Summer 2020 ISMT S-117 Final Project Report
To be or not to be Shakespeare?
Student Name: Haitao Shang

Introduction

Shakespeare, who has been broadly considered as one of the greatest bard and playwright in humans' history, changed the direction of theater for centuries across different languages and cultures. Shakespeare penned 37 plays over the course of his lifetime and almost all of these plays have been classical masterpieces. However, it has been doubted that if it is really possible that Shakespeare was able to possess the creative strength and focus to craft all of these great works only by himself [1]. Some researchers have suggested that, *rather than a single genius, Shakespeare probably was actually a team of playwrights or the joint effort of some collaborations* [1]. However, (obviously) there has been no conclusion on this question. In this project, I will use the techniques from natural language processing (NLP) to explore this challenging problem.

Dataset

The dataset that I will use for this project is from Kaggle [2]. This dataset is a collection of all the sentences in all 37 plays by Shakespeare and it is in a CSV format. Each row in this file is one line from Shakespeare's plays, and there are 111596 rows in total. For each line in a row, some other relevant information is also provided: the name of the play where this line is from, the actual line being spoken at a given time, the Act-Scene-Line where this line is from, and the name of the player who is saying this line. This dataset provides the low-hierarchy information (words), intermediate-hierarchy information (sentences), and high-hierarchy information (plays). This structure of the dataset nicely provides information at different levels, which allows us to explore the data of Shakespeare's plays from different perspectives and therefore can provide us some deep insights into the question I want to answer.

Exploratory Analyses and Methodology

Indeed there is some risk to use the NLP techniques to study this question, because it is possible that these plays will be classified into different categories: comedy, tragedy and history, and eventually we cannot get a classification of the writing styles of the plays. However, we do not know if a technique is useful or not before trying it. As shown below, my results show that the plays are NOT simply classified into three categories (comedy, tragedy and history) when using several different machine learning methods.

After some exploration in the structure and characteristics of the dataset, I think my objective and the dataset together define an unsupervised learning task. The question that *if Shakespeare is one person or multiple people* is very complicated and cannot be answered directly. Therefore, I proposed some metrics for solving this problem. Given the information in the dataset, studying the patterns on the 111596 lines may provide us some hints. But what patterns are related to authorship? Here I propose several possible metrics: (1) writing style, such as the choice of words used and the

common length of sentences (the latter may be a weak metric), (2) ways of expressing emotions throughout the plays including comedies and tragedies, and so on.

Different NLP approaches were attempted: (1) "word2vec", "doc2vec", "TF-IDF": these methods were used to assign feature vectors for each observation (an observation can aggregate different levels; (2) "sentiment Analysis" and "average length of the sentences": these methods were used to create new numerical features from the textual information, besides the vectors. I assume that different authors normally have different ways of expressing emotions, and thus the sentiment analysis can be used for clustering. In the same way, different authors may write sentences with very different lengths, and so it may be a way to make them out. But please notice the latter looks like a weak metrics. Because, based on my personal experience and observations, people can decide to use long or short sentences in an arbitrary way.

However, the approaches listed above still have not answered the question: if Shakespeare is one person or multiple people. Given the *unsupervised* nature of my problem, Therefore, I used clustering as an approach of analyzing the authorship of Shakespeare's plays based on the results from previous steps. One of our goals is to show that there is a clustering structure within these plays that leads us think that there might have been "more than one Shakespeare" or not; that is, there might have been more than one writers behind these plays. The further the distance between different clusters (while the closer the distance between plays in the same cluster), the more likely the plays were written in different styles, and therefore the more probable there were more than one authors behind these plays. To construct clusters, I used some techniques such as "Agglomerative Hierarchical Clustering", "K-Means" and "Mean-Shift", and compare the clustering results obtained using different algorithms.

Results

Notes: The results obtained using NLP techniques (such as word counts, TF-IDF, LDA, cosine similarities, and so on) and play classifications are not shown in the ipynb notebook rather than this report. Because they are for explorations and do not directly answer our question that if Shakespeare is one person or multiple people. Please refer to the ipynb notebook for these analyses.

The analysis by `Play` should give us a first intuition about the existence of clusters in Shakespeare's plays. To do that, we are concatenating all the lines in a play and extracting the features from each pay using several methods: Word2Vec, Doc2Vec, and TF-IDF. Finally, we are using hierarchical clustering to see how the plays are related.

