Assignment2：Text Analysis and Text Mining

1. Project Overview

In this project, I did some research in two books of Charles Dickens which are *Oliver Twist* and *A Child's Story of England.* The txt files were obtained from Project Gutenberg. Data cleaning, text summary statistics and sentiment analysis were accomplished. Some interesting results are illustrated in this report.

2. Implementation

The whole project can be divided into three parts:

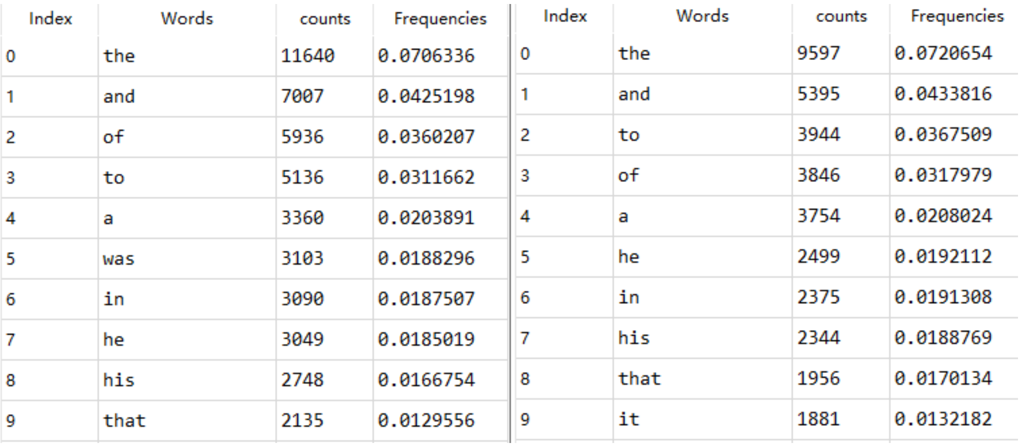
The first part in to download required txt files. A list contains all urls are initialized and thus we can use a loop to download all txt files. A time delay was added in the loop in case the website denies too frequent request. In this case, I downloaded two books and this block of code can be reused in the future.

The second part is about cleaning the data. There are two main tasks. Firstly, remove all the characteristics that are not alphabetic and space. The other task is to remove irrelevant contents before and after the main body. In this part, we use list() to transform string into list that contains every single character. The reason to do that it is more convenient for list to operate elements than string. After special characters were removed. A string with no special characters was obtained by join() and then it was split to get a list contains words. Then we can iterate the list to find the start and end of the main body and slicing is.

The last part is to analysis the clean data. I firstly count the word occurrence of every unique word in the book and calculated the frequencies of these words. Then plotted several graphs to illustrated the finding. These processes were handled with pandas data frame and matplotlib. I am also interested in how a sentiment analysis techniques would conclude about these two books, so I did a NLP sentiment analysis.

3. Results

Firstly, by comparing the top ten word with highest frequencies in two books as shown in figure 1 below:



Index	Words	counts	Frequencies
0	the	11640	0.0706336
1	and	7007	0.0425198
2	of	5936	0.0360207
3	to	5136	0.0311662
4	a	3360	0.0203891
5	was	3103	0.0188296
6	in	3090	0.0187507
7	he	3049	0.0185019
8	his	2748	0.0166754
9	that	2135	0.0129556

Index	Words	counts	Frequencies
0	the	9597	0.0720654
1	and	5395	0.0433816
2	to	3944	0.0367509
3	of	3846	0.0317979
4	a	3754	0.0208024
5	he	2499	0.0192112
6	in	2375	0.0191308
7	his	2344	0.0188769
8	that	1956	0.0170134
9	it	1881	0.0132182

Figure 1: Text Statistics. Left:A Child's Story of England, Right:Oliver Twist

We can see that the words and their frequencies are very similar. We may raise an idea that the occurrence of these articles and prepositions like ‘the’, ‘a’, ‘of’, and ‘to’ takes a stable part in articles if we stand in the macro view (such as all human’s literary works).

To prove this idea to some extent, we plotted the frequencies of words that in the first books’ top 100 that also appears in the second book as shown in figure 2.

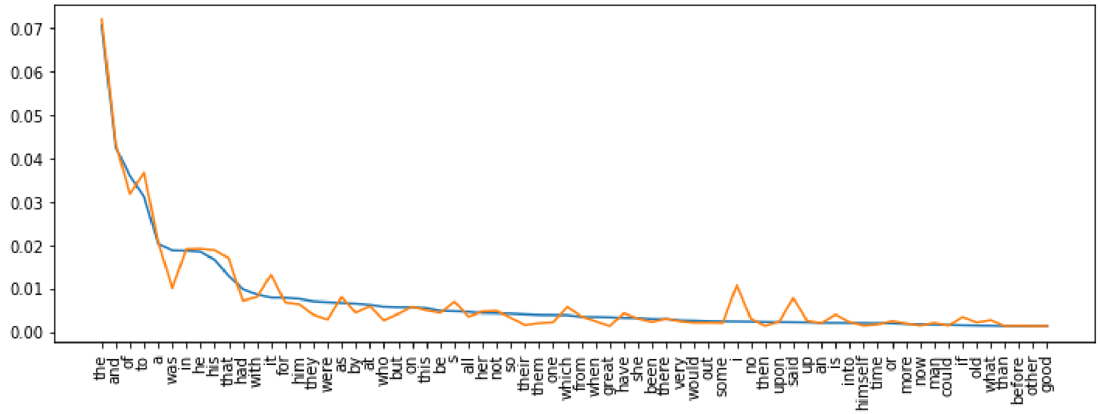


Figure 2: Frequencies of 68 common words in two books.

We can see in figure2, there are 68 works are in common in the top 100 words in two books and their distribution are quite similar which supports our conjecture.

Finally, used the nltk library and we got the algorithm conclusion:

{'neg': 0.123, 'neu': 0.75, 'pos': 0.127, 'compound': 1.0} for *A Child's Story of England*

and

{'neg': 0.101, 'neu': 0.761, 'pos': 0.138, 'compound': 1.0} for *Oliver Twist*

It seems *A Child's Story of England* is a more negative book than *Oliver Twist.*

4. Reflection

I think this project was appropriate scoped. In a process point of view, this project is inspirational. It started with an interesting topic. Then following the path that get data, operate data and then analyze the data. I established a sense that how a text mining project proceed and learned several techniques commonly used in text mining.