Abby Rooney
Wayne Lee
Applied Data Mining
7 May 2022

Using Text Analysis to Explore Questions of Genre and Gender in Shakespeare's Plays

Introduction

If you attend school in the United States, chances are you have been assigned to read a Shakespeare play for a high school English class or college distributional requirement. While Shakespeare's dense early modern English verse may not be every student's favorite reading material, his plays are widely taught not only because they are masterful works of literature, but also because they are deeply woven into the fabric of our culture. You may not realize that Shakespeare invented countless now-commonplace phrases and cliches, including "break the ice," "method in the madness," "all of a sudden," "what dreams are made of," "come full circle," "too much of a good thing," and "love is blind," just to name a few. You may not realize that some of your favorite movies and TV shows riff on Shakespeare's plots (did you know *The Lion King* is based on *Hamlet*?). In short, Shakespeare is all around us, and his works continue to shape how we think and communicate over 400 years after they were first published. That is why, much to many students' chagrin, these plays remain an ever-relevant object of study.

While you have probably been asked to analyze Shakespeare's work in essays or exams for your English classes, there are many ways to gain insight into these plays that do not involve a ten-pages-double-spaced close reading. By using statistical analysis tools to mine the text of Shakespeare's plays, we can make discoveries about these texts that are not apparent without the help of an algorithm. We can also consider this text data alongside other "metadata" about the plays, from genres to line counts to gender classification, to take our analysis a step further.

In an education system in which English classes have long been structured around a familiar literary canon and the same litany of assignments, the goal of this project is to show English teachers and students that classic literary analysis may be complemented with statistical analysis. I hope they may cite the results of my analysis to bolster their arguments about these plays and inspire further academic queries into the aspects of these plays that textual analysis unlocks.

Questions of interest

A question of genre

I will use text analysis to investigate a couple of well-worn literary analysis questions from a fresh angle. One classic literary studies topic is generic classification. How do we assign plays to genres? Are dramatic genres really as distinct as we think they are, or is there more overlap and grey area between them than we often acknowledge? Can an algorithm group these plays into genres based on the words they use? I will use clustering analysis to explore these questions and examine how features of the text, such as word choice and word frequency, may contribute to

our sense that a play belongs in a particular genre. Or, perhaps I wil discover that these features do *not* corral plays into genres as easily as we might expect.

A question of gender

A separate, but in some ways related, question is: How does Shakespeare represent women? To tie that question to genre: Are women represented differently in plays of different genres? Many essays have been written on these topics, focusing on particular heroines like Rosalind in the rollicking comedy *As You Like It* and Cleopatra in the epic tragedy *Antony and Cleopatra*, but my analysis will delve into the text to discern trends that characterize how Shakespeare writes women, and perhaps if that representation varies across genre. Again, I will use clustering analysis to explore these questions, with the goal of adding a data-driven dimension to the already-existing scholarship on these topics.

Data overview

Text data

My main data set is the text of Shakespeare's plays, courtesy of the Massachusetts Institute of Technology and available at www.shakespeare.mit.edu. MIT published the first online edition of the complete works of Shakespeare, which you can read online, but it also offers a version of the text in a CSV file, with each row corresponding to a line in a play. These are the columns in the data set:

- **Dataline:** Keeps track of the row index.
- **Play:** Title of the play to which the line belongs.
- **Player:** The name of the character who says the line.
- **PlayerLine:** The line of text.
- **PlayerLineNumber:** Keeps track of the line's index relative to the lines spoken by that character (for example, a PlayerLineNumber of 2 means that this is the second line spoken by that character).
- **ActSceneLine:** The act, scene, and line numbers which show you where in the play you can find this line of text.

I will focus on the variables Play, Player, and PlayerLine, which are most suited to answering my questions of interest.

