CS 6140 Project Final 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. Additional comparative literature has been obtained through the Guttenberg project, from works such as Alice’s Adventures in Wonderland, Moby Dick, The Great Gatsby, and The Tragedy of Dr. Faustus.

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 modern day.

Semi-Contemporary Literature Analysis

Background

This part of the report, done by Ariel, will analyze the similarities between the works of Shakespeare and popular books from after Shakespeare's time to see how alike works since Shakespeare are to his works and how much of an impact he has had. The books selected to be analyzed against Shakespeare’s were “A Room With a View” by E.M. Forster, “Middlemarch” by George Eliot, “Moby Dick” by Herman Melville, “The Enchanted April” by Elizabeth Von Arnim, “The Blue Castle” by L.M. Montgomery, “Cranford” by Elizabeth Cleghorn Gaskell, “Frankenstein” by Mary Wollstonecraft Shelley, “Pride and

CS 6140 Project Final Report

Jim Morgan, Foad Namjoo, Ariel Oconitrillo

Prejudice" by Jane Austen, "Alice's Adventures in Wonderland" by Lewis Carroll, and "The Great Gatsby" by F. Scott Fitzgerald. The transcripts of each book were obtained from the top 25 most downloaded books on Project Gutenberg, which has copies of copyright free books. Books were selected to contain a good amount of variety in genres and in order to get a variety of authors only one book from an author was allowed to be used. Unfortunately, due to copyright concerns, no books from the current century were used.

Bag of Words

After getting the manuscripts of the books we used the Misra-Gries algorithm to find the 50 most common words in each work. To get more accurate results punctuation was stripped from the text so words followed by or starting with any sort of punctuation were not viewed as different from those without. All words were also completely in lower case so that placement didn't affect words.

Here are three examples of the most common words found in the books (the order is not specific):

The 50 most common words in "The Great Gatsby":

> 'we', 'it', 'will', 'and', 'run', 'faster', 'stretch', 'out', 'he', 'his', 'i', 'our', 'arms', 'further', 'one', 'to', 'fine', 'morning', 'so', 'was', 'beat', 'boats', 'against', 'current', 'borne', 'on', 'back', 'ceaselessly', 'into', 'of', 'for', 'the', 'past', 'green', 'light', 'orgastic', 'future', 'year', 'in', 'by', 'recedes', 'before', 'us', 'eluded', 'then', 'but', 'that', 'a', 'that's', 'no'

The 50 most common words in "Pride and Prejudice":

> 'from', 'had', 'been', 'means', 'uniting', 'visits', 'uncle', 'his', 'her', 'aunt', 'to', 'of', 'city', 'with', 'gardiners', 'they', 'as', 'were', 'always', 'most', 'intimate', 'the', 'terms', 'darcy', 'well', 'was', 'elizabeth', 'i', 'really', 'loved', 'both', 'a', 'ever', 'at', 'which', 'she', 'them', 'in', 'sensible', 'warmest', 'gratitude', 'and', 'towards', 'persons', 'who', 'by', 'bringing', 'you', 'on', 'into'

The 50 most common words in "Moby Dick":

> 'padlocks', 'when', 'in', 'their', 'great', 'mouths', 'savage', 'with', 'i', 'sea-hawks', 'coffin', 'sailed', 'sheathed', 'beaks', 'second', 'sail', 'drew', 'and', 'near', 'it', 'nearer', 'picked', 'me', 'floated', 'at', 'last', 'was', 'up', 'devious-cruising', 'a', 'rachel', 'the', 'her', 'retracing', 'of', 'day', 'that', 'search', 'on', 'to', 'after', 'missing', 'children', 'only', 'found', 'by', 'another', 'orphan', 'as', 'its'

The list of Shakespeare's most common, across all his plays, words that were used for analysis are:

> "the", "and", "i", "to", "of", "a", "you", "my", "in", "that", "is", "not", "me", "it", "with", "for", "be", "his", "your", "this", "he", "but", "have", "as", "him", "will", "so", "what", "her", "no" , "all", "do", "by", "shall", "if", "are", "we", "our", "on", "good", "now", "lord", "oh", "from", "sir", "come", "at", "they", "enter", "or"[1i]

[1] (Note: because Shakespeare's English had some words that are not common in more recent times, translations were made to match a more modern language. Changes include 'hath' to 'have', 'tis' to 'is', 'thou/thee' to 'you', 'o' to 'oh', and 'thy' to 'your'.)

