Do William Shakespeare's men and women employ significantly different diction?

Simran Vatsa, 804644668, ENGL 150, Prof. Robert Watson

3/25/2018

Introduction:

I've been quite interested in the potential of text data-mining for about three months now, so I've been reading articles, blog posts and studies that explore or demonstrate it. I've gathered that it gives one the ability to distant-read a corpus in ways human reading and language processing cannot. Humans are best at processing small amounts of information at a time; but for machines, the larger the corpus, the more likely it is that useful, interesting insights will be produced. With this in mind, I kept my research question quite broad: *Do William Shakespeare's men and women employ significantly different diction?*

Technologically (and theoretically), I'm very far from an expert on natural language processing, so this project is far less incisive that it could have been in someone else's hands. For instance, I've read papers on machines predicting gender from dialogue. I admittedly do not have the knowledge to code something like that (I hope to soon though!), so I wanted to use this project to solidify and apply some basic techniques I've read about.

The cleaning process:

I attempted to download the plays from Gutenberg, but they were quite messy; while tokenizing and tidying was not difficult, there was no convenient way to assign speaker names and play line numbers to each line or word. I was thus happy to find a dataset of Shakespeare's plays on Kaggle, with the following variables:

Dataline: row number in dataset

Play: name of play

PlayerLinenumber: represents the nth change in speaker

ActSceneLine: the play's act, scene and line number, as seen in physical copies

Player: character speaking the corresponding line of speech

PlayerLine: text of the play, line by line

The dataset first contained all Shakespeare's known works. I filtered it to include just six of the eight plays we read in class. I then tokenized the variable PlayerLine into single words, labelling the variable "word." Tokenizing converts data to tidytext format, the most convenient way to analyze text in R.

At first, I included "Henry IV, Part One" and "Henry V" but these skewed my analysis due to their numerous auxillary characters - the plays contained around 70 characters combined - and lack of women. Removing them from the analysis, I was left with 166 characters. I exported these characters to an Excel spreadsheet and manually coded them each 0 or 1, representing female and male respectively, reimported the sheet, and joined this dataset with my original one, "tidy_shakes."

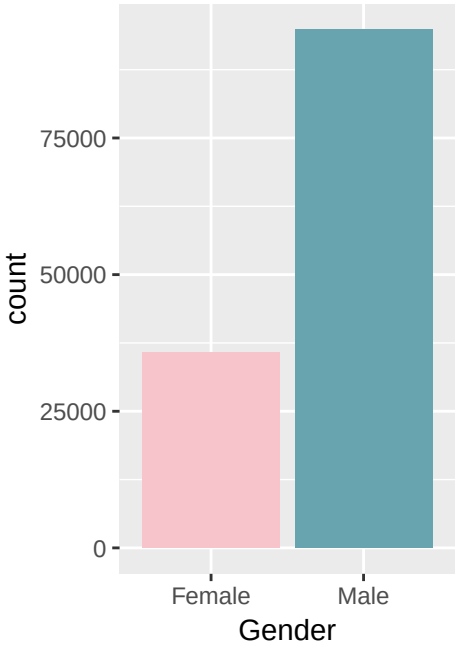
How often do men and women each speak?

I now had a gender assigned to every character, and this showed I had 26 female characters and 134 male. Plotting the number of words said by men and women in the plays shows men speak 2.80 times more words than women in total. By including "Henry IV, Part One" and "Henry V" in the dataset, this number goes up to 3.96, showing the extent to which the two plays would affect the answer to my thesis question. Women

were never at the forefront of either diplomatic strategizing or warfare - thus, Shakespeare's dramatized histories are limited, in ways out of Shakespeare's hands, in how they portray women.

```
## Number of characters in the six plays: 166
```

```
## Number of each gender:
## Female: 26 Male: 134
```