Figure 1 is a hierarchical clustering with Word2Vec method returns 2 main clusters. It shows that there is a difference among the groups of plays. Let's see what happens with other methods. It is interesting to see that several comedy-tragedy pairs are grouped together as the closet plays in their writing styles. For example, Pericles (comedy) and Hamlet (tragedy), A Midsummer nights dream (comedy) and Troilus and Cressida (tragedy), and so on. Two big clusters appear on above dendrogram; each of them includes several tragedy and comedy. This implies that the comedy/tragedy category probably does not influence much on our analyses of the styles of plays for

determining if the plays are written by one person or a group of people. Similar results are also observed in other analyses, which are shown in Figure 2 and Figure 3.

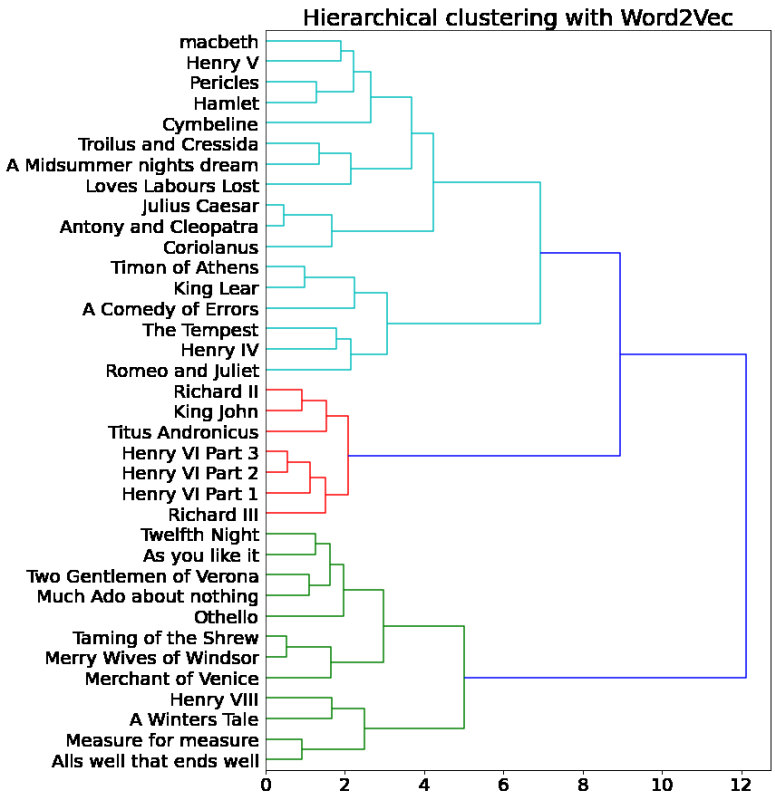

Figure 1. A hierarchical clustering of plays with Word2Vec.

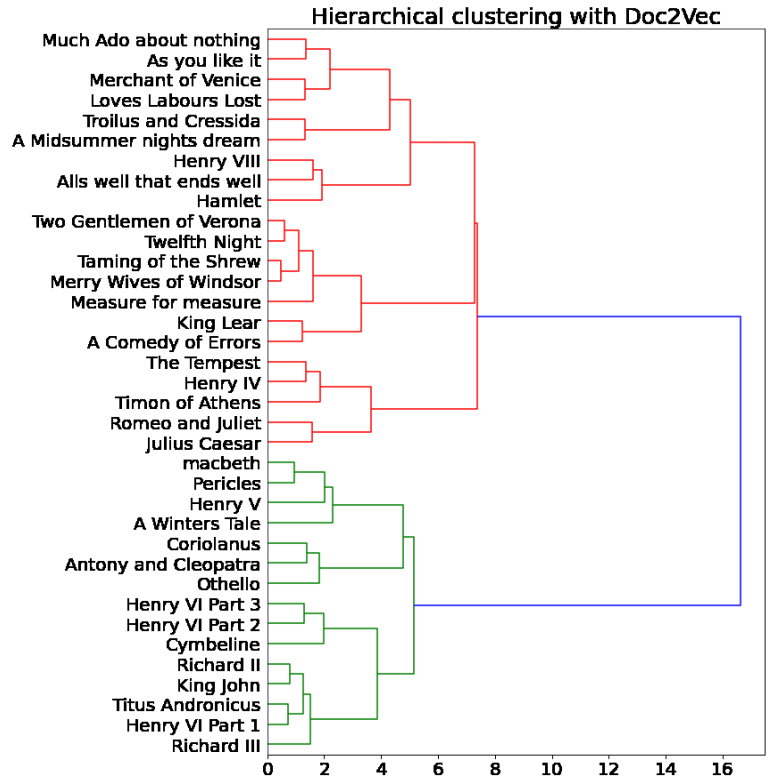

Figure 2. A hierarchical clustering of plays with Doc2Vec.

