Text Mining for the Analysis of Shakespeare

Nicholas J. Ruta
Harvard University
nruta@g.harvard.edu

Peter V. Henstock
Harvard University
peter.v.henstock@pfizer.com

ABSTRACT
William Shakespeare is considered one of the greatest playwrights in the English language. He is attributed with 38 plays, 154 sonnets, and other works. Even 400 years later, his writings are still being studied in most high schools and colleges in the US and elsewhere.

Is it possible that a single person could have written such a diverse set of masterpieces over a window of just 24 years or so? Perhaps some of the acts or scenes were written by other notables of the day or perhaps a group of his students. This project uses machine learning/data mining techniques on the language of Shakespeare's and his contemporaries' works to address the question of authorship.

CCS Concepts
Computing methodologies → Machine learning → Learning paradigms → Supervised/Unsupervised learning.

Keywords
machine learning; data mining; model selection; feature selection; natural language processing.

1. INTRODUCTION
The motivation for this research is to explore the possibility that Shakespeare's corpus of work was authored by multiple people. We started by collecting the works of Shakespeare and several of his contemporaries, including plays and sonnets, that literary scholars have suggested may have helped in the creation of the famous author's corpus. These contemporaries included Ben Jonson, Thomas Middleton, John Fletcher and Christopher Marlowe. The works were broken down into plays, acts and scenes.

2. DATA PREPARATION
Producing high quality data proved to be a challenging aspect of this assignment. We used mechanisms from libraries like BeautifulSoup and NLTK, as well as implementing our own naïve string parser using regular expressions and if statements. The majority of the data preparation work came down to a huge amount of trial and error in addition to navigating a large amount of HTML tags. The plays were broken down to three levels of granularity including play, acts and scenes. We removed all capitalization and "\n" characters from the text during the process of feature engineering.

3. FEATURE ENGINEERING
We spent a while thinking about/reading up on what might distinguish one author from another. The only thing that was consistent is that there is no real consensus and that it depends on the circumstances. Instead, we focused on the kinds of things that Shakespeare is known for.

3.1 Features for Authorship Attribution
First of all, his work has an enormous vocabulary so obviously, we had to include some feature that measured vocabulary size. We came up with dividing the number of unique words by the total number of words in the play to give a "percentage unique" measure of word count. Additionally, we include the notion of "big words". We initially thought of discarding the shorter words, but noted that short words are too distinctive of Shakespeare's style so we left everything in.

In addition to words/vocabulary, we know that Shakespeare wrote his plays prior to the existence of stage direction available in scripts, so all of his intent was captured by his line breaks and punctuation. With that in mind, we examined various metrics for evaluating punctuation use in his plays. Again, we came up with the idea of averaging. In this instance we chose to break up the text by line and by sentence (so ending with a ?, !, or .) and then we took the average number of punctuation marks by line and by sentence to produce 4 more features. To further strengthen our ability to capture the linguistic structure of the works, we incorporated 2, 3 and 4 length n-grams.

We found the top 5 words by word frequency. We chose 5 by producing a histogram of the most common stemmed words and looked for the "first elbow" in the 5-20 range.

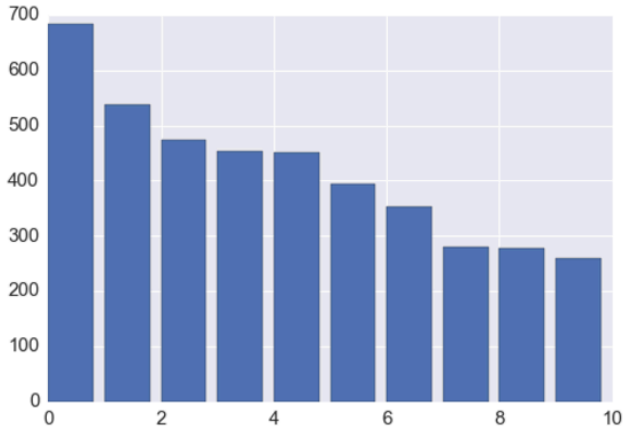

Figure 3.1 Histogram showing the most common stemmed words. It shows a "first elbow" at 5 words.

Our features are named in our data set as: "vocabulary_richness", "average_word_length","average_line_length","punctuation_per_line", "average_sentence_length", "punctuation_per_sentence", "unique_word_count", "word_frequency" and "2 3 and 4 ngrams."

