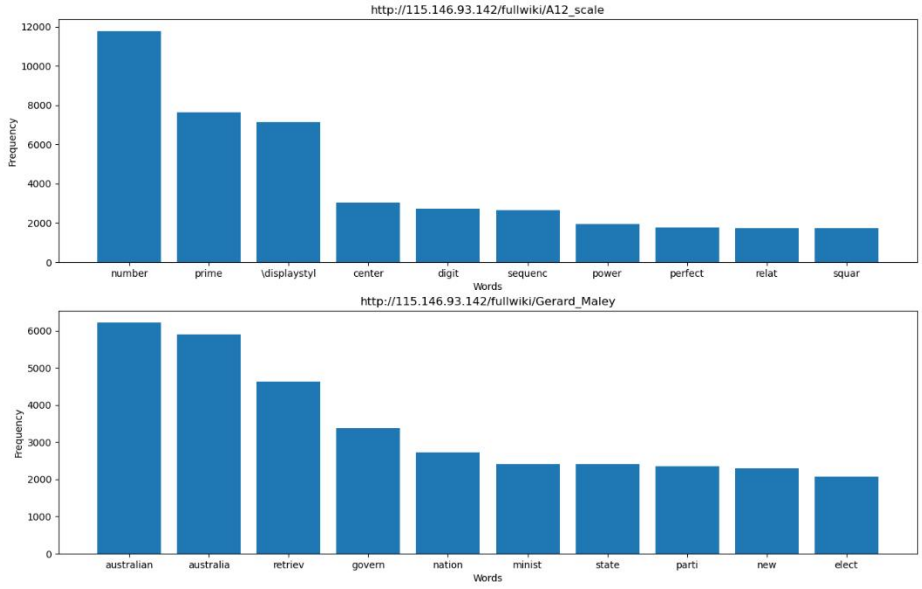
Task6

This brief report summarizes the outcomes and include the analysis of the outputs of this assignment. I used various data processing techniques to get the result data from webpages. This report also includes the plot produced in task4 and task5.

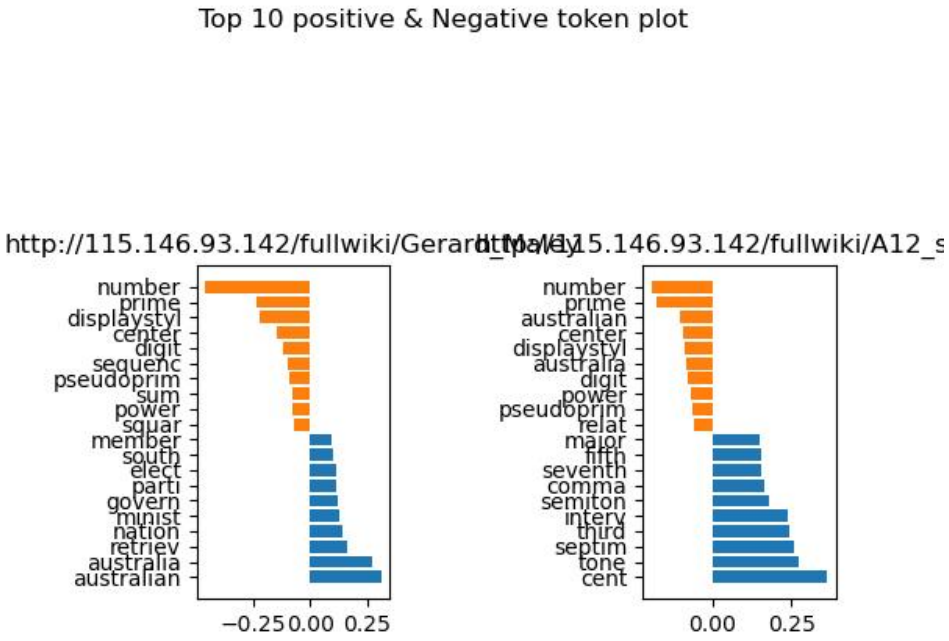
In the task4, according to the plot graph generated in task4 and the dictionary result returned by task 4 the ten most frequent words in seed URL http://115.146.93.142/fullwiki/A12_scale are "number", "prime", "\displaystyl", "center", "digit", "sequenc", " power", " perfect", "relat", "squar". Find out that those words come from music area. Which also demonstrate the content on this webpage mainly is about music. For the seed URL http://115.146.93.142/fullwiki/Gerard_Maley the ten words who have the most frequency is " australian", " australia", "retriev", "govern", "nation", "minist", "state", "parti", "new", "elect". All ten words are related to the politics of Australia. Thus, can predict that http://115.146.93.142/fullwiki/Gerard_Maley has the theme of Australian politics which is significant different from the theme of http://115.146.93.142/fullwiki/A12_scale whose theme is about music.