Next, I removed stop words from the dataset, adding common archaic words to the modern English corpus of stop words after some exploration of the data. The ratio of men's words to women's remained around the same, 2.76. Dividing the number of significant words said by number of characters for each gender showed each woman was given a good deal more significant words on average. This is a result of most auxillary characters being male; if a woman features, she is usually important to the plot, or important to a woman who is important to the plot: Juliet's nurse, Celia in "As You Like It" and Nerissa in "The Merchant of Venice" are all examples of the latter. When coding gender to character names, written in order of appearance, it was interesting to note that very often, when one woman appeared, another often followed immediately. This supported the idea of breaking away from comfortable same-sex friendships into the world of romance, sex and the other gender, as was discussed in class - women are often cocooned together in their first appearances.

Gender	n()
Female	12090
Male	33424

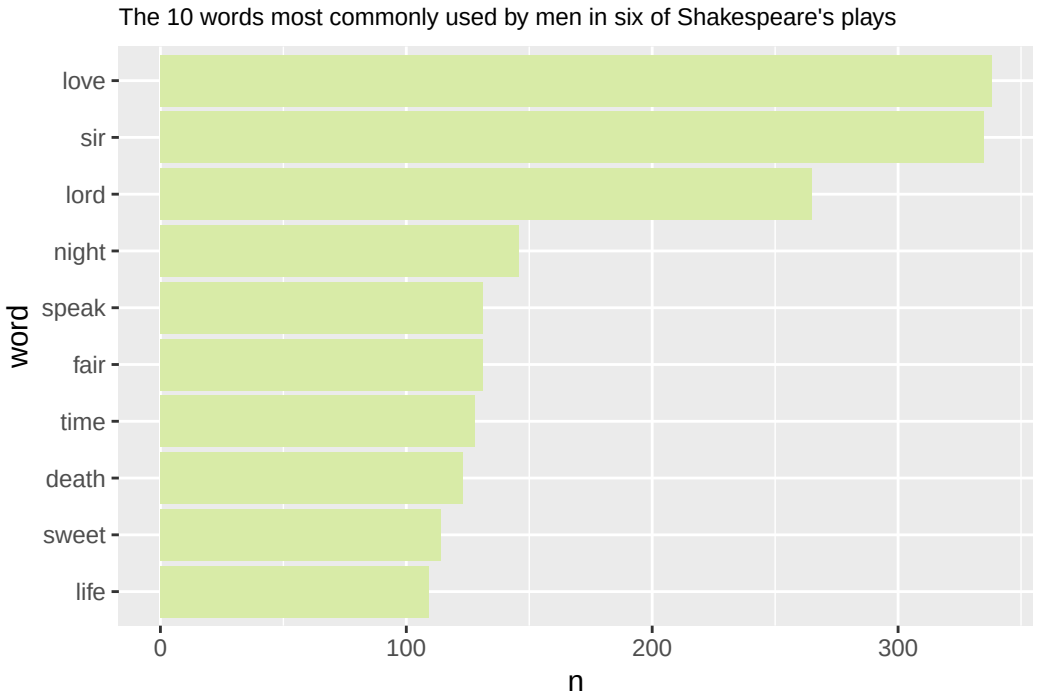
```
## Average number of significant words a character says:
## Female: 465 Male: 249.4328
```