Metadata

To complement the text data, I used metadata that is available on GitHub courtesy of Allyson Turner, a University of Denver graduate student who works at the intersection of literature and data science. "Genre" and "Number_of_Speeches" particularly stood out to me as features that would help me consider how these plays are categorized and compare women's speech to men's speech. The features of the metadata are:

- **Play:** Title of a Shakespeare play.

- **Year.1:** Year that the play was published.
- **Year.2:** Alternate year that the play was published, either due to uncertainty about the first date or because the second publication included significant changes. Many plays have an NA value in this column.
- **Genre:** The genre to which the play belongs (Tragedy, Comedy, History, or Romance).
- **Number_of_Speeches:** The number of speeches in the play, defined as chunks of text with 20 or more lines.

Name and gender data

For a third data set, I obtained data from the University of California, Irvine Machine Learning Repository which attributes first names to genders, giving counts and probabilities that each name corresponds to "male" or "female" based on open-source government data from the US, UK, Canada, and Australia. I incorporated this data set to keep track of the genders of Shakespeare's characters. This data set is limited in that many early modern English names that Shakespeare used are no longer popular enough to be part of this recent data set. Another obstacle is that many of Shakespeare's characters are not referred to by their first names, but rather by titles such as "Duke" or functional occupations such as "Messenger." Despite these limitations, this data set will still help me consider some differences in how male and female characters speak in these plays. The columns are:

- **Name:** A first name.
- **Gender:** M or F.
- **Count:** Count of people who have that given name and gender.
- **Probability:** Probability that a person with that given name identifies with that gender.

Feature engineering

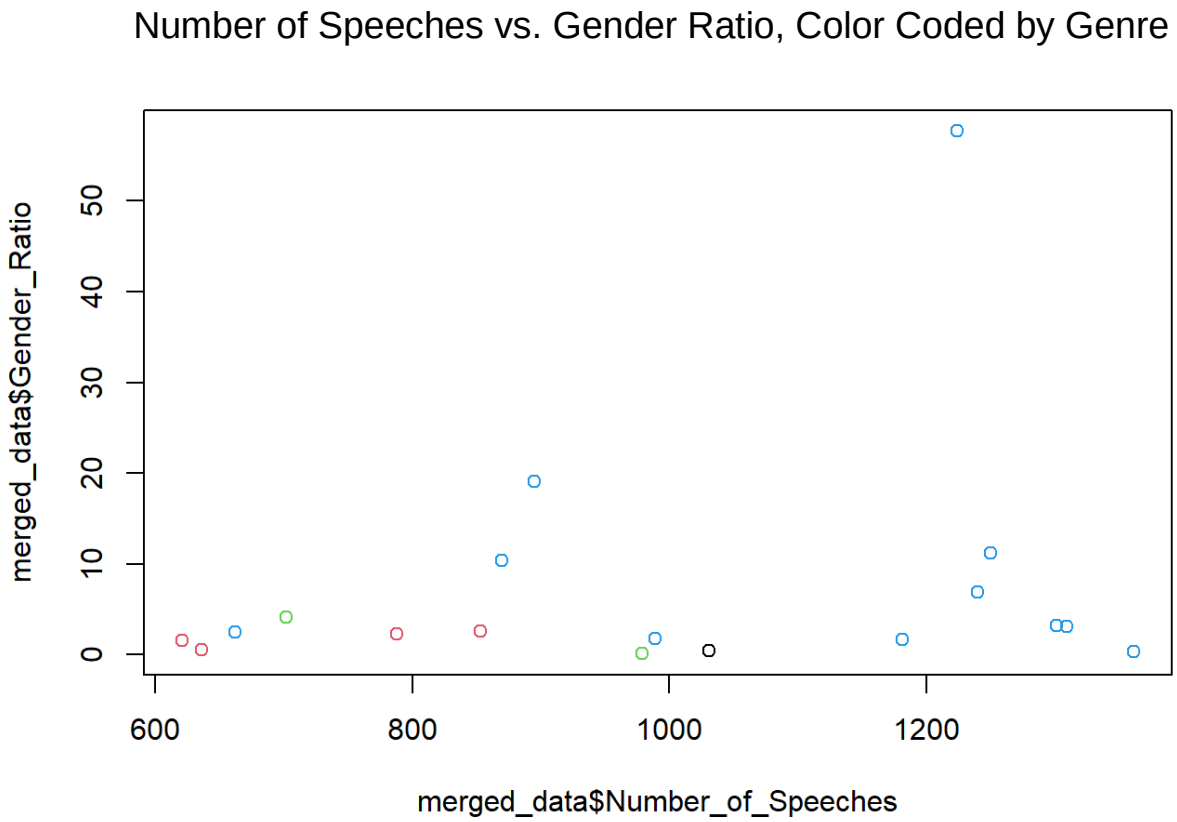
I engineered a feature called Gender_Ratio, which computes the ratio of men's lines to women's lines for each play (based on how genders were assigned to names by the above data set from University of California, Irvine). Again, there are limitations to this feature because many names used in these plays cannot be assigned a gender from the UCI data set, but this ratio helps me get a general sense of how often women and men speak in these plays. Please see the output of my code to see how I cleaned the name and gender data and constructed this feature.