CS 6140 Project Final Report

Jim Morgan, Foad Namjoo, Ariel Oconitrillo

Jaccard Distance Analysis

After having lists for the most common words in each book and across all of Shakespeare's plays, we used Jaccard's Similarity to find how alike the works were to Shakespeare. The results are reported below.

Title	Jaccard Similarity with Shakespeare
Alice's Adventures in Wonderland	10/90 = .111
Cranford	17/83 = .205
Frankenstein	22/78 = .282
Middlemarch	22/78 = .282
Moby Dick	16/84 = .19
Pride and Prejudice	17/83 = .205
Room With a View	15/85 = .176
The Blue Castle	14/86 = .163
The Enchanted April	10/90 = .111
The Great Gatsby	20/80 = .25

Table 1 – Comparison of Jaccard Similarity with Shakespeare's plays

We next used Jaccard's Similarity of Pride and Prejudice with the other nine books. We picked Pride and Prejudice because it is the oldest of the books, meaning it was the only one with a chance of influencing future publications.

Title	Jaccard Similarity with Pride and Prejudice
Alice's Adventures in Wonderland	9/91 = .099
Cranford	14/86 = .163
Frankenstein	12/88 = .136
Middlemarch	17/83 = .205
Moby Dick	14/86 = .163
Room With a View	14/86 = .163
The Blue Castle	15/85 = .176
The Enchanted April	14/86 = .163
The Great Gatsby	12/88 = .136

Table 2- Comparison of Jaccard Similarity with Pride and Prejudice

From Table 1 the average Jaccard similarity to Shakespeare's works was 0.198. From Table 2 the Jaccard similarity of the nine other books to Pride and Prejudice was 0.156. Though it may not be large, the books are more like Shakespeare than Pride and Prejudice. Another interesting find is that the books from the 1900s (Room with a View, The Blue Castle, The Enchanted April, and The Great Gatsby) had lower average similarities to Shakespeare than the ones from the 1800s, .131 vs .213 respectively. These comparisons would need to be made with more books, but it is possible that as time goes on Shakespeare's works are having less impact on writers.

Summary

Overall, the results were perhaps a bit unexpected. Looking through the list of the most common words in Shakespeare's works it might be expected that the average Jaccard's Similarity would be higher than

CS 6140 Project Final Report

Jim Morgan, Foad Namjoo, Ariel Oconitrillo

just 20 percent. Some words such as “lord” and “sir” are probably not expected to be used as much in the times since, but the rest of the words are common. Also, names that are frequently used in a book will also lower the similarity to Shakespeare, as no names appear in the most used words in Shakespeare's works. A problem could be with the output of the Misra-Gries algorithm which can at times output infrequent values, however most of the words it output seemed to be correct. Regardless, 20 percent similarity is still high, so it seems that Shakespeare’s influence was still strong when it came to future popular works of literature.

Dataset Description

Foad performed the bulk of Shakespearean data analysis. The dataset contains 934 unique players and 6 columns. The columns are Dataline an identical number, ‘Play’ to identify plays, ‘PlayerLinenumber’, ‘ActSceneLine’ for scene number, ‘Player’ for the character speaking, and ‘PlayerLine’ for the line of dialogue.