3.2 Feature Normalization

The many works used for our analysis varied greatly in length. There were pieces that were over 3,000 words long and others of around 500 words. In order to address these varying distribution sizes, we normalized by the document length for the unique word count and word frequency features. For the unique word count, we took the unique word count divided by the total word count to produce a 0 to 1 value. For the word frequency, we normalized the top five most frequent words by taking the word's count divided by the total word count. For example, the word "i" appears 184 times in a play with 5093 total words: 184 / 5093 = 0.0361.

4. MACHINE LEARNING TECHNIQUES

We decided to experiment with a variety of supervised and unsupervised machine learning techniques. These techniques included Linear Regression, Principal Component Analysis with KMeans clustering and Random Forest Classification. We experimented with various visualization techniques, such as isomaps and MDS, and then implemented the kmeans and Gaussian Mixture Models clustering methods once again.

4.1 Linear Regression to Predict a Categorical Response

We created a category to contain a 1 for Shakespeare and a 0 for non-Shakespeare prediction of each piece of work. We used a linear regression model to make the predictions based on features involving sentence and word structure.

We looked at the scores and found some Shakespeare play/act/scene selections that had less than 50% for the prediction score based on the fitted linear regression model. For example, Shakespeare's play 'Much Ado about nothing' had a score of 0.426790 (42% likely to be classified as 'shakespeare') for the normalized unique word count prediction.

4.2 Principal Component Analysis & KMeans Clustering

We used PCA to reduce the dimensionality of our data in order to conduct a KMeans clustering. We used the "elbow curve" technique to determine the cluster size and ran our clustering algorithm. The results were that Shakespeare plays were found in each of the three clusters.

4.2.1 Elbow Curve - Determine Optimal Cluster Size

The KMeans algorithm was a clustering technique that we tried. The first step was to determine the number of clusters, or k value, to use. First, we created a range of test clusters from 1 to 10. Then, we used scipy.cluster.vq.kmeans to run the kmeans function and compute centroid and the distortion between the centroids and observed values associated to the distortion that is computed between the centroid and the observed values of the cluster. Using the centroid in each of the group o clusters, we compare the Euclidean distance (2-norm) from all the points in space to the centroids of the cluster using the SciPy provided 'dist' function. We printed the 1 cluster and 2 clusters results to verify. It showed us the distance of each of the observed points from the different centroid. We got the distance of each of the observed points from the different centroids and found the minimum distance that relates to the closet centroid. We could see the cluster that is closest to each observed point. We computed the average of the sum of the square of the distance for each observed point. We printed the results. It was an array of values which represents the average sum of the square from one to ten cluster groups. We plotted an elbow curve chart to determines the k size for the kmeans clustering technique using the above calculated averages. The elbow curve shows that there are bigger jumps from one cluster to the next until slight jumps are observed starting from cluster 3 to 4 and onward. For this reason, we chose a cluster size of 3 to segment the data.

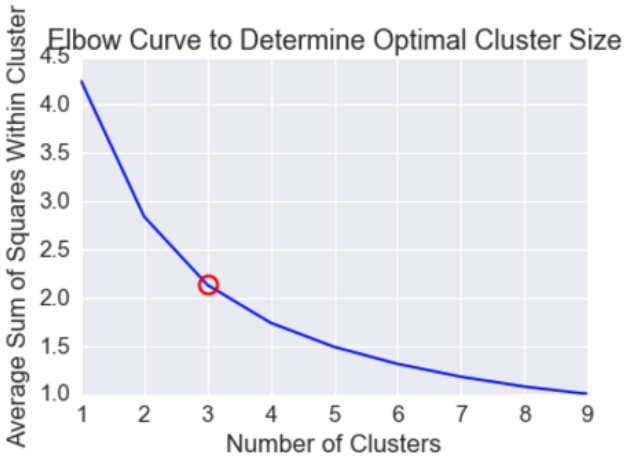

Figure 4.1 Identifying the optimal cluster size using the "Elbow Curve."

cluster_group	author	
0	fletcher	22
	jonson	60
	marlowe	29
	middleton	136
	shakespeare	454
1	fletcher	4
	marlowe	35
	middleton	2
	shakespeare	75
2	fletcher	18
	jonson	6
	marlowe	38
	middleton	23
	shakespeare	425

Figure 4.2 The results of the KMeans clustering showed Shakespeare works present in each cluster.