Exploring and cleaning the data

Preliminary exploration of feature relationships

I did some preliminary data exploration to get a sense of how my features might be related to each other. For one example, I looked at the correlation plot (pictured below) between gender ratios and numbers of speeches and colored the plot by genre to see if certain genres tended to have a higher gender ratio and a higher speech count. The plot suggests that men speak much more often in tragedies (represented by the blue dots), which also tend to have more speeches. Comedies, romances, and histories have lower gender ratios, which means that they are more balanced between men's and women's speech (and sometimes, women even speak more often

than men!). This plot suggests to me that there may be a relationship between a play's genre and how that play represents women, and I will investigate that idea more thoroughly with my text analysis.



Data quality issues

I noticed some data quality issues during my exploration: There is not genre information, nor is there information about the characters' genders, available for every play in Shakespeare's oeuvre. So, when I do analysis that limits my data set to plays that provide this information, I cut some plays out of my data set, which makes my conclusions less generalizable to his body of work. Still, I'm working with tens of thousands of lines of text, so I have a lot of data—it is just important to note that a few plays are left out of the mix.

Text cleaning process

To prepare the text data for clustering analysis, there are a few cleaning steps I needed to take:

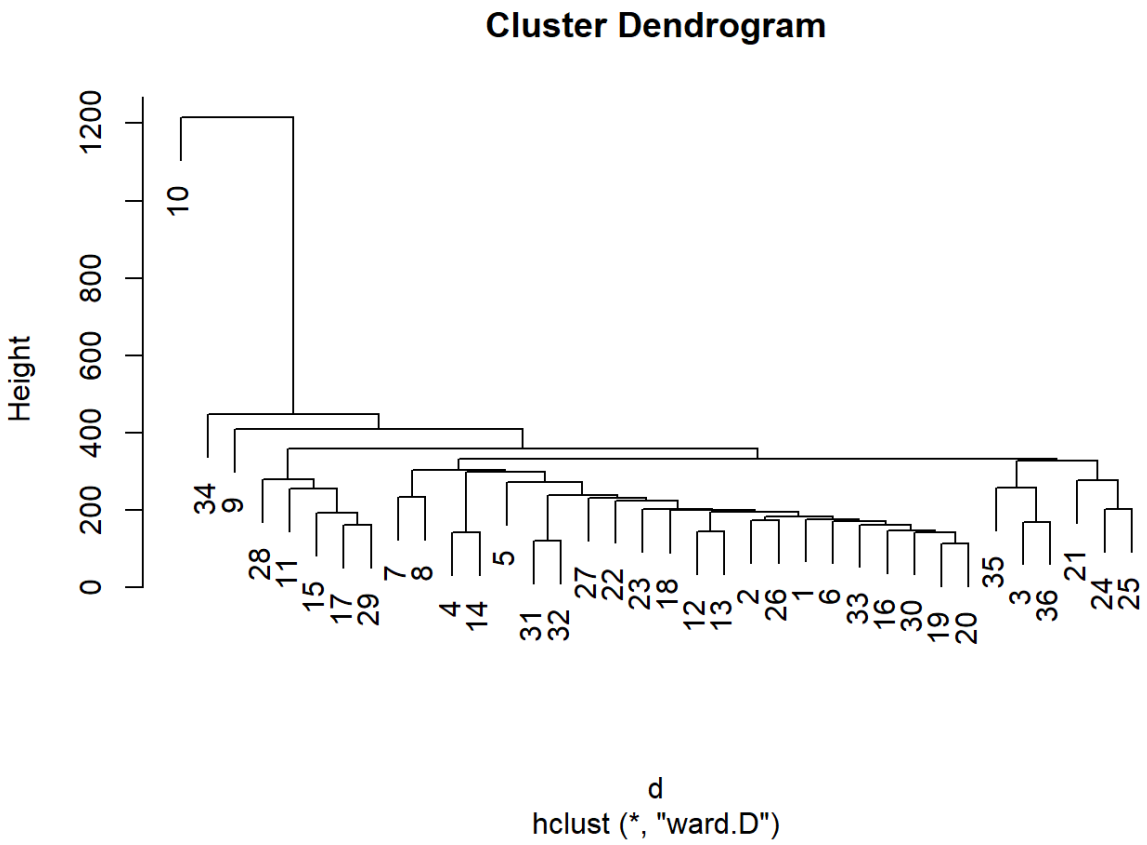
- Remove act and scene markers (ex: "Act I," "Scene IV.") so that they are not counted as meaningful words.
- Make text lowercase so that words are not taken to be different words simply because of a different format (one is uppercase and one is lowercase, for example).
- Remove punctuation and numerals, by the same logic.

Clustering

General analysis leading to genre analysis