	Dataline	Play	PlayerLinenumber	ActSceneLine	Player	PlayerLine
0	1	Henry IV	NaN	NaN	NaN	ACT I
1	2	Henry IV	NaN	NaN	NaN	SCENE I. London. The palace.
2	3	Henry IV	NaN	NaN	NaN	Enter KING HENRY, LORD JOHN OF LANCASTER, the ...
3	4	Henry IV	1.0	1.1.1	KING HENRY IV	So shaken as we are, so wan with care,
4	5	Henry IV	1.0	1.1.2	KING HENRY IV	Find we a time for frighted peace to pant,

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

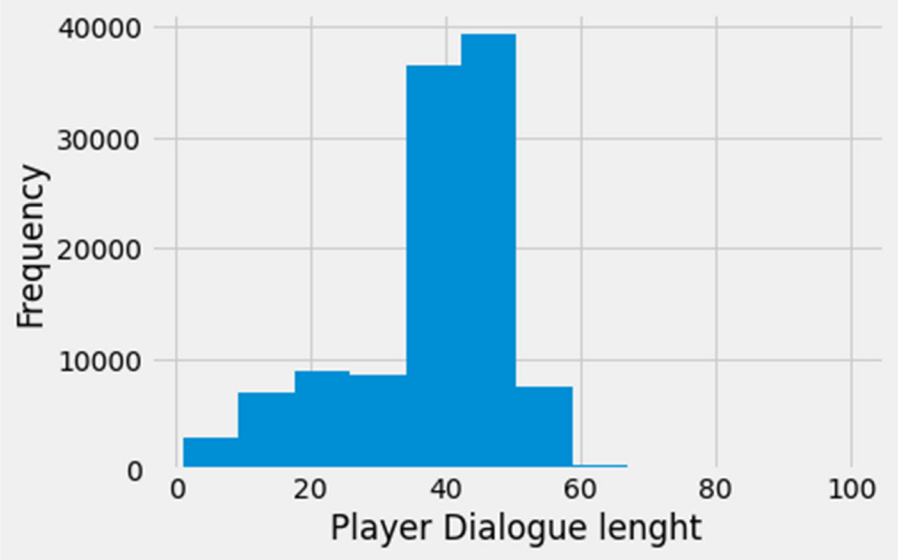
	Dataline	Play	PlayerLinenumber	ActSceneLine	Player	PlayerLine	PlayeLineLength
0	1	Henry IV	NaN	NaN	NaN	ACT I	5
1	2	Henry IV	NaN	NaN	NaN	SCENE I. London. The palace.	28
2	3	Henry IV	NaN	NaN	NaN	Enter KING HENRY, LORD JOHN OF LANCASTER, the ...	96
3	4	Henry IV	1.0	1.1.1	KING HENRY IV	So shaken as we are, so wan with care,	38
4	5	Henry IV	1.0	1.1.2	KING HENRY IV	Find we a time for frighted peace to pant,	42

Dialogue Length Analysis

To better illustrate we need to derive dialogue length distribution from our dataset. In the histogram below we have shown the length of dialogues frequency. For example, a dialogue length between 40 to 50 had the highest frequency at about 40,000 occurrences.

CS 6140 Project Final Report

Jim Morgan, Foad Namjoo, Ariel Oconitrillo



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).

	Player	PlayerLine
3	KING HENRY IV	So shaken as we are, so wan with care, Find we...
36	WESTMORELAND	My liege, this haste was hot in question, And ...
114	FALSTAFF	Now, Hal, what time of day is it, lad? Indeed,...
115	PRINCE HENRY	Thou art so fat-witted, with drinking of old s...
221	POINS	Good morrow, sweet Hal. What says Monsieur Rem...
...	...	...
109776	FLORIZEL	These your unusual weeds to each part of you D...
109781	PERDITA	Sir, my gracious lord, To chide at your extrem...
109967	DORCAS	Mopsa must be your mistress: marry, garlic, To...
109969	MOPSA	Now, in good time! I was promised them against...
110662	Shepard	So 'tis said, sir, about his son, that should ...

934 rows × 2 columns

TF-IDF Analysis

Now, we import TfidfVectorizer from sklearn.feature_extraction.text; and RegexpTokenizer from nltk.tokenize. Using the default method parameters, we converted all words to lower-case and removed stop words in the supplied ‘english’ dictionary. We reached a final total of 22,309 unique words in 934 players, comprising a matrix with 22,309 columns and 934 rows. This matrix is then tokenized using TF-IDF. Now, we use a 1-word gram on the new dataset without stop words and create a data frame.