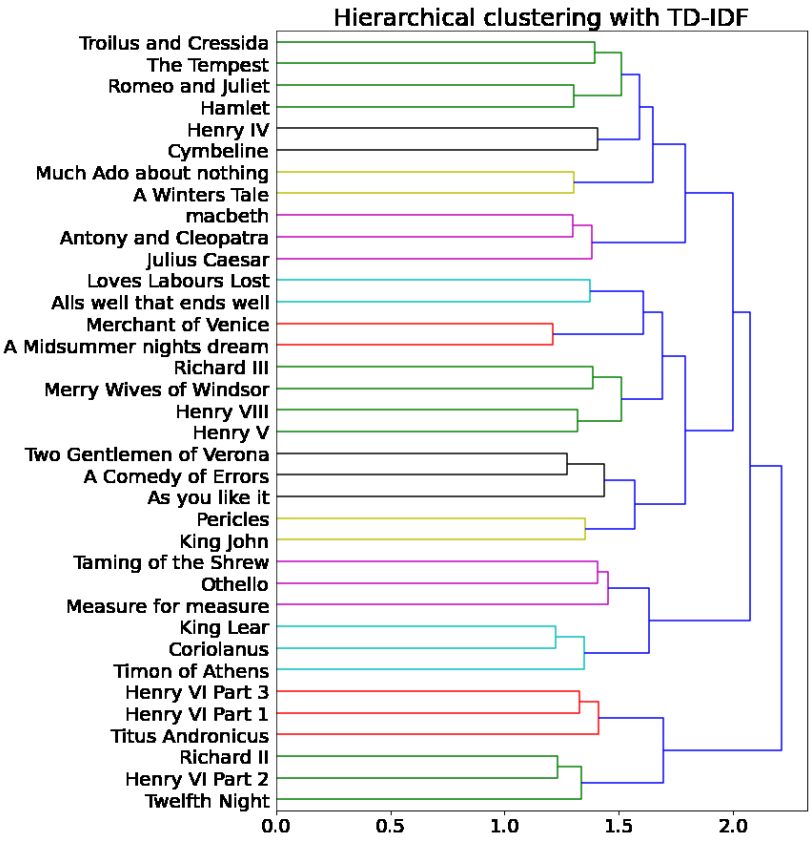

Figure 3. A hierarchical clustering of plays with TD-IDF.

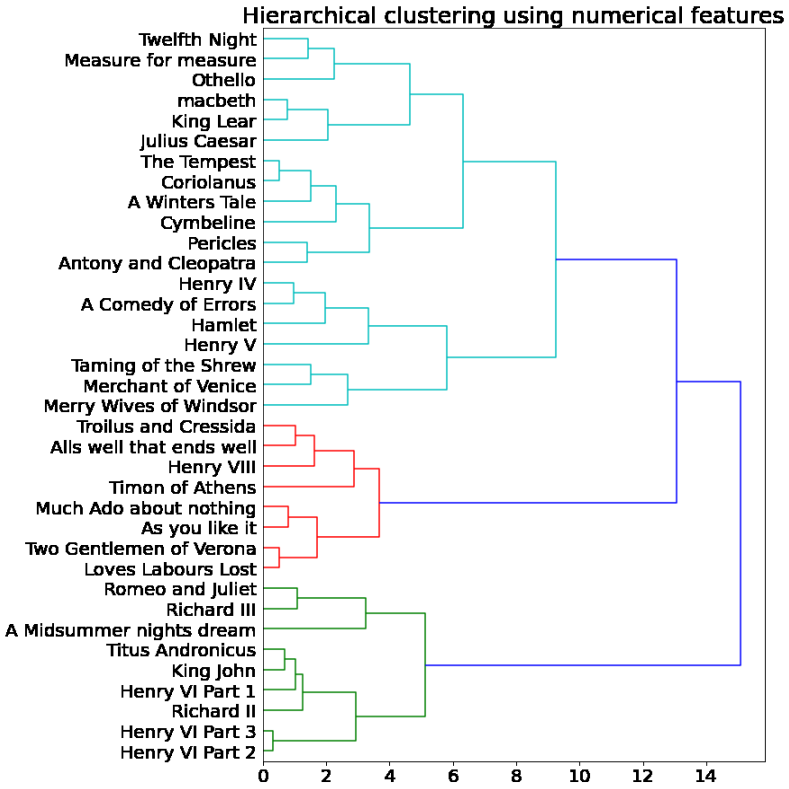

Figure 4. A hierarchical clustering of plays using numerical features (Average length of sentences and Sentiment).

