Portfolio 3
Stat 479
Bella Wu

I personally enjoy reading books written by Jane Austen, and in this project, I am going to explore the relationships between vocabulary in Austen's book to draw interesting findings using a correlation network graph in a shiny app. In terms of relationships between words, I am interested in if there are any words tend to follow others immediately and why this kind of pattern occurs. To achieve the above objective, I used a package called 'janeaustenr' which contains the text of 6 different novels written by Jane Austen, including *Sense and Sensibility, Pride and Prejudice, Mansfield Park, Emma, Northanger Abbey, and Persuasion.* I tokenized the novel into pairs of adjacent words rather than divide the novel into individual words to explore the relationship between pairs of adjacent words. This technique is similar to bigram in the language model, but instead of calculating the conditional probability in bigram, I generated pairwise correlation of word pairs.

In the shiny app I developed, users are able to choose one of novels by Jane Austen to explore vocabulary relationships. On the right-hand side of the shiny app, I created a network graph of vocabulary to visualize the co-occurrence relationships between words. Each blue node represents each word in the novel of Jane Austen and the grey line connecting the nodes indicates that there is a certain degree of relationship between words. The darker the edge in the network graph demonstrate the strong relationship between words. Users are also allowed to select the minimum correlation between words to visualize. If a small minimum correlation is selected, the network graph might be denser with more nodes and edges and sometimes could be hard to visualize due to the high information density. In contrast, if a large minimum correlation is selected, there might not be any nodes or edges in the graph since few words have such a high correlation. Therefore, users may play with the minimum correlation to select the most appropriate correlation to visualize, and this value might vary depends on the specific novel.

There are many interesting finds in the shiny app. In the novel *Persuasion*, there are a lot of correlation between names of people and their social position, like "lady russell" and "wentworth captain". Correlation between family members also occurs since "mother", "father", "sister", and "brother" are connected in the graph. As a tradition in Jane Austen's novel, there is always party at night and as a result, "evening" is related to "party". We also could find a lot of correlation between time like ("half", "hour"), and ("twenty", "hours"). In *pride and prejudice*, we also see a lot of relationship between character names and their social position like ("catherine", "lady") and ("william", "sir"). Furthermore, words like "pounds", "thousand", and "ten" are also strong correlated, which is not surprising as the a lot of the characters in the novel emphasize the role of money. "Money" and "marry" also have a strong correlation, demonstrating those characters respect people with great wealth and desire to marry these kinds of people.

# Correlation of Vocabulary in Books Written by Jane Austen

**Select the minimum correlation between words to visualize**

0 0.14 0.4

0 0.04 0.08 0.12 0.16 0.2 0.24 0.28 0.32 0.36 0.4

**Select the name of the book**

Sense & Sensibility