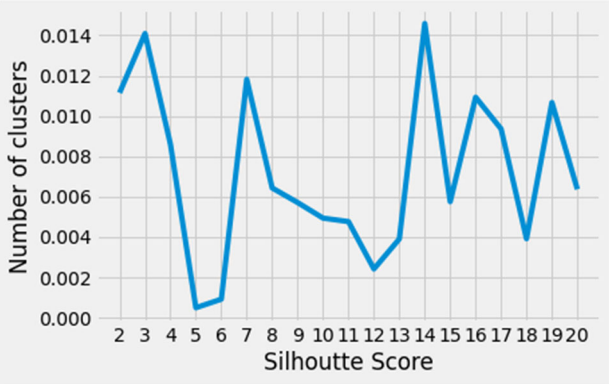
CS 6140 Project Final Report

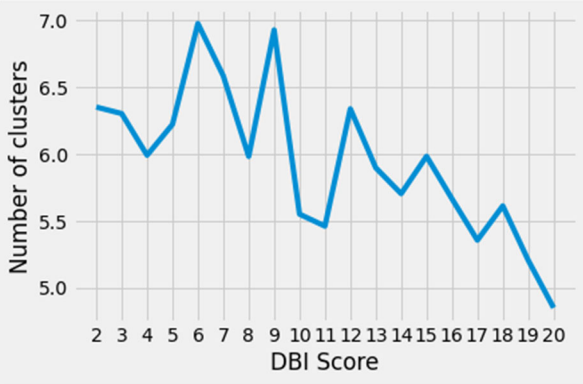
Jim Morgan, Foad Namjoo, Ariel Oconitrillo

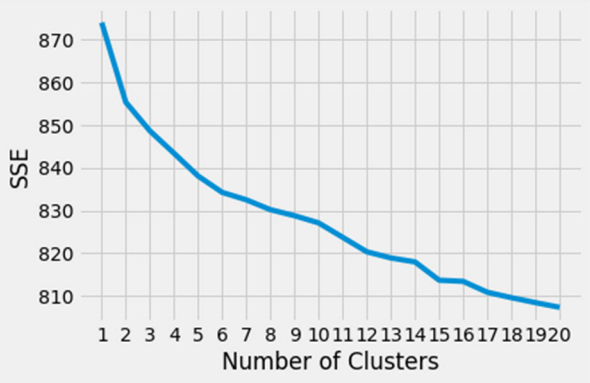
```
10                1
machine           1
mackerel          1
maculate          1
maculation        1
                ...
good           2724
thee           3025
shall          3485
thy            3727
thou           5273
Name: 934, Length: 22281, dtype: int64
```

Clustering

We performed an additional K-Means clustering using a slightly different approach. We applied the K-Means algorithm to cluster the Shakespeare plays. Now we can perform clustering on this matrix by the K-Means clustering algorithm. K-Means put the users with similar dialogue in a cluster. K-Means put the dataset in the k cluster. To identify the optimal k-value, we screen values of 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.