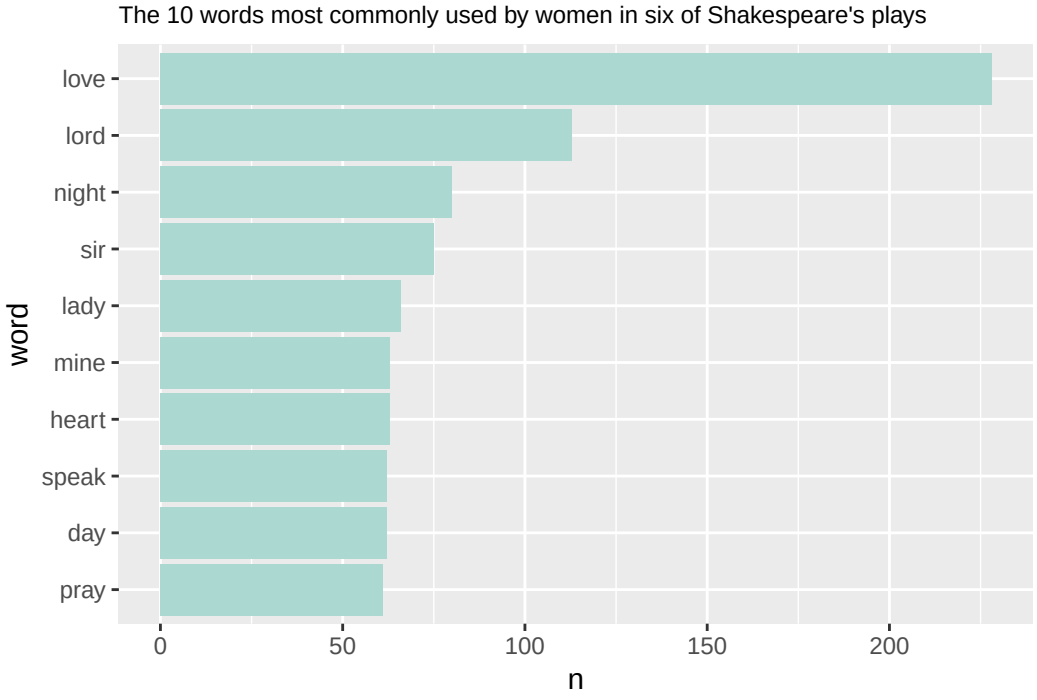
What words do men and women say most?

Graphed below are the 10 words most commonly used by men and women. The first four are the same; thus they serve more as testament to Shakespeare's voice and preoccupations than any indication the genders have different vocabularies. Often, in sorting by frequency with text data, the first few-ranking results are too ubiquitous to be significant, and this is the case here too. Of the final six results on either graph, however, five are unique to each gender and display rather large differences: men are concerned with life and death, and women more so with heart and possession. This could be seen as illustrative of how, when women have a stake in plot, it is mostly romantic.

Women also say the word “lady” more relatively frequently, which was unexpected, considering men say a good deal about women. This is probably another consequence of the presence of the Nerissa-Maria-Nurse character, who constantly addresses her mistress as “lady” - in an analysis of gendered words in Shakespeare, which did not yield particularly interesting insights, it turned out “mistress” was a fairly common word for a woman to refer to a woman by.

A note on the word “pray” - not in its modern context but in its somewhat archaic meaning of “entreating,” often in the phrase “Pray, tell.” While a common phrase, it is interesting it cracks the top 10 with women; perhaps it was a sign of feminine courtesy.