First, I ran a couple of clustering analyses on the entire data set to see how successful it might be at meaningfully grouping plays. I created a TF-IDF matrix to identify meaningful tokens across all these plays, taking care to account for stop words (I used regular English stop words as well as a set of early modern English stop words, courtesy of http://earlymodernconversions.com/wp-content/uploads/2013/12/stopwords.txt; I also added character names and stage directions like "enter" and "exit" to my stop words list).

Using a data frame of meaningful tokens (stop words removed, high TF-IDF value), I tried to cluster the plays. K-means was unsuccessful, but the hierarchical method, though certainly not perfect, did a little better.



At first glance, this tree seems to show that clustering was unsuccessful—the tail on the right side of the page is not what we like to see. However, there are a couple of genuine clusters in there, and when I used the metadata to discern which plays belonged to the clusters, I was surprised by what I found. The six plays that are tightly clustered on the right side of the

dendogram are *Twelfth Night*, *A Winter's Tale*, *Two Gentlemen of Verona*, *Measure for Measure*, *Much Ado About Nothing*, and *Othello*. All of these are comedies of a similar bent, with the exception of *Othello*, which suggests that there is something about the words that Shakespeare chooses for his comedies that make them identifiable as a group.

There are other mini-clusters of interest, such as the clustering of play 2 with play 26 toward the center-right of the diagram. Those plays are *A Midsummer Night's Dream* and *Pericles*, which are both a different brand of comedy, heavily focused on romance and magic. It is telling that the clustering algorithm picked up on their similarities. Even play 10, all alone on the left of the diagram, is interesting—it is *Comedy of Errors*, which funnily enough is not a true comedy, but rather a tragicomedy. Its mix of comedic and tragic elements makes it difficult to categorize, which this clustering output reflects. Even though this dendrogram does not look particularly successful, it does in fact identify characteristics of these plays that contribute to genre classification.