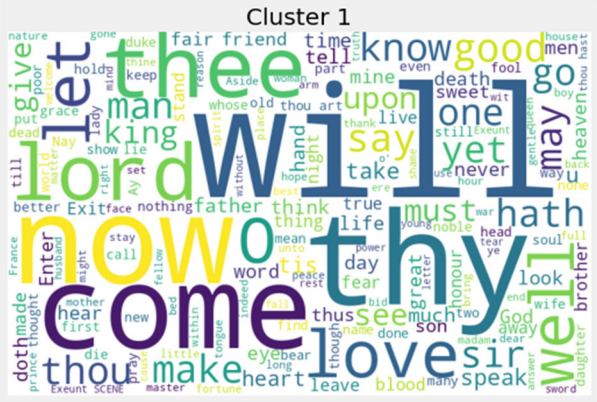


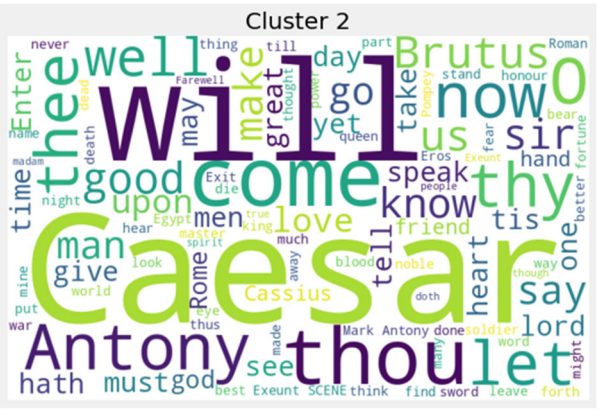


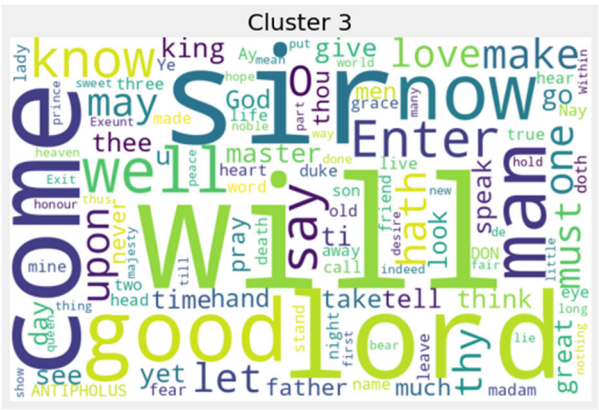
CS 6140 Project Final Report

Jim Morgan, Foad Namjoo, Ariel Oconitrillo

Based on DBI, SSE, and Silhouette 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.







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

In the following, Frequent words in plays have been calculated. We use CountVectorizer from sklearn for constructing 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 lines of dialogue.

	10	2d	2s	4d	5s	6d	8d	aaron	abaissiez	abandon	...	zenelophon	zenith	zephyrs	zir	zo	zodiac	zodiacs	zone	zounds	zwaggered
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	5	0
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	1	0

CS 6140 Project Final Report

Jim Morgan, Foad Namjoo, Ariel Oconitrillo

Connection to Film

There was an objective initially to look for connections between Shakespearean plays and their modern film equivalents. Such films as “Warm Bodies” (Romeo and Juliet), “She’s The Man” (Twelfth Night), and “The Lion King” (Hamlet) all have been derived from Shakespearean works. These recent films, unfortunately, bear little lexicographical similarity to their respective inspiring works. Extensive work was done building one, two, and three n-gram analyses with little quantitative measures to show. Nonetheless, some themes do still show through.

Romeo and Juliette – n-Gram Analysis

For instance, in “Romeo and Juliet”, study of the unigrams shows common usage of the terms ‘love’ and ‘Romeo’, with death a fairly distant third. Juliet didn’t even make the top 10!