Now we are going to try to find cluster structure using other numerical features in the dataset, such as `PlayYear`, `AvgSentLength`, `NegativeSentiment`, `NeutralSentiment`, `PositiveSentiment`. In this first approach, we are going to do it at Play level, that is, considering all the lines included in a given Play. The result is shown in Figure 4, which demonstrates that even though the structure of the clusters may be quite different (comparing the cluster by the eye the partitions can be very different), the first analysis aggregating the lines per `Play` says that there are few clusters, but well separated.

The analysis per Act is going to be used in more detail shortly. In this section, we only are repeating the hierarchical clustering to see if there are different clusters when considering the Act as the unit to concatenate the lines of a play. *Notice that here we focus more on how many big clusters will be bifurcated from the root of the dendrogram rather than whether comedy and tragedy plays will be grouped together.* Because the latter question has been investigated in previous parts. The results are shown in Figure 5, Figure 6 and Figure 7. Both Word2Vec and Doc2Vec return the same number of clusters as the analysis per Play. TF-IDF returns some more clusters, but we can make out that 5/6 clusters might be a good structure to research. Analysis per Scene does not seem to be very useful, as it is quite likely that local structures distort the results and could be misleading. Again, we are searching for cluster structure using `PlayYear`, `AvgSentLength`, `NegativeSentiment`, `NeutralSentiment`, `PositiveSentiment` previously calculated. The result is shown in Figure 8. At Act level, that is, concatenating all lines included in an Act, we can make out 3 clusters very well defined.

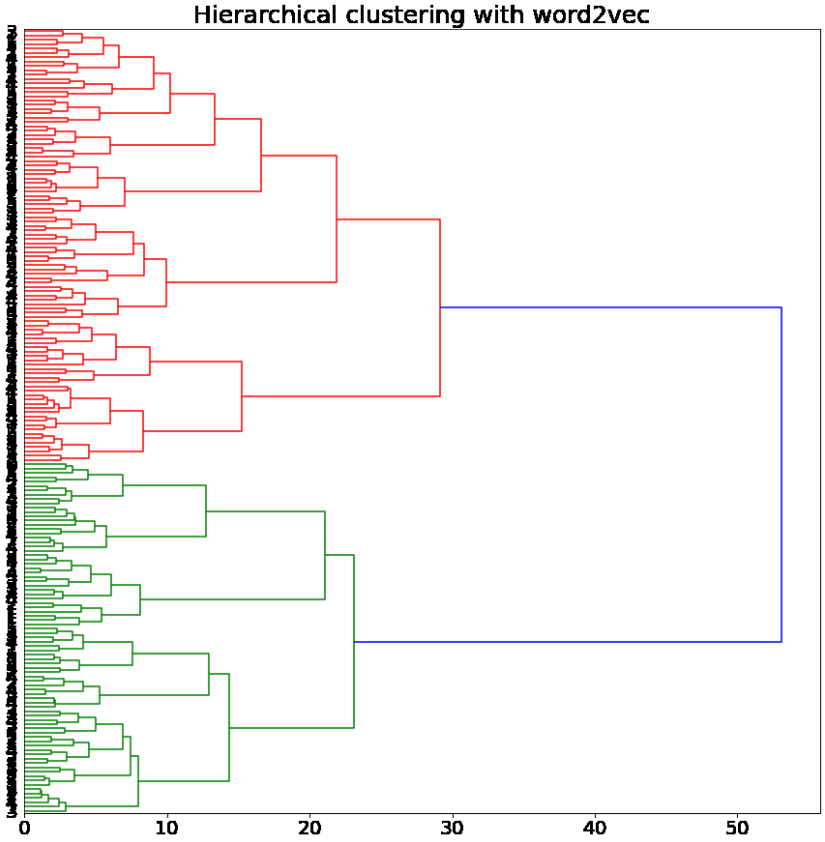

Figure 5. A hierarchical clustering of acts using Word2Vec.