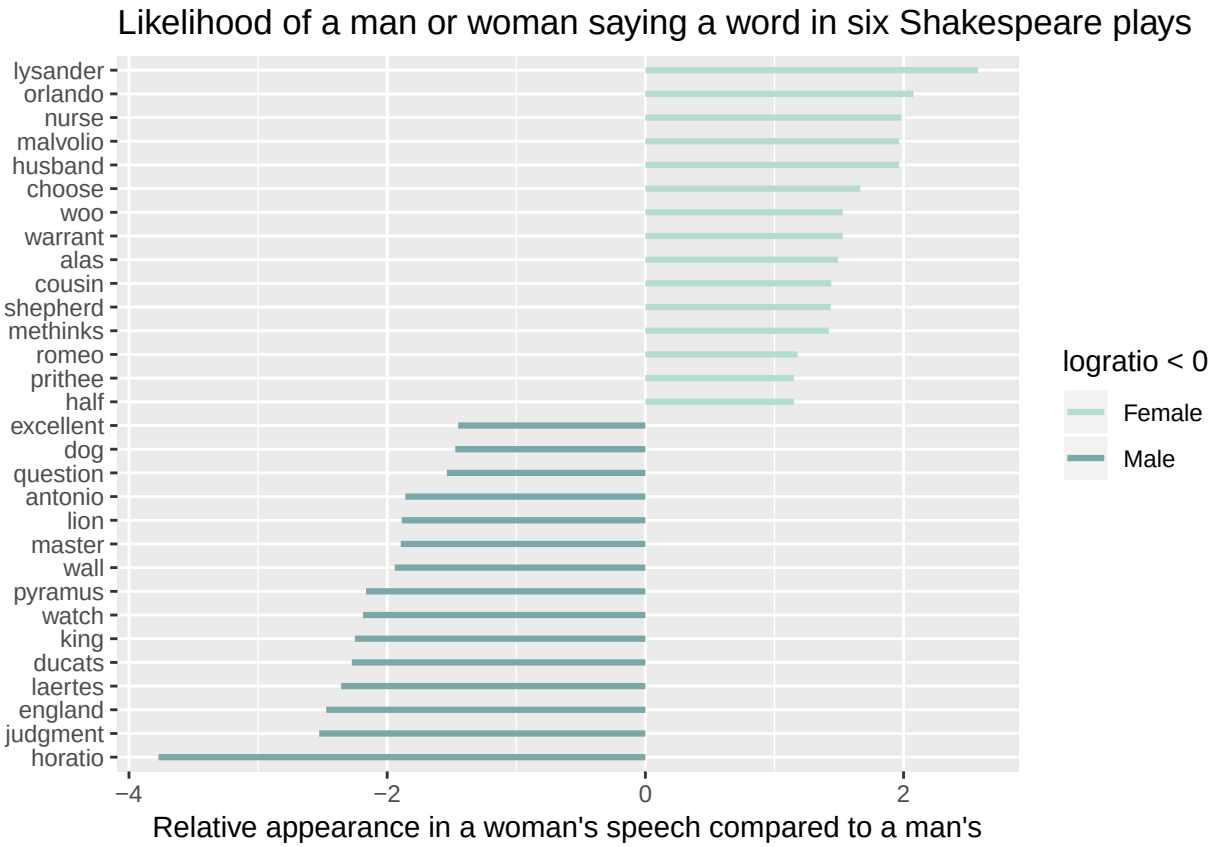
What words are more likely to be said by a man or a woman?

I read a wonderful analysis of "he" and "she" bigrams in Jane Austen novels at https://juliasilge.com/blog/gender-pronouns/, and wanted to incorporate her methodology into this project. I adapted Dr. Silge's code for speech patterns in Shakespeare, generating a table of 10 words almost equally likely to be said by either gender and a graph of the 30 words most likely to be said by one gender or the other, 15 for each. This segment is a step toward predicting speaker gender through dialogue; a lot more linguistic processing would necessary though.