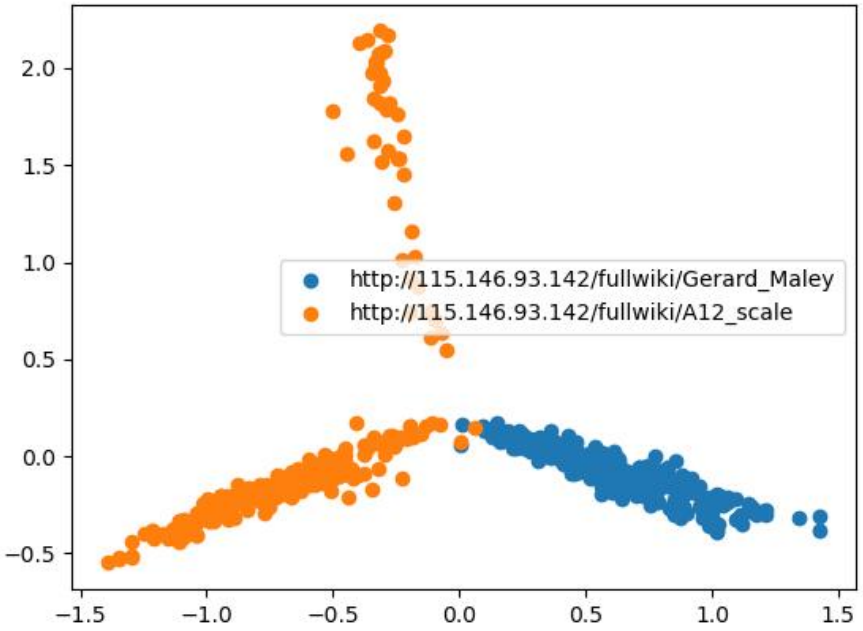


Words we might not be surprised to find in articles for each Seed URL based on the information in task5a.png and task5b.png: From task5a.png, and task5b.png, we can observe that for both two seed URL which are http://115.146.93.142/fullwiki/A12_scale and http://115.146.93.142/samplewiki/Gerard_Maley the positive words all have “number”, “prime”
Thus, it is easy to find “number” and “prime” in both webpages.

In task5b.png, we can see that the URLs for http://115.146.93.142/fullwiki/A12_scale are

clustered on the left side of the graph, while the URLs for http://115.146.93.142/samplewiki/Gerard_Maley are clustered on the right side of the graph. However, there is some overlap between the two clusters in the middle of the graph. Based on this, we might be able to determine which seed URL a new unseen link originated from when plotted in the 2D space after applying PCA, but there might be some ambiguity in cases where the link falls in the overlapping region. Overall, the results of task5 suggest that the word frequencies of articles can be used to distinguish between different topics and sources.





Though the dataset produced is quite comprehensive but there are still some limitations and shortages. The datasets contain noise in the form of irrelevant words, typos, and misspellings. This may affect the accuracy of the results and make it difficult to draw meaningful conclusions. And also, the dataset only can summarize the basic words and the frequency of the words that showed in the webpages, cannot make a summarize the themes or other particular information from the webpages. To the aspect of processing techniques there are also some drawbacks, such as tokenization and stemming which is the crucial part in crawling. They have limitations that can affect the accuracy of the results. For example, tokenization may not capture all the nuances of language use, and stemming may not always accurately identify the root form of a word.