Subsetting by gender ratio, then clustering

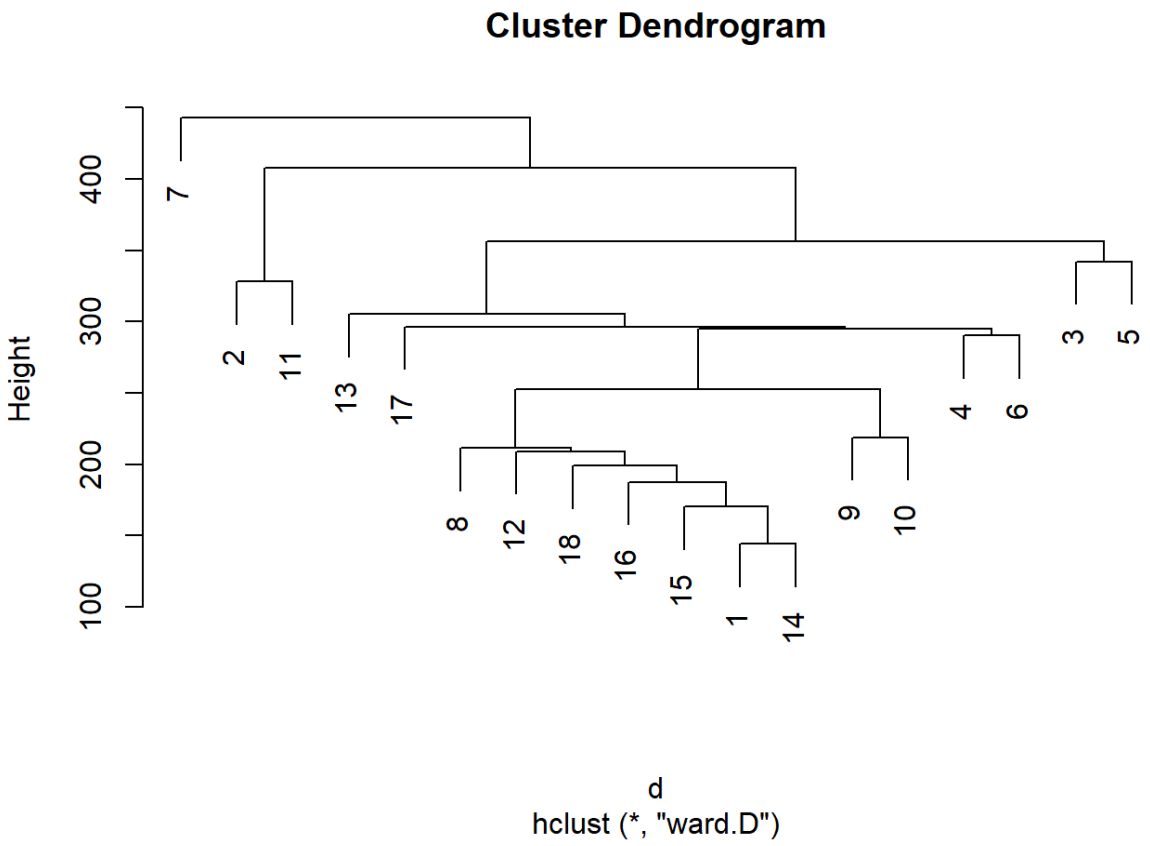
Since my clustering algorithm had some success clustering plays into genres, I then decided to explore the relationship between genre characteristics and gender representation. I proceeded to split my data into two sets: One set in which the plays have higher ratios of men's lines to women's lines, and one set in which the plays have lower ratios of men's lines to women's lines (which indicates that women speak more). Then, I tried clustering the plays and looking at the top words in each cluster.

I found that the plays became a little bit easier to cluster when they were split into these gendered groups, especially within the group of plays with a high gender ratio (more men speaking). Looking at the top words in the main clusters of this group, they seem to reflect particular themes: worldly affairs of the state, with country names like France, Denmark, and Cyprus popping up frequently; classical epic themes, with words relating to Greece, Troy, and gods dominating the list; and themes of violence and warfare, with the word rape being the most prominent. Please see my code for the full output.