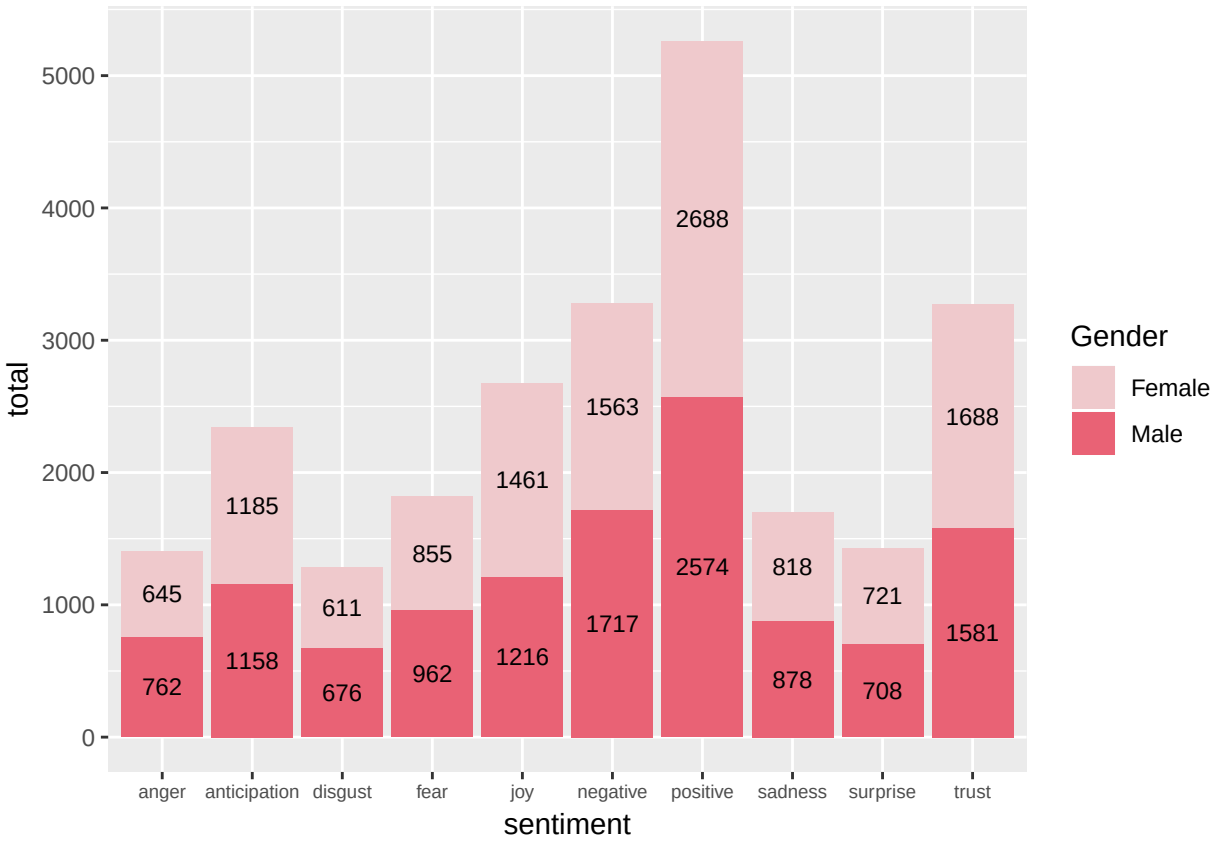
Both the graph and the table can be greatly modified by changing the lower bound for a word's occurrence. I have set it at 25; at 10, most words were character names.

word	Female	Male
breath	0.0022442	0.0022350
lord	0.0213204	0.0212324
mind	0.0033664	0.0033525
wit	0.0039274	0.0039112
drink	0.0020572	0.0020754
lay	0.0020572	0.0020754
word	0.0050496	0.0051086
seek	0.0024313	0.0023946
air	0.0018702	0.0018359
pale	0.0018702	0.0018359

The table shows interesting, if not cohesive results; for instance, "wit" is equally likely to be said by either gender, possibly because Rosalind says the word often, and she and Celia refer to Touchstone as "wit" - he accepts the title and often discourses on wits and fools. Similar analyses could possibly be carried out with other words, and this technique could be expanded into a standalone project as Dr. Silge did in her post.



The graphs still contain a fair number of character names; Lysander, Orlando, and amusingly, Malvolio, appear popular with the women. Rosalind and Celia discuss Orlando by name far more than any man in the play does. Lysander declares his love to two women, both of whom refer to him by name quite often. While Malvolio is far from the object of anyone's desire, Olivia calls him by name quite often, and his interactions with male characters are limited to being ridiculed by Sir Andrew and Sir Toby. Romeo also makes this list - Romeo was actually the 12th or 13th most spoken word by women in total, possibly because Juliet is in the habit of saying his name rather often ("O Romeo, Romeo! Wherefore art thou Romeo?", Romeo and Juliet, II.ii.33). As for Horatio, who tops the list for men, he speaks to and is addressed by no one besides Hamlet; therefore, it's very unlikely dialogue from a woman will contain his name. And Antonio, in both cases, is not interested in women, so it is unlikely he will be addressed or referred to by one. The fact that women are more likely to say "woo" strengthens the argument that their involvement in plot is often only romantically, or toward a romantic end. Men are more likely to say more "serious" words, such as "king," "master," "judgement" and "question."