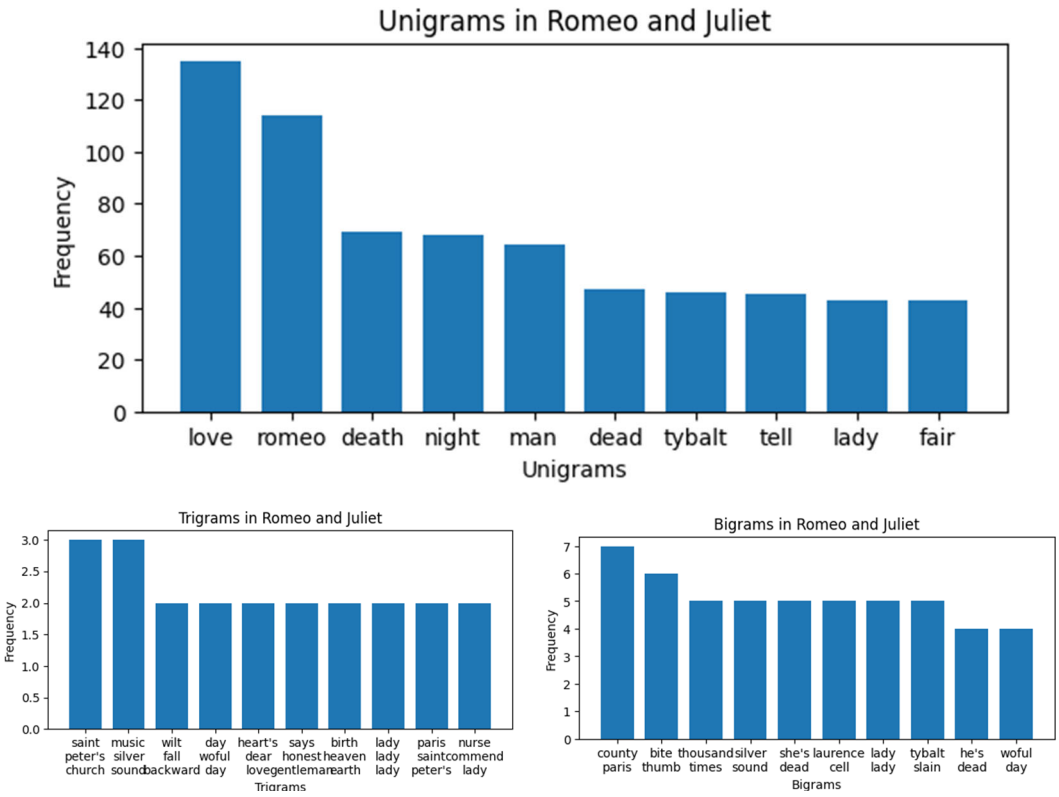
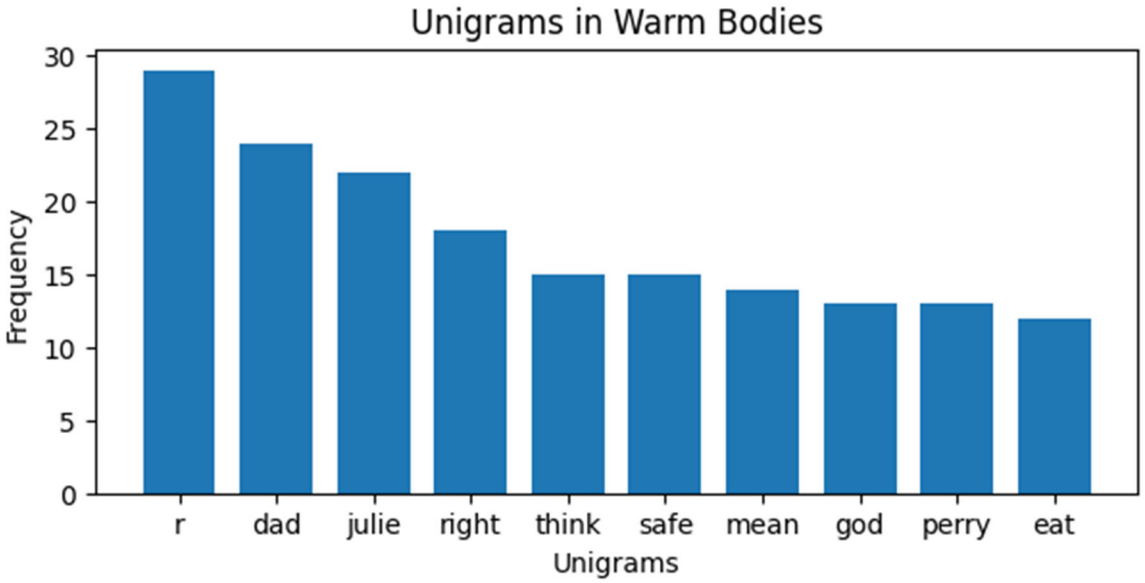