While I set out to learn something about how Shakespeare writes women, I learned more about how he writes men—the main themes that men seem constrained to talk and think about in this group of plays. The clustering exercise was less successful on the plays in which women spoke more, suggesting a difference between men and women's speech. Perhaps women do not have such a clearly defined set of themes to speak on, so it is more difficult to cluster plays in which they speak more.

Below is the Cluster dendrogram for high gender ratio, which is not perfect but satisfactory at grouping pairs of plays here and there by theme.

(See next page)



Verifying my results

The patterns that I have uncovered, especially with regard to the gender-split data, are not especially strong, and I think that may have something to do with the gender-split data only analyzing about 19 plays rather than the full 36. If I can find gender and genre data somewhere else—or perhaps produce it myself—that will allow me to use all of Shakespeare's work in my analysis, my results would likely be more robust.

There is also a good deal of uncertainty involved in how text data is processed. While I did my best to process it such that stop words are filtered out and every word is appropriately lemmatized, there are hundreds of thousands of words in these plays, and even just looking at some of my code output I can tell that I did not catch every issue that prevents a word from being appropriately lemmatized. These are just a couple of ways in which I could improve on this project in the future.

Conclusion

My text analysis suggests that the boundaries of genre—much as I personally despise them—do hold up on a textual level, as there are characteristics of these plays inherent to the words that

Shakespeare chose for them that contribute to their gender classification. My analysis also suggests that women's and men's speech is characterized differently in these plays, with different topics dominating men's discourse. These insights, and more insights like them, can be useful for literary scholars at all stages of inquiry to support or refute arguments about genre and gender in these plays and hopefully spark new ideas about how we can use algorithms to shed new light on familiar plays.

It is important, as always, to be wary of data snooping in projects like these. In my project, data snooping might entail looking at the text data and trying to subset it in particular ways so that the plays cluster how I want them to. It could also involve removing tokens that seem out of place or like they are interfering with my idea of what the top tokens should be showing. With text data, there are many text processing choices you can make that will affect the results, so someone who wanted to data snoop could experiment with these choices until the results show what they want to see. Of course, that is not my goal, and I find it much more exciting that the data revealed these patterns on its own.

Critique: "Exploring *Star Wars: The Clone Wars* episode scripts" by Cyrus Jackson III

- What was the initial motivation for tackling the project?

As a fan of *Star Wars: The Clone Wars*, Cyrus' motivation is to understand how an episode's script is predictive of its rating. Through analyzing the text of the episodes' scripts, Cyrus wants to gain insight into which particular characters, settings, and themes—as discerned from the text—are most preferred by viewers as reflected by their IMBD ratings.

- What datasets were used?

Cyrus used two datasets: IMDB ratings of *Star Wars: The Clone Wars* episodes scraped from the IMBD website, and a database of episode scripts scraped from a fan site.

- What aspect of the project is considered a data mining and what is discovered?

This project is considered data mining because Cyrus is mining the text for patterns, looking at how words are grouped together how those groupings might be tied to positive or negative ratings. He discovered that characters associated with action words appeared to garner more positive ratings. The value that analysts and writers may take away from this project is that prioritizing such characters when writing *Star Wars* projects, focusing on action language, could be beneficial for ratings.

- Is there anything you would have done differently?

One thing that I might have done is consider scripts for *Star Wars: The Clone Wars* within the larger corpus of scripts for all *Star Wars* projects to get an even stronger sense of which settings, characters, themes, and other story aspects that we can mine from the text are important in the

grand scheme of the *Star Wars* universe. There are ratings available for those projects, too, so it might be useful to see how those texts map onto their ratings to get a better understanding of how reliable this analysis is and confirm if this finding—that action characters garner better ratings—is reproducible in other contexts.