4.3 Random Forest Classification

We did not include the n-grams features since they are considered sparse data and not well suited for random forest. The number of estimators used was 100 for the random forest. Our research showed that 64-128 is an acceptable range for most cases.

872 out of 954 of the Shakespeare play/act/scene rows returned at least one feature as being classified not Shakespeare.

4.4 Multidimensional Scaling (MDS)

4.4.1 KMeans Clustering

We performed an additional KMeans clustering using a slightly different approach. This time, we applied the kmeans algorithm to cluster the Shakespeare plays. Two key features that we focused on for this were the number of unique words and the number of total words found in the play. Using those key features, we compared the averages found in Shakespeare works with three of Shakespeare's contemporaries. They were Thomas Middleton, Christopher Marlowe and Ben Jonson. We calculated the averages of the unique word counts and total word counts for the works of

each author. The final average used was the sum of the total word count and unique word count divided by two. We compared the averages and determined a 3-group cluster of Jonson, Middleton and Shakespeare. With the clusters defined, we used the Shakespeare corpus and added a column to represent which cluster each Shakespeare work was most aligned with.

	unique_wordcount	total_wordcount	cluster_group
Measure for measure	3610	30803	Thomas Middleton
Antony and Cleopatra	4267	35583	Ben Jonson
The Tempest	3484	26435	Thomas Middleton
A Comedy of Errors	2746	21675	Thomas Middleton
macbeth	3632	24932	Thomas Middleton
Taming of the Shrew	3513	29503	Thomas Middleton
Loves Labours Lost	4086	33763	Ben Jonson
Much Ado about nothing	3249	34362	Ben Jonson
King Lear	4548	43503	William Shakespeare
Henry VI Part 3	3827	37205	Ben Jonson

Figure 4.3 Chart showing Shakespeare plays assigned to a cluster group.

The results seem to express that features, such as the unique words and total word count, found in the data are associated with several authors and therefore provide a strong argument for the theory that multiple authors worked on the Shakespeare corpus.

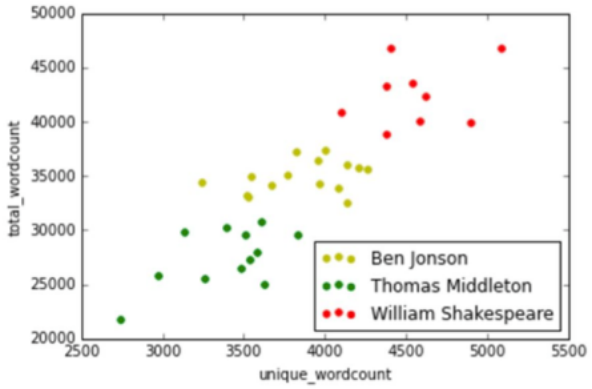

Figure 4.4 Plotting a visualization to see the clustering.

4.4.2 Gaussian Mixture Model (GMM)

The Gaussian mixture model proved to be incredibly accurate in clustering into populations based on our features set. In particular, when clustering into two groups, we found that the author prediction was typically accurate on 75%80% of the scene data points when using either the line based punctuation features or all 6 features.

We experimented with a isomaps and MDS, but in the end we found that MDS produced remarkably distinct visualizations of our clusters. The overlap between authors was apparent. That being said visualizations are clearly much easier to interpret when dimensionality is less than 3. As stated before, the accuracy of the GMM was good enough that we chose to focus on it based on the following two plots.

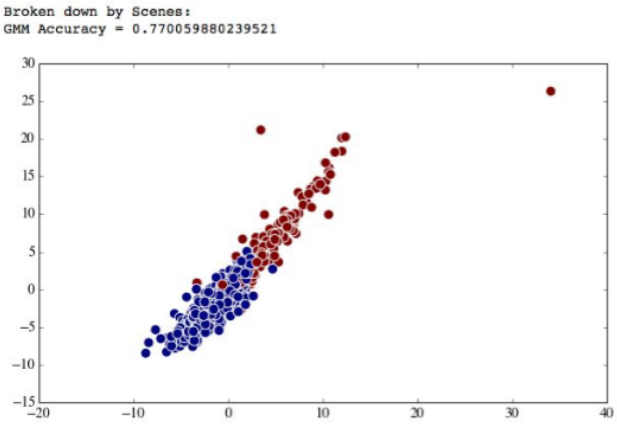

Figure 4.5 Using 6 features with GMM clusters, plotted with MDS. The overlap and differences are clear.