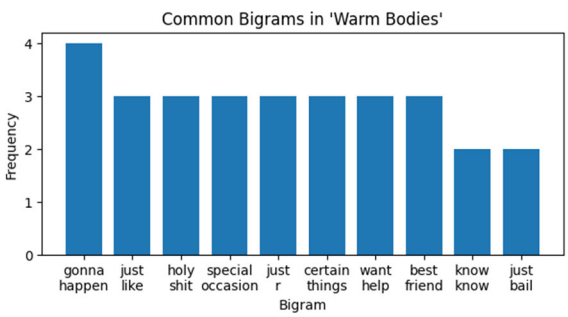
Figure 1- N-Gram analysis of Romeo and Juliet

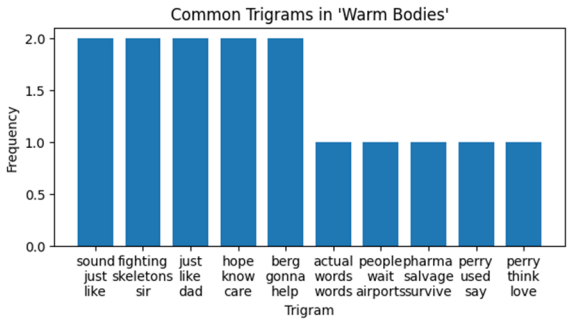
Similar unigram analysis of “Warm Bodies” show a markedly different tone. Despite being a movie revolving about zombies, the only implied reference is the word ‘eat’ cracking in the number 10 spot. Instead, words like ‘R’ (the film’s protagonist and equivalent to Romeo), ‘Dad’ and ‘Julie’(Juliet) are most predominant. This shows a very character-driven approach to the story, with less overt emphasis on thematic elements.

CS 6140 Project Final Report

Jim Morgan, Foad Namjoo, Ariel Oconitrillo







Bigram analysis of “Romeo and Juliet” begins to show some of the Shakespearean prose. The insult regarding biting one’s thumb (the 16th century equivalent of flipping the bird) is one of the most prevalent phrases. Other phrases lend themselves to the tragedy of the piece, such as “he’s dead”, “she’s dead”, and “woful day.” The bigrams found in “Warm Bodies” are less creative, though still quite colorful, with 50% less repetition.

