Project Write-up and Reflection for Mini Project 3 – Text Mining and Analysis

**Project Overview**

Starting this project, I didn't know for sure what I wanted to analyze. I decided to start with a book from project Gutenberg, in my case Wuthering Heights by Emily Bronte. My goal with this project was to better understand what kind of text mining was possible to do and how it would work. Therefore I talked to a few former students about their projects to get inspiration on what to create. I ended up creating a frequency counter that was giving me the number of words appearing in the text for all frequencies.

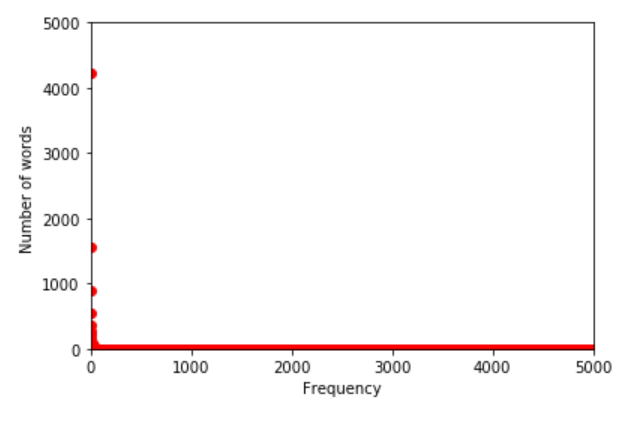
**Implementation**

I started by choosing the book I wanted to analyze. The language is a little bit old fashioned but still very similar to modern English. Since I wanted to analyze the words in some way, I started by creating a histogram of my text, to create a base. I created a function displaying the most common words and their frequency. The result was almost exclusively very common English words like 'the' and 'and', so I decided to create a new dictionary with the 50 most common words in English to subtract them from my dictionary. I got some more interesting words in my new histogram, but it wasn't interesting enough.

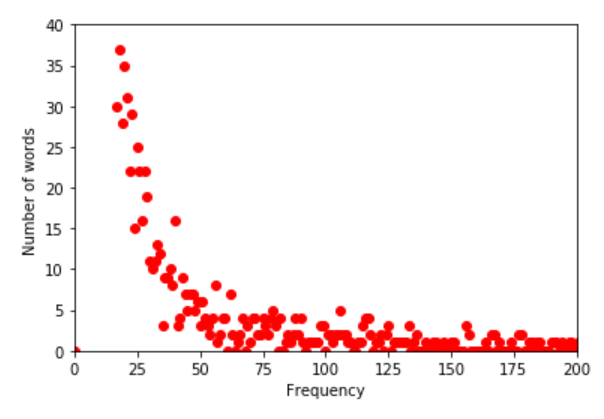
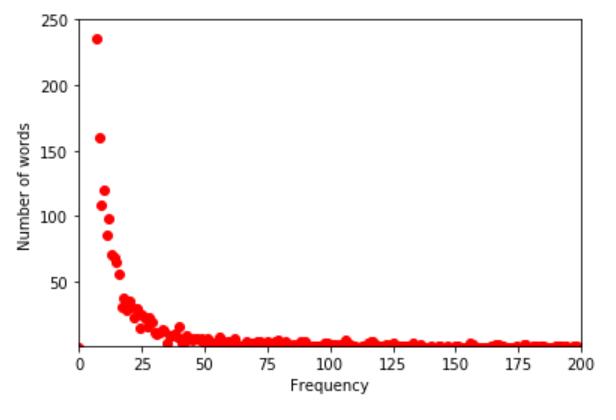
Therefore I tried a new direction, to look at the least common words. When I printed them, I saw that there were far more than 50 words appearing once, so I thought it might be interesting to see how many words only appeared once in the text. The result were 4217 words only appearing once, which surprised me. Then I came up with the thought of exploring how many words appeared a certain time. Then I made it more general and created a function that output how many words appear for all frequencies (setting the interval based on the frequency of my most common words). To get a good overview of the data I gathered, I tried to plot it. I chose to do that in matplotlib which was an easy way to plot a simple graph with many inputs.

**Results**

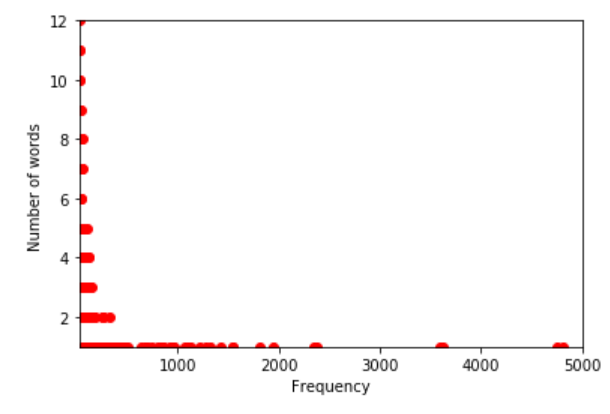
I have chosen to present the data I gathered though plotted graphs. As you can see, there are a lot of data in the graph, therefor I have done some scaling of the x and y axis to display different things. The first graph just underneath is an overview of my data, you can see that there are multiple words in the lower frequencies that are decreasing.



To show that further, I created these two zoom-ins. You can see that there are a lot of words which are printed 1-25 times. It is decreasing more after 50 times.



This is a graph to show the high frequency words. There are a lot of words in the higher frequencies that don’t have any number of words. But there are a few words appearing ~4800 times or ~3700 times.



From this you can make the conclusion of Emily Bronte writing with a wide vocabulary and are using multiple different words. This data is a little bit misleading since if you print the words mentioned once, there are a few words that are just symbols. This is something I could try to figure out how to eliminate for the next project.

**Reflection**

The things that went well with my project are that I reached my goal of learning about text mining and got good insights from talking to others about their projects, both students in this class and former students.

I think my project was appropriately scoped for the level of programming I am on. I learned a lot, this project made me understand dictionaries much better and understood that you can take data from the 'real world' when we are programming (which I thought would be much harder).

Improvements you could do with this project is to eliminate the extra symbols I mentioned in the results. If you wanted to develop it further, you could compare this book with another Emily Bronte book to see if there are any difference. After that, you could compare it to another author to see if here is a difference.