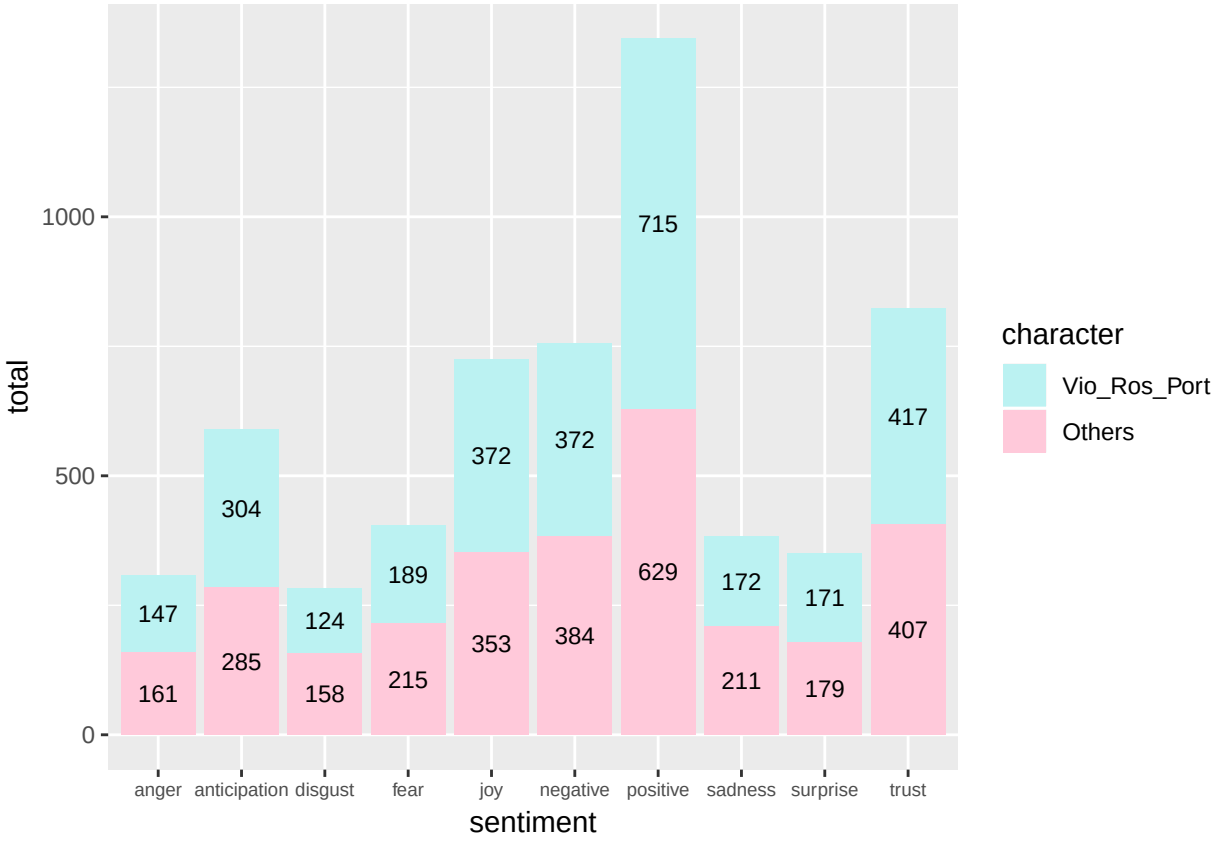
Sentiment analysis: what do men and women feel?

