1. Project Overview

In my project I wanted to find out more about books by famous authors, and which one had the most romance in it. I love romantic novels, so this will help me determine which book to read next!

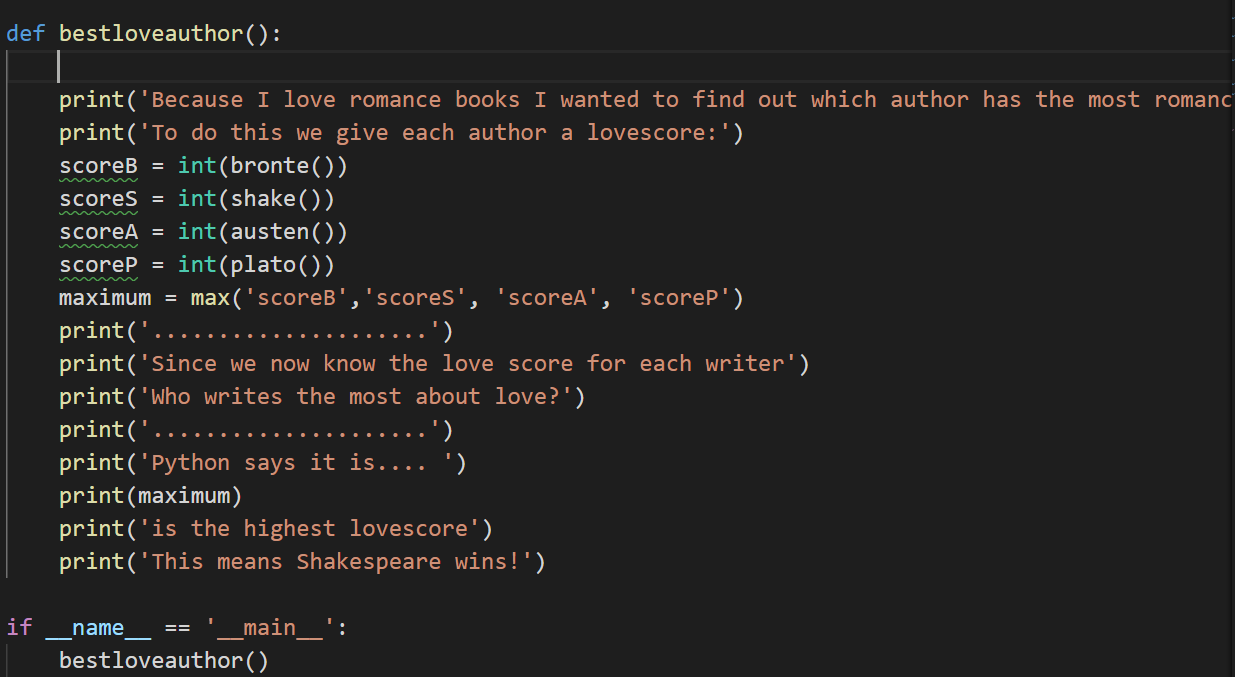
The data source that I chose for this assignment is the Guttenberg Project. I chose a few famous novels that were featured on the website. This includes 3 pieces by the Bronte sisters, 2 from Shakespeare, 1 from Jane Austen, and also 1 from Plato. I wanted to pick one author that did not write romance(Plato) to see if there was any relation between usage of love words and romantic nature of the book.

I wanted to use the nltk package and cosine similarity to see how similar each novel was to each. I struggled with this package, but I played around with it.

1. Implementation

Overall I tried to make the functions as simple as possible so I could reuse them when looking at a different text, but by just switching around the names. I designed it to first take a look at the authors where I could find more books, so I did Bronte first, then Shakespeare, then others with just 1 book. The first time I did the I realized the “lovescore” of the author was not taking into account that I was looking at three books from one author and only one from the other. This is why I went back in and added an extra function for Bronte and Shakespeare to calculate the average score of their books.

For each of the different authors I was able to determine the lovescore based on how many times the words in the love file appeared in their piece.

I then created a function to help calculate the highest lovescore by assigning each of these values to a new score variable: 

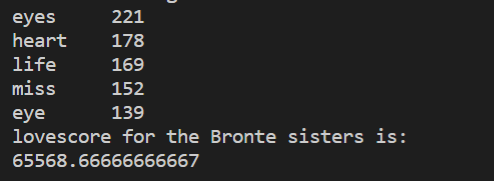
I chose to design this book by book, and then give an overall answer for which one is best, so people can see the breakdown. It would be nice to know how many lovewords, but it was nice to see the top 5 for each author because it gave you a flavor for the type of words this author is using.

\*I added the sentiment analysis after this portion of the project, but I struggled to get valid results from it. That is why it is not included in my results section.

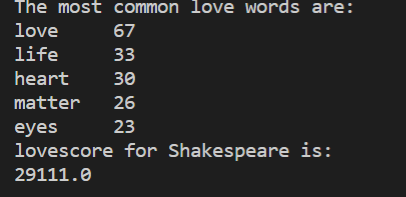
1. Results

I had difficulty in getting my sentiment analysis to work properly, that is why parts of it are # out.

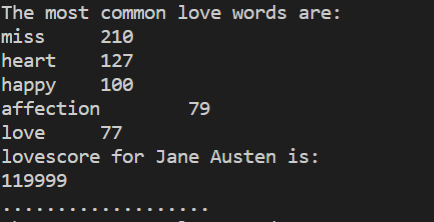
The results for the novels by the Bronte sisters are as follows:



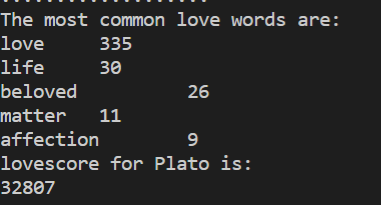
The results for Shakespeare’s pieces are as follows:



The results for Austen’s piece is as follows:

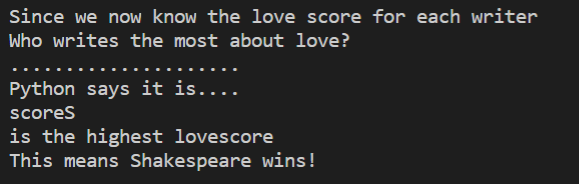


The results for Plato’s work is as follows:



I found it interesting that although Plato’s work is not considered a romantic piece, it still has a lot of words related to love! Had I not run this book, I would have never known! It might be on my next books to read list now because of this finding.

Then I tried to get Python to calculate which of these authors had the highest lovescore. This is what the results looked like in the terminal:



From a process point of view the lovescore went well and it gave me the results I wanted after I played around with the functions. I worked alone so I did not have to split up the tasks. As I mentioned earlier I struggled with the nltk package, so had we done more of that in class it would be helpful.