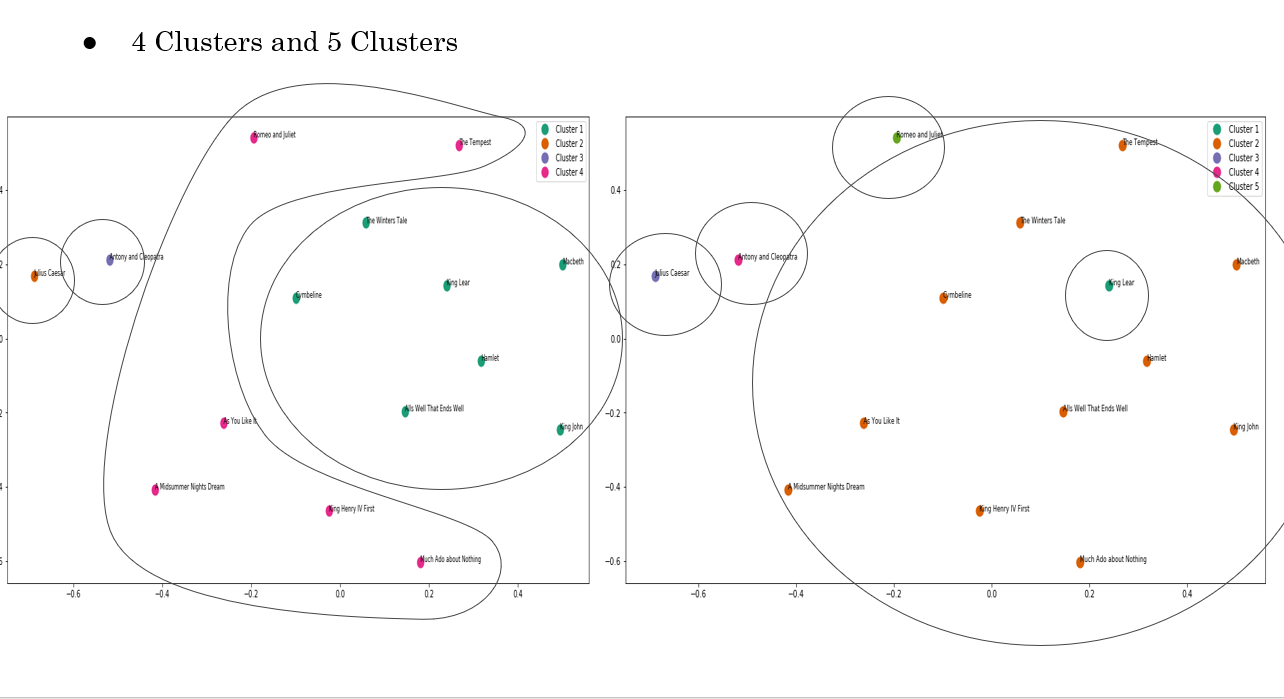
**Result**

Plotting each play by Euclidean distances gives the first glance that it is plausible that more than one person authored these plays. Typically, works from a given author should be closely grouped as they tend to use the same writing style.



Further analysis using plots with various clustering options (K-Means) indicates that it is most likely that **two or three authors** wrote these plays. Any more than that is improbable with the current analysis.

|  |
| --- |
|  |



In order to evaluate the **feasibility of performing a Gender Analysis** on Shakespeare’s Characters the Word Frequency Plot for the 20 most used words for each Gender was created and analyzed to see if any of the common words was used disproportionately by one gender. As the screen shot below indicates, there is some **gender-imbalance in the** words used in the dialogues as indicated by the plot.

|  |  |
| --- | --- |
|  |  |

Using **Matplotlib**’s **pyplot** function a bar chart of the top 20 words was plotted. Looking more closer to the top 20 words used by both genders, it is clear that there are some words that were used commonly by both Genders like “**love**”, but the frequency however was very different. The word “**love**” was used in over **150 speaking lines** for **Female** characters in the Dataset but only about **100 speaking** lines for **Male** characters. In addition, there are words like “**death**” the was commonly used in Male speaking lines but not so much in Female speaking lines.

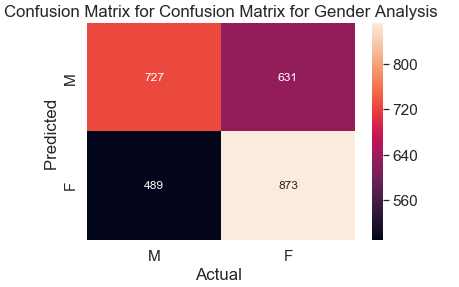
|  |  |
| --- | --- |
|  |  |

The analysis clearly suggests some distinction in the dialogues between the Male and Female characters. The next step was to see if commonly used Text Mining Algorithms can be trained to distinguish these differences.