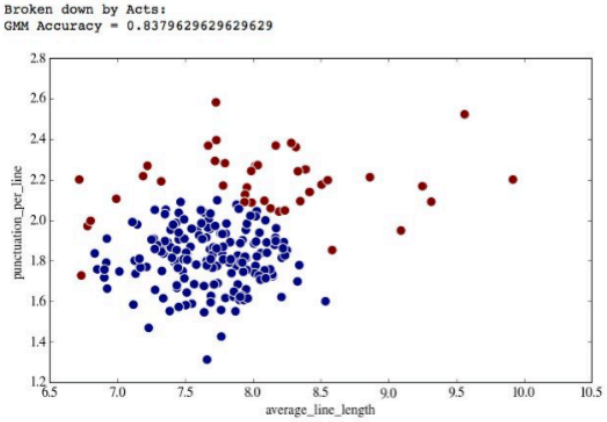

Figure 4.6 Plotting/clustering with two punctuation based features.

We trusted the accuracy of the GMM with the punctuation and all features to the point that we decided to see if we could generate a stable list by repeatedly clustering and intersecting incorrectly assigned items. Most other methods we tried wouldn't produce stable lists after a few iterations, though since we knew they were producing results with less than 60% accuracy we just discarded them. We were able to produce a list of 24 scenes that were consistently classified as Middleton's work, even though they are in Shakespeare plays. Of particular note were 3/10 scenes in Henry VI act II.

4.5 KMeans Clustering w/ Stop Words Removed, Stemming and TF-IDF

We used sklearn and nltk to do a KMeans clustering involving each play/act/scene's full document. We used a Snowball Stemmer and removed 'STOP' words according to nltk's default English list. Since plays, acts and scenes vary so much in length, we used term frequency-inverse document frequency to adjust. The entire corpus consisted of 19,417 words. The cluster results were:

full_doc_kmeans_tfidf_cluster_group	author	
0	marlowe	86
	middleton	3
	shakespeare	169
1	fletcher	33
	jonson	66
	marlowe	16
	middleton	116
	shakespeare	318
2	fletcher	11
	middleton	42
	shakespeare	467

Figure 4.7

We did the same as above but for 2,3,4 NGrams. Here were the results for all three NGram sets -

bigrams_kmeans_tfidf_cluster_group	author	
0	fletcher	6
	marlowe	38
	middleton	24
	shakespeare	191
1	fletcher	4
	jonson	40
	marlowe	14
	middleton	82
	shakespeare	184
2	fletcher	34
	jonson	26
	marlowe	50
	middleton	55
	shakespeare	579

Figure 4.8 The clustering results for bigrams.

trigrams_kmeans_tfidf_cluster_group	author	
0	shakespeare	8
1	middleton	3
	shakespeare	23
2	fletcher	44
	jonson	66
	marlowe	102
	middleton	158
	shakespeare	923

Figure 4.9 The clustering results for trigrams.

fourgrams_kmeans_tfidf_cluster_group	author	
0	shakespeare	2
1	shakespeare	4
2	fletcher	44
	jonson	66
	marlowe	102
	middleton	161
	shakespeare	948

Figure 5 The clustering results for fourgrams.

5. CONCLUSION

We drew several conclusions based on the entirety of our research and implementations. From our last run of the kmeans algorithm, the results seem to express that features, such as the unique words and total word count, found in the data are associated with several authors and therefore provide a strong argument for the theory that multiple authors worked on the Shakespeare corpus. Namely, from the GMM and our features choices, we believe that Thomas Middleton had a hand in writing these scenes:

A Winters Tale:5:2

Coriolanus:4:7

Cymbeline:2:5

Cymbeline:3:3

Henry VI Part 1:4:6

Henry VI Part 2:2:2

Henry VI Part 2:3:1

Henry VI Part 2:4:10

Henry VI Part 3:2:1

Henry VI Part 3:4:4

Henry VIII:1:0

Henry VIII:5:5

King Lear:3:3

Merry Wives of Windsor:4:6

Pericles:1:2

Pericles:5:6

Richard II:1:2

Richard II:2:1

Richard II:5:1

Richard III:3:6

Richard III:5:1

Richard III:5:5

Romeo and Juliet:1:4

Romeo and Juliet:2:3

Twelfth Night:4:3