The graph above uses a sentiment lexicon to partition words into eight emotion categories. The lexicon is modern English, so some words' connotations are very different from they were in Shakespeare's time. The graph uses word values that have been proportioned per the number of words analyzed for both men and women. Interesting insights that came to light:

- Women say less negative and more positive things than men. (This may or may not be entirely Hamlet's fault.)
- Women express far less anger and fear than men. The first is not surprising - it did not become a woman to express her anger - but the latter reveals an interesting lack of the "damsel-in-distress" trope in Shakespeare's work. Since we only read one tragedy that did not feature a woman in a lead role, and

most comedies display a somewhat equitable distribution of dialogue between men and women, this cannot entirely be a result of men on average being in more perilous situations than women.

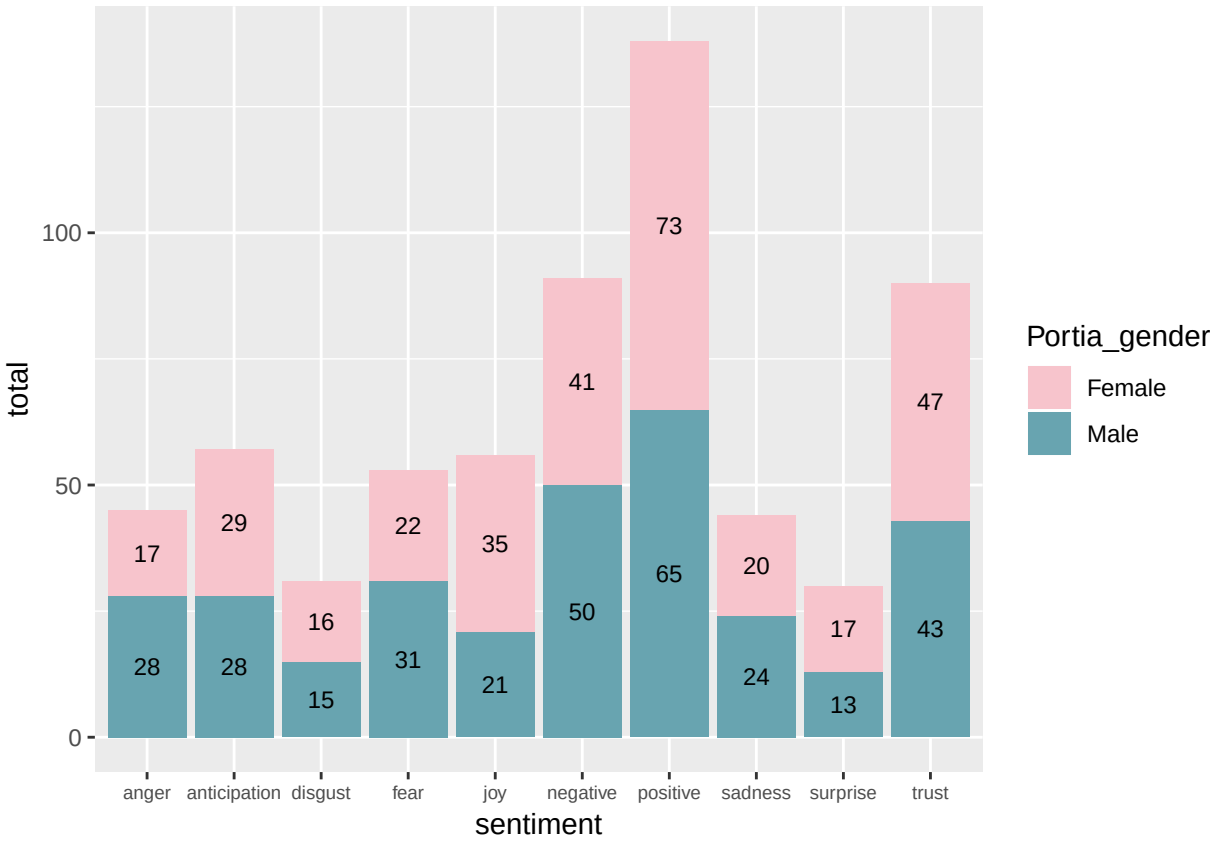
- Women use words that connote joy and trust far more than men, perhaps again a result of their involvement almost purely in love plots. However, they use similar numbers of words for each negative emotion (besides "negative" itself), meaning women do not show a lack of negative emotional depth, even in their love plots; Viola silently yearns, Rosalind expresses despair and Juliet - well, Juliet is in a constant state of anguish and worry.