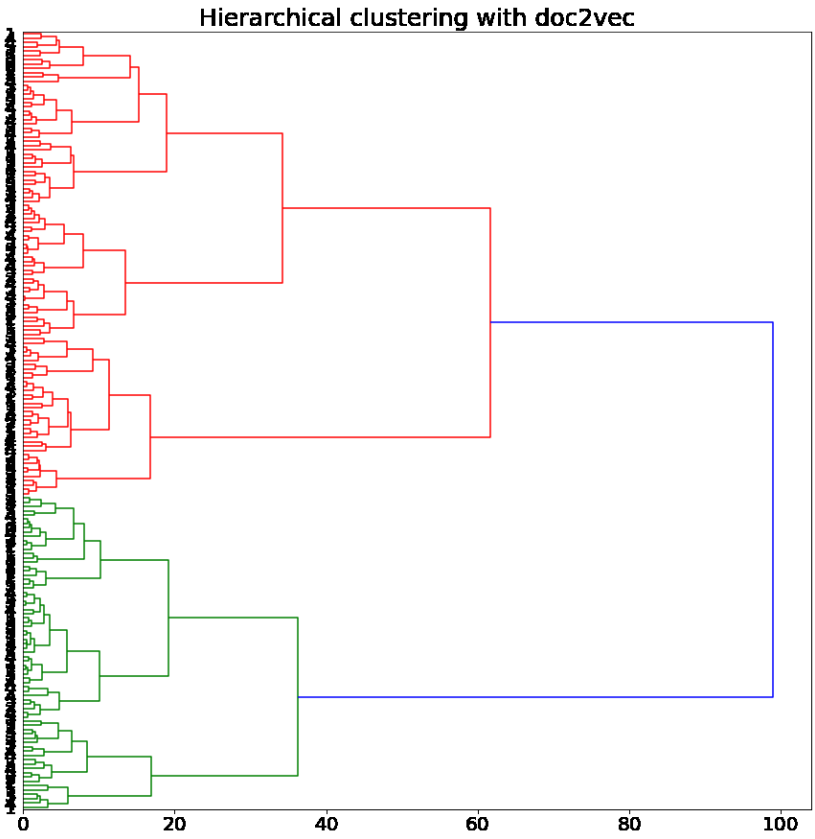

Figure 6. A hierarchical clustering of acts using Doc2Vec.

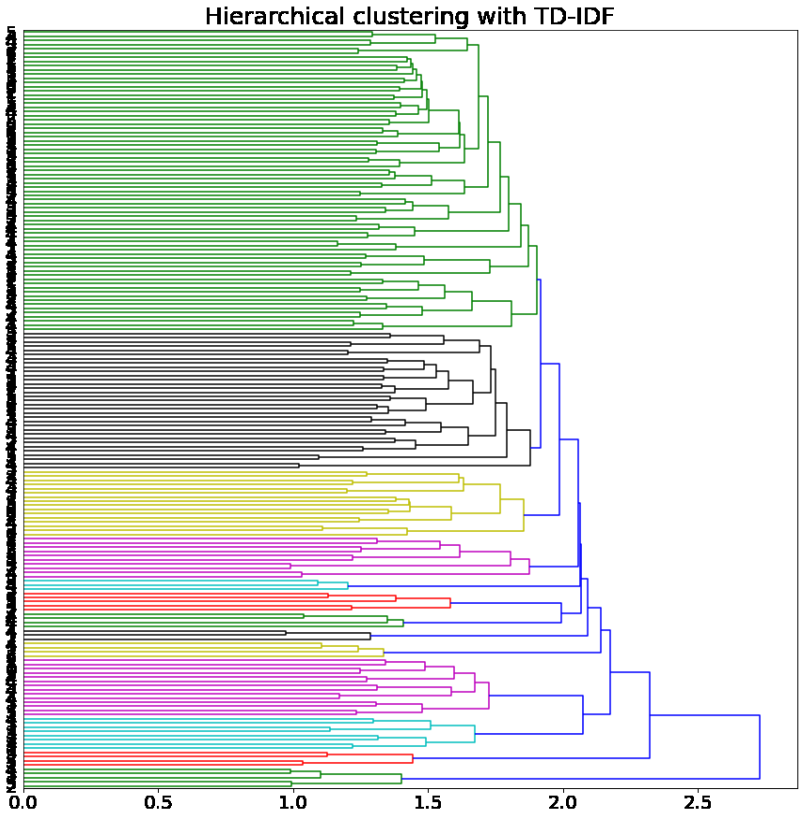

Figure 7. A hierarchical clustering of acts using TD-IDF.

