**Q3 Report**

Referring the script named ‘Q3.py’ in the same folder, the ideas of the code are as follow:

The ‘with open’ function was used to read a .txt file containing the paragraph and stored as a string object. Then the split function was used to split the string according to the ‘. ’ separator and the individual sentences were stored in a list.

1. It is required to determine the probability of the word ‘data’ to appear in each line of the paragraph. What I did was to check each and every sentence in the list to see the string ‘data’ appears or not. If yes, the counter was increased by 1. The probability can be computed through the following equation:
2. It is required to show the distribution of distinct word count in each of the sentences. What I did was first convert the sentences into lower cases. Then I removed all the punctuations such as comma and brackets to homogenise the words using regex. After that, we can split the sentence into individual words and store them is a list. Finally we can use the set data structure to retrieve distinct words and the size of the set can be viewed as the number of distinct word in the sentence. The results can be plotted into a histogram.
3. It is required to determine the probability of the word ‘analytics’ to come after the word ‘data’. What I did was count the number of the occurrence of the word ‘data’ is the whole paragraph (converted to lower case). Similar thing was done to the word ‘data analytics’. By comparing the ratio, we can obtain the probability using the following equation.

Limitation: This method does not consider the case when the period appears in the middle of the sentence. For example:

*EDA is often compared to detective work, while CDA is akin to the work of a judge or jury during a court trial -- a distinction first drawn by statistician John W. Tukey in his 1977 book Exploratory Data Analysis.*

Similarly, dealing sentences that do not end with a period (maybe an exclamation mark) can also be a problem. The code can be improved by using natural language processing (NLP) libraries (such as spaCy) that use some models that can split the sentences based on the entity.