The Dataset was split using the Hold-Out method to use 70% for Training and 30% for Testing. Three vectorization options were analyzed, the first was using an N-gram Count Vectorizer, the second using an N-gram Boolean Vectorizer and a third using a N-gram TfidfVectoizer with min\_df set to use .0001%. Although both SVM and MNB algorithms rendered similar results for the different vectorization options, MNB consistently scored higher compared to SVM with MNB **using N-gram Count Vectorizer** scoring the highest Accuracy of **59%.**

|  |  |  |
| --- | --- | --- |
|  |  |  |

Looking at the Confusion Matrix for the **MNB using N-gram Count Vectorizer**, it is clear that 631 speaking line that belonged to the Male Character were wrongly predicted as Female character and about 489 Female speaking lines were also wrongly predicted as Male.



Using Uni-gram Tfidf vectorizer, the **top 10 indicative words** using SVM and MNB algorithms clearly show that Male character’s dialogues were relevant to ‘**England’** and security ‘**Guards**” and army “**General**” whereas the Female character’s speaking lines had to do with **family** as indicated by “**husband**” and “**child**”.

|  |
| --- |
|  |

To build a **case for alternate authors,** multiple models using different methods, parameters, and seeds would be used to compare their consistency in predicting each other. The two main model types used were support vector machine (SVM) and multinomial Naive Bayes models (MNB). The types of parameters used for each model were unigram only Boolean, unigram only term frequency (TF), and unigram + bigram TF. Finally, a total of 7 different seeds were used.

|  |  |  |  |
| --- | --- | --- | --- |
| Confusion Matrix Average | Christopher Marlowe | Edward de Vere | William Shakespeare |
| Christopher Marlowe | 1.19047619 | 0.380952381 | 2.142857143 |
| Edward de Vere | 0.785714286 | 4.071428571 | 1.571428571 |
| William Shakespeare | 2.547619048 | 0.642857143 | 2.642857143 |

When comparing the confusion matrices across all the models and seeds, one thing that stood out was the difficulty in the models differentiating between Shakespeare and Marlowe. Looking at the confusion matrix average across all models, the average misclassification between **Marlowe** and **Shakespeare** stands out as the **most misclassified poems**. The misclassifications averaged over 2 poems either way between the two when the test sets consist of 16 total poems. The only other misclassification of note is classifying Edward de Vere’s poems as Shakespeare’s poems at roughly 1.5 poems out of 16. Interestingly though, Shakespeare’s poems were not frequently classified as de Vere’s poems.

To further build evidence for potential alternate authors of William Shakespeare, the conditional probabilities of the MNB models would be compared.

|  |  |  |
| --- | --- | --- |
| **Histogram of Conditional Probabilities** | **Bin** | **Frequency** |
|  | Christopher Marlowe | 32 |
|  | Edward de Vere | 72 |
|  | William Shakespeare | 56 |
|  | Marlowe or de Vere | 30 |
|  | Marlowe or Shakespeare | 60 |
|  | de Vere or Shakespeare | 36 |
|  | Marlowe or de Vere or Shakespeare | 50 |

The probabilities would be distributed into 6 potential categories. For each model, they would be classified based on the authors that had over 10% likelihood predicted. Based on the counts, the models were good at classifying de Vere’s and Shakespeare’s poems but not Marlowe’s. Marlowe’s poems instead appeared to be confused with Shakespeare’s poems or Shakespeare’s and de Vere’s poems, but not solely with de Vere’s poems**. This is just further evidence that Marlowe could be a potential alternate author for Shakespeare**.

**Conclusion**

|  |  |
| --- | --- |
| William Shakespeare was the most influential writer of all-time, bringing a lyrical element to plays about great kings and poor paupers alike. He wrote his earlier plays in the traditional style of the time and about people who seemed real instead of using stock characters as was common in the theater during his days. Additionally, Shakespeare’s work deviated from that of his contemporaries in that he wrote for every type of person who came to the theater or read poems, not just for the upper class as was common | https://lh4.googleusercontent.com/Fm3JMqtAxQG78cse651QgJO-BGhTTsttP2ANbqOvDJ-9hGwXZw7gutXWm4dkAtg40oCot8wmWapi9rXIfTozRR0b4fN5eBQSMhZnNS85Yfqakb9TCSWBrNGfiRucnfGGmtbP56xRmOg |

Although Shakespeare’s plays support the stereotypes of women and men and their various roles and responsibilities in society, he was also a writer who questioned and challenged those representation. His plays indicated that he seemed have raised questions about the standard images of males and females, about what the characteristics of each gender are, about what is defined as masculine and feminine, about how each gender possesses both masculine and feminine qualities and behaviors and about the roles women and men should play in acting out the stories of their lives. Whether Shakespeare intentionally wrote his plays so that female characters could be distinguished from his male characters based on language and linguistic style, **text mining algorithms were clearly able to distinguish some subtle differences** as indicated with the analysis on a sample corpus.

There is so little known about the real [William Shakespeare](https://www.telegraph.co.uk/culture/theatre/william-shakespeare/) unfortunately and with no papers found that document the writing of his many works or any hand-written notes or manuscripts it is hardly surprising that there are plenty of theories that suggest that his works may be that of other authors. However the analysis using a sample set of Shakespeare Plays with **Text Mining using clustering clearly indicates that there is evidence to support the theory of multiple authors**. The Classification Algorithms further supports the case that **Christopher Marlowe** could have written some of Shakespeare’s works indicating that further research with more historical evidence should be conducted to build a case for the theory of alternate Shakespeare authors!