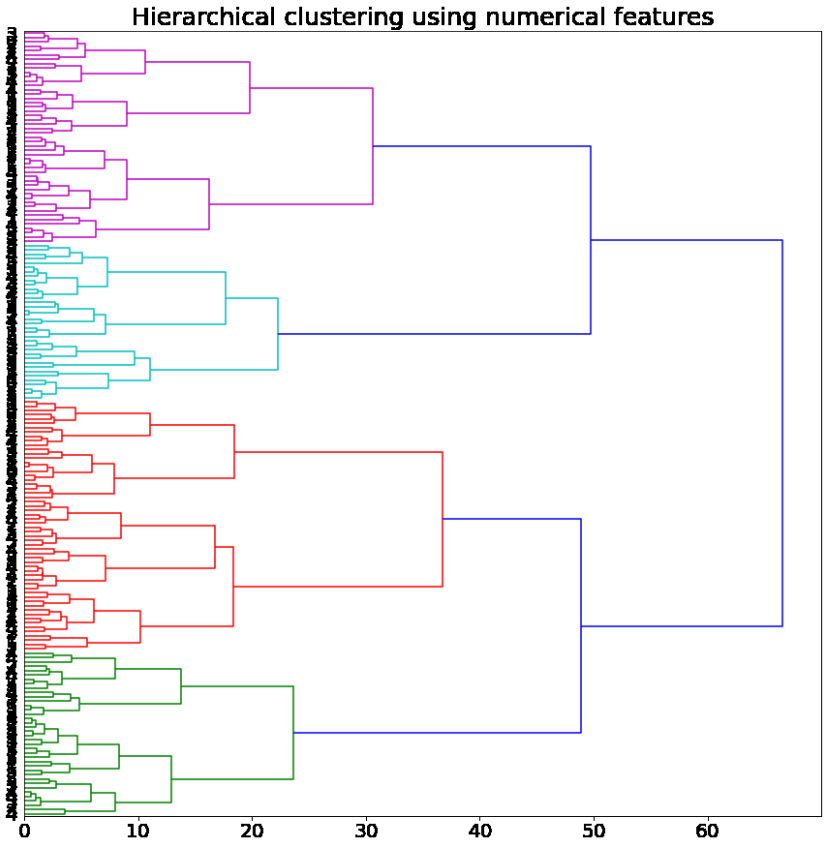

Figure 8. A hierarchical clustering of acts using numerical features (Average length of sentences and Sentiment).

Aggregating at `Act` level we continue finding balanced clusters, that is, there are features that clearly distinguish Plays or Acts from the others, and the reason can be that they were written by different authors. We are going to drill down in our analysis of Shakespeare's plays, by using other clustering methods over the plays, aggregated by `Act`.

The results shown in Figure 9 corresponds with a K-Means with 3 clusters – that is, the most consistent number of clusters obtained in previous steps. The method is executed using the word2vect transformation of the Plays, aggregated by Act, that is, every observation is a concatenation of all the lines included in an Act. In the analysis, we can see that there is kind of a dominant cluster, plus two smaller ones. If we analyze Play by Play, we can see that besides being mono-cluster Plays (something expected), there are Plays including 2 clusters, but rarely 3 clusters. This might be an indication that a Play is split in two authors, but normally not more.

In Figure 10, I performed similar analyses as for Figure 9, but this time using the other numerical features that we had extracted from the Plays: Year, Average Length of the sentences (maybe different authors tend to write sentences with very different lengths) and Negative/Neutral/Positive sentiment. It is interesting that again there is a dominant cluster, plus 2 smaller ones. The distribution of clusters within a Play show the same pattern as before, that is, there are "pure" plays, but there are also plays with 2 clusters, but it is more rare to have a Play with 3 clusters.

Figure 9. Pie Chart of statistics for the clusters of play acts using K-means based on Word2Vec transformation.