Do Rosalind, Viola and Portia, disguised as men, show different speech patterns from other women?

To answer this question, I split the female-character corpus into two; one contains all of Rosalind and Viola's words, as well as Portia's when she is disguised as a man. The other contains the speech of all other female characters. I ran sentiment analysis on the two corpuses and made their results proportional to each other. The results are recorded in the graph; much as I dislike gendered color coding, I've done that here to make it less confusing.

This is quite a specific question, and thins the corpus quite significantly, so I was aware it might not lead to a discovery. However, it did, in that showed the opposite of what might have been expected given the male/female trend. These three women are more positive than the others, expressing more positive emotion and less negative emotion. However, this is likely because they are in comedies, and play very self-assured men always ready with a quip or an argument. They show very little disgust and fear, and, despite Rosalind and Viola's feminine yearnings for Orlando and Orsino being factored in - it was quite impossible to split the corpus in their cases, as they continuously alternate between playing men and women - they show significantly less sadness than other women.



Does Portia speak differently / use different words when she's playing a man?

Since Portia's time as a man is more easily isolated, I decided to focus on her for a bit, at the risk of thinning the corpus into oblivion. I plotted her sentiments as a man against hers as a woman, and found, this time, that her language was significantly different for the two cases, better matching the male/female trend than the previous analysis did. This shows that Rosalind and Viola's intermittent female language interfered with sentiment plotting, skewing it in an unprecedented direction.

The tables below depict clearly the situational differences Portia the woman and Portia the man find themselves in, which is in part why the graph that follows will likely show significant sentiment differences for the two.

word	n
jew	9
bond	8
flesh	8
justice	8
mercy	8

word	n
nerissa	16
choose	14
lord	13
love	13
ring	12

Conclusion:

The question I asked was very broad, so the conclusions that can be reached are similarly so.

- Firstly, and unsurprisingly, men spoke a great deal more than women, but each woman character spoke a larger share of significant words.
- A recurring conclusion seen in each analysis was women's greater use of and preoccupation with love language than men. From "woo" and "husband" appearing in women's likelihood graphs, with "wife" absent from the men's but "king" and "master" being present, it also became clear that both men and women talk more about men than women. In fact, it's possible some of these plays do not pass the Bechdel Test - more advanced NLP could probably devise a way to test this out!
- The sentiment analysis is certainly colored by the fact that language has changed greatly since Shakespeare's time, but the male/female trends it depicts are still valid, since each gender's words would be equally affected.

The two case studies, Portia's in particular, served to back up the male/female sentiment analysis in different ways. The analysis of Viola, Portia, and Rosalind showed that Viola and Rosalind's fluid genders flawed their categorization as "men" and led to graphs that went against the original trend. Portia's analysis, however, supported the original trend observed, since her character's lines were easily bisected into male and female. This essay serves as an exploratory starting point for more exhaustive and intensive quantitative analysis of gender in Shakespeare. While Shakespeare plays with the fluidity of gender in his plays, it remains that hidden beneath some overarching similarities (the first four words in the frequency analysis) in the language he uses, there are important differences in the words he gives men and women.