Name: Zi Liang, Ong

Project 3

I used the data source that was used as an example for the mini project 3. It is a book from Project Guntenberg called Oliver Twist. The technique I used to project the books was counting the frequencies of each word and the top 10 most frequently used words in the text. In order to shorten the amount of time Python needs to process the text, I shorten the story up till Chapter 4. I hope to be familiarised with dictionary because I have never deal with dictionary in C programming previously.

The first thing I did was to extract each word and put them in a list for processing. I chose to deal with list at first because I find it easier to deal with than dictionary because I can access the individual words in the list easily. Each word in the list have to be uniform meaning that there is no punctuation and all letters are in lower case. I assumed that punctuation only appears at the end of the word so I code a algorithm that looks at the last character of a string and remove it if it is a punctuation. However, it is not the case for speech - "..". Therefore, I had to look up a new function called translate which removes anything that I do not want in a string.

I used the exercise in Think Python and make use of get function to find the frequencies of each word. It is clear and easily understandable instead of Histogram function in the book. In order to sort the dictionary from the smallest to largest value, I used the idea of selection sort because it is a fixed algorithm I had used previously which will always work. However, it is not as concise and error free as the sorted function incorporated in Python. However, the sorted function returns a list without the frequencies of word appearing. I used a for loop to combine the values and keys into the dictonary but I do not know the reason why in line 37 of my code, it does not input keys and values systematically into the dictionary. The second input is sloted into the first index of the dictionary and the third input is slotted into the middle of first and second index. Therefore, I had to used the sorted function once again. In order to do this, I had to go a step further and look into Tuple in order to use the sort function once again. It is because now I have to form a list in a list out of a dictionary.

I found interesting things while analysing the text such as using words like "the" so many times in a book. I think that the most fruitful learning objective is to realise even more useful functions that Python has.

I think that Monday's class has helped me a lot to get the project started. It is because it is pretty complicated to extract sources from internet, once the text is available to analyse, it will be easier to think of ways to process the text. I think that text analyse is one of the most important skills in engineering because analysing data is crucial. I wish that I could have more time so I can implement text similarity as well because I think that it is pretty interesting.