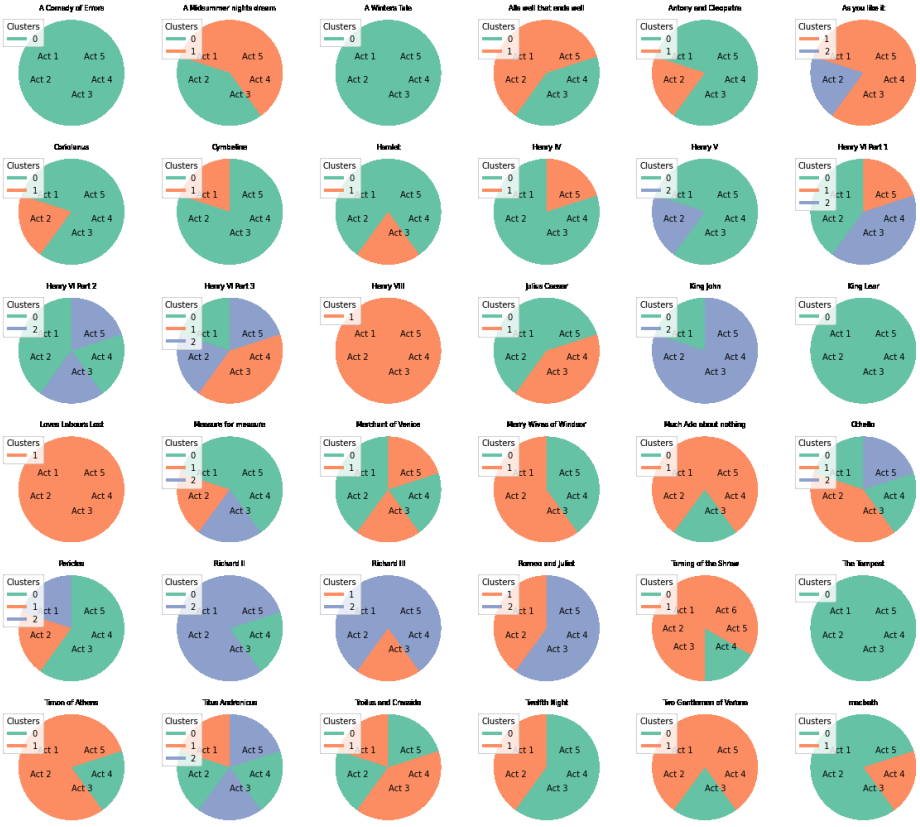

Figure 10. Pie Chart of statistics for the clusters of play acts using K-means based on numerical features (Average length of sentences and Sentiment).

We have used another clustering method, Mean Shift, to see how the clustering structure looks like. Mean Shift is returning 6 clusters, but we can see that there is 1 dominant cluster and then another smaller cluster, being the remaining 4 clusters very residual. Again, we found that there are Plays whose Acts are assigned to the same cluster (expected), and other Plays split into 2 clusters. The number of plays including 3 or more clusters is low.