Shakespeare’s Connection to Contemporary Authors

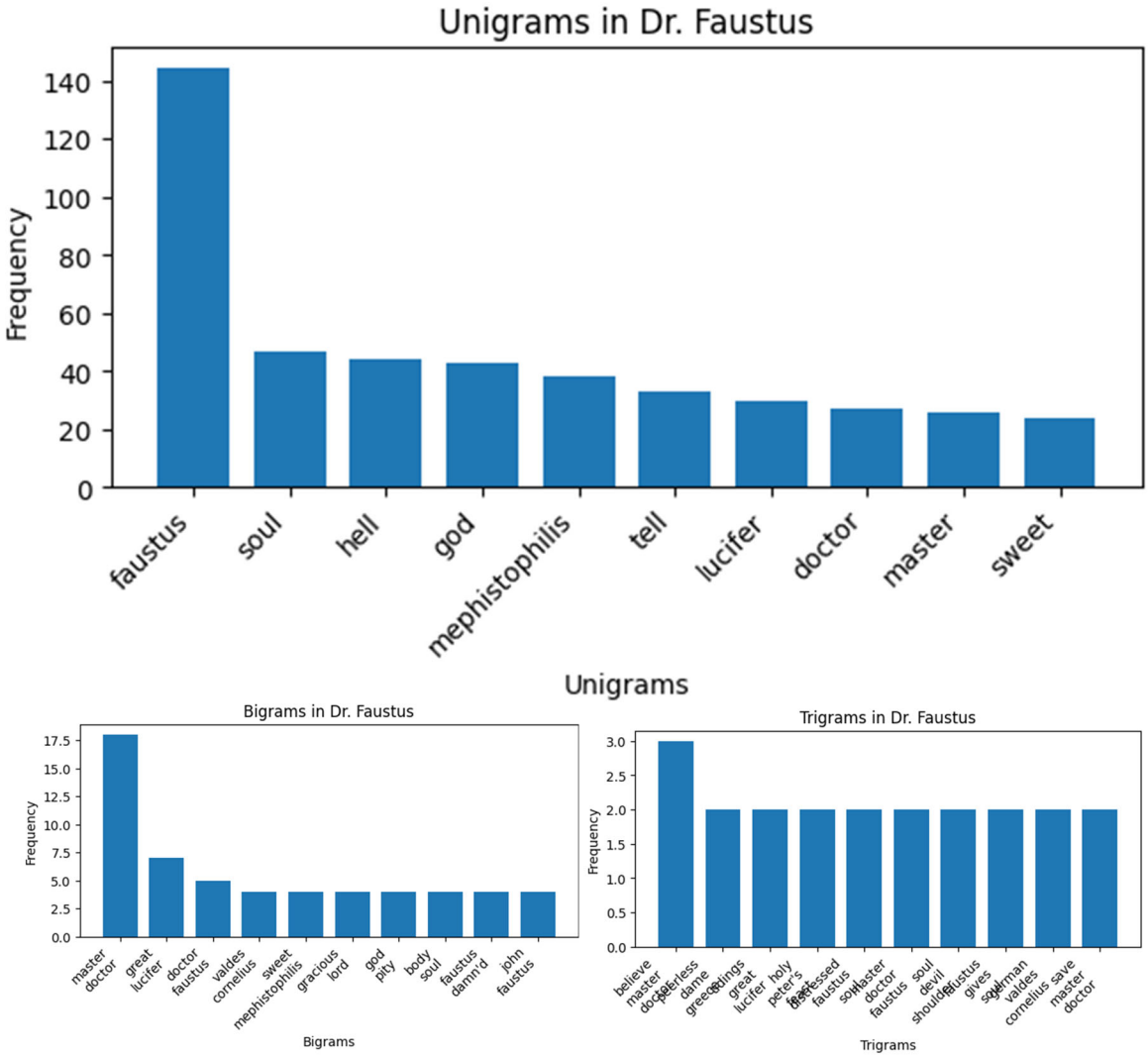
There has been some speculation that William Shakespeare either didn’t exist or was not as prolific a playwright as history would lead us to believe. One alternative scrivener that has been proposed as a potential alternative identity for Shakespeare was Christopher Marlowe. Christopher Marlowe is famously identified with “The Tragedy of Doctor Faustus,” (hereafter referred to as “Dr. Faustus” a play revolving around one man’s deal with a devil and the resulting sin associated with it.

CS 6140 Project Final Report

Jim Morgan, Foad Namjoo, Ariel Oconitrillo

N-gram analysis of Dr. Faustus showed that there is already a great deal of difference between Marlowe's and Shakespeare's styles. Whereas one of Shakespeare's title characters didn't even make the top ten of his most frequently used words, Doctor Faustus dominates. Indeed, the good doctor speaks to himself in the third person frequently enough to warrant making his own name a stop-word.

Like Shakespeare, Marlowe's n-grams indicate a heavy reliance on themes, though the bi- and trigrams indicate that most of these themes were very character-centric and displays a lack of the panache that Shakespeare demonstrates.



CS 6140 Project Final Report

Jim Morgan, Foad Namjoo, Ariel Oconitrillo

Conclusion

Shakespeare was a prolific author and poet, and any rigorous analysis of his works would require far more tools and techniques than can be supplied in this paper. We were able to draw up comparisons between Shakespeare and other authors, both contemporary to his time and ours. Jaccard similarities showed a possible decrease over time in Shakespearean prose, though this could as likely be ascribed to a shift in diction and grammar as a loss of interest in emulating the Bard. Future studies would look at themes explored in Shakespeare's works, such as death and mourning, as well as comedy and farce. Next steps could include labeling some of the words used and associating them with literary themes. As more language modeling tools become available, it would be interesting to see future directions analyzing these works in ways that exceed lexicographical. Until then, closing in the words of Shakespeare "Good night, good night! Parting is such sweet sorrow, that I shall say good night till it be morrow."

---

[i] This list was found online at https://lmackerman.com/AdventuresInR/docs/shakespeare.nb.html.