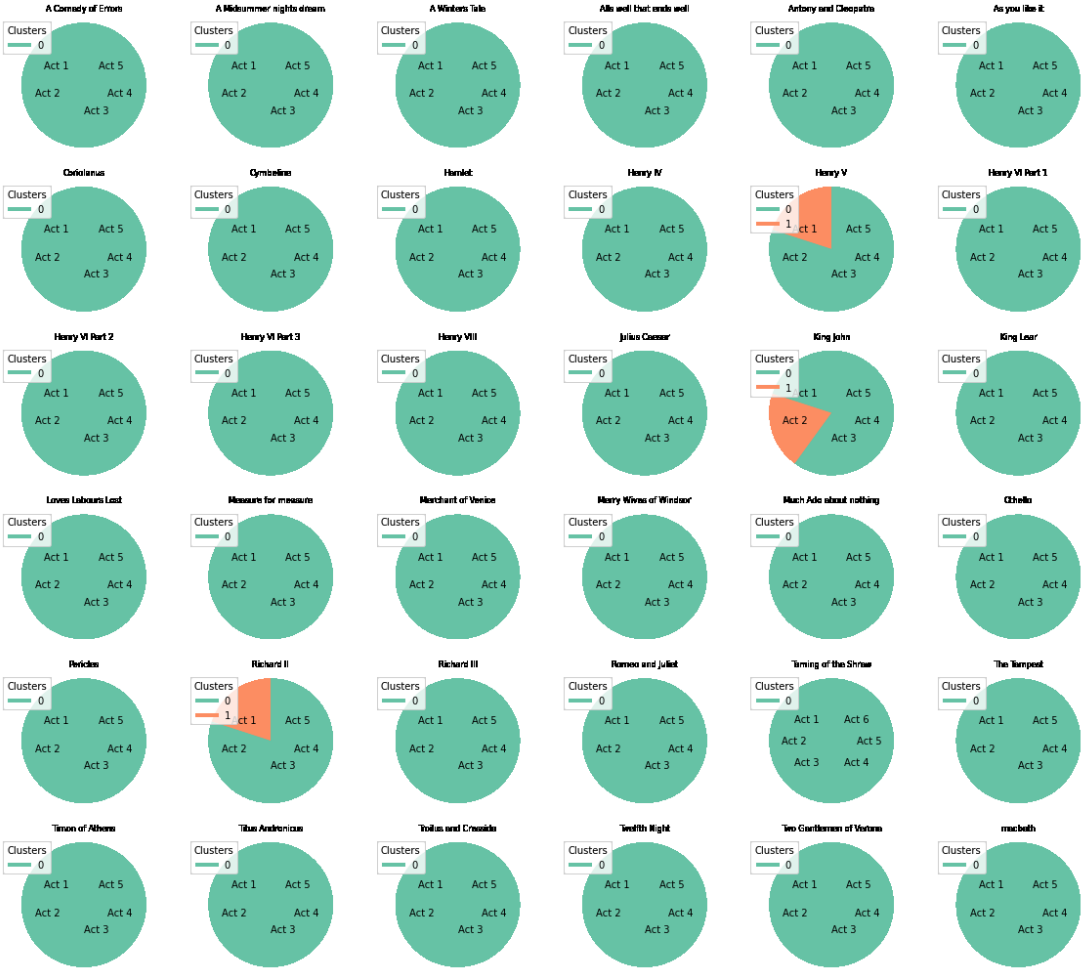

Figure 11. Pie Chart of statistics for the clusters of play acts using Mean Shift based on numerical features (Average length of sentences and Sentiment).

Discussion

Knowing exactly whether or not Shakespeare's plays were written by a genius or a group of people is a very complicated task. At the beginning, we tried several different NLP techniques (including word counts, TF-IDF, LDA, cosine similarity, and so on) to understand the structure and contents in these plays, and also perform some basic supervised learning (classification) using GloVe-based features. Then we used unsupervised learning method (clustering) to explore our initial question that whether one person or a group of people wrote the plays. Indeed we will never know the ground truth to check our model. However, building models that extract clustering structure from the data is possible.

The PlayerLine is a good source of information that we can use, and we can aggregate (that is, concatenate) the lines at different levels. In this project we analyzed Shakespeare's plays at `Play` and at `Act` level. In both analyses we observed that there are some relatively clear clustering structures. This probably is not enough for us to determine exactly if the plays were written by only one or many people, but it provides us a window to investigate that whether Plays and Acts within the Plays have structures and characteristics that make them belong to different clusters.

We analyzed the PlayerLines extracting features using `Word2Vec`, `Doc2Vec` and `TD-IDF`. Also, we extracted negative, neutral and positive `Sentiment` and performed clustering analyses. Then we used Agglomerative Hierarchical Clustering, K-Means and Mean Shift to evaluate the clusters. It is interesting that several comedy-tragedy pairs are grouped together as the closet plays in their writing styles. For example, Pericles (comedy) and Hamlet (tragedy), A Midsummer nights dream (comedy) and Troilus and Cressida (tragedy), and so on. This implies that the comedy/tragedy category probably does NOT influence much on our analyses of the styles of plays. Finally, we found that there are consistently 3 to 4 clusters when we use different methods. This implies that it is possible that Plays (and Acts) were written by a group of people rather than one single person.

Deployment Strategy of Potential Applications

Although what I conducted in this final project is an academic study, the methodology and framework (of code) I developed may find practical applications. Maybe we can extend what I did in this project to a useful package for library managers and readers who need to classify book categories. The other possible application is detecting academic misconduct. Directly copying from other people is easy to detect. However, it may be not that easy to detect that someone modifies or replaces some (key) words and/or re-arrange the sentences from others' papers. With the methodology I proposed above for authorship testing using NLP techniques together with unsupervised learning methods, we may be able to develop an application interface for scholars to check if some suspicious papers are really written by the authors themselves or are simply modified from other people's papers. The null hypothesis behind this deployment would be that assuming a paper we are going to test is modified from others' paper(s). Then we test the statistical significance of how the writing styles (i.e., words, sentences, ways of expressing opinions and emotions and so on) of this paper are similar to the writing styles of papers written by other authors. If the statistical significance is low, then we reject the null hypothesis and conclude this suspicious paper is actually unlikely to be written with academic misconduct. If the statistical significance is high, then we fail to reject the null hypothesis and conclude this suspicious paper is likely to be written with academic misconduct. Although this way is definitely not enough to conclude if someone really has academic misconduct or not, it can be used as an auxiliary tool for such detections.

Closing Remarks

Due to the lack of ground truth, knowing exactly whether Shakespeare is a single person

or a team of playwrights is a very difficult task. The final conclusion may will never be obtained forever. However, using NLP techniques to extract relevant information from the dataset and then providing some inference for this problem are executable. As reported above, I aggregated the lines in the dataset at different levels and investigated the structures, characteristics and styles of the plays and performed clustering analyses on these data. My results suggest that Shakespeare's Plays (and Acts) were probably written by a group of people rather than one single person. But indeed this conclusion and the methods I attempted in this report require further evaluation by future work.

.

References

1. Emma Smith. The Shakespeare Authorship Debate Revisited. *Literature Compass.* 5(3) 2008: 618-632.
2. https://www.kaggle.com/kingburrito666/shakespeare-plays/version/4