Data from a site hosting over 500 pages was taken from Wikipedia and through techniques of data processing, data was retrieved and visualised. The first half of the project focused on retrieval of information from two seed URLs in which crawling, scraping and bag of words representation was used. The latter half of the project focused on producing plots to visualise the data from the webpages. Through the process, plots describing and comparing the top 10 most common words from seed URLs and tokens with the highest weights in the principal components obtained from PCA were created, as well as a scatterplot showing PCA applied to the seed URLs.

Focusing on the visualised results of the project, task 4 produced a plot comparing the top ten most common words from two seed URLs. The most common words from "http://115.146.93.142/fullwiki/A12_scale" seem to be heavily related to mathematics. Comparing this to the most common words from "http://115.146.93.142/fullwiki/Gerard_Maley", which all seem to be related to Australia and politics, we see that these two seed URLs contain information on completely different topics. From the graph it is evident that the two seed URLs were highly likely to belong to different categories, as the grouped bar chart showed no overlap in common words, seen by the fact that there was no overlap in common words and also how there were no grouped bars in chart. The most likely reason that there is such a drastic difference between the top 10 most common words is that the seed URLs branched off into different categories, in this case one branch into maths and another into Australia and politics.

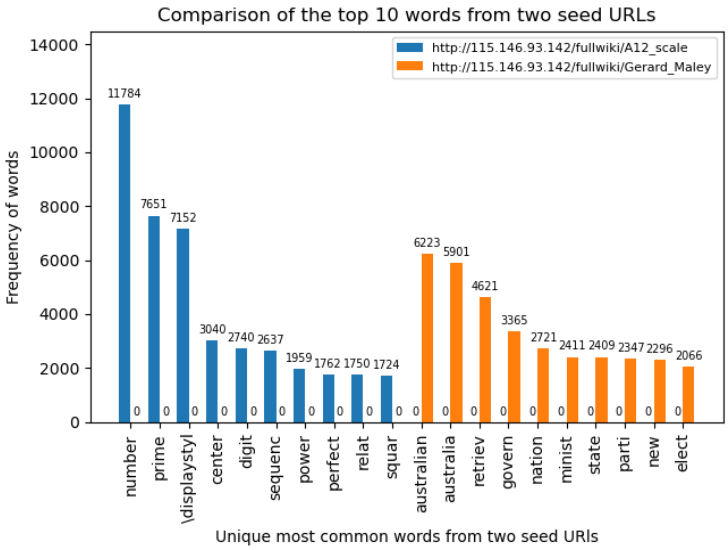

Figure 1. task4_my_full.png

Looking at the plots produced from task 5. The Task5a plot appears to have three types of tokens. On the left and right sides of the graph we see the tokens' weights that mainly affected only one component. In the middle we see some overlap, indicating words that weighted into both components. Looking at how the graph is presented, the words that overlap and don't is not surprising. The three groups' words appear to be dissimilar and may stem from different categories. Furthermore, in the scatterplot it can be seen that there

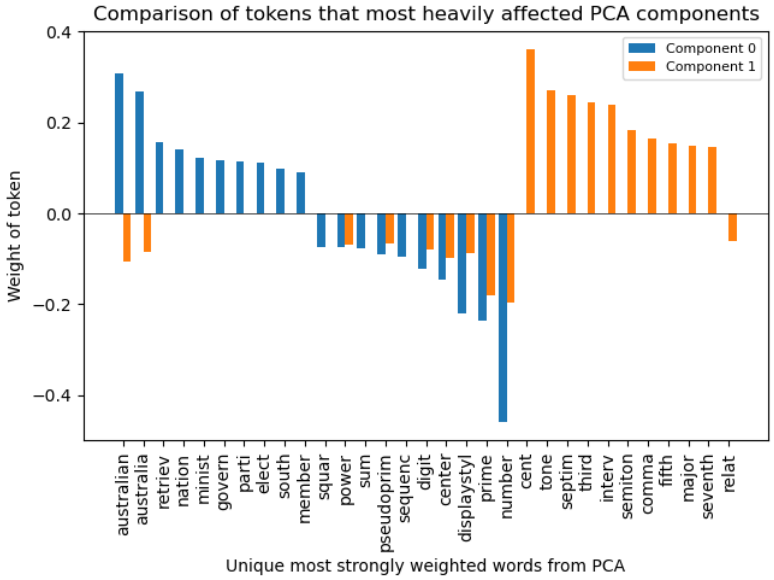

Figure 2.a Task5a.png

appears to be 3 clusters of URLs. This further implies that the words we find in articles would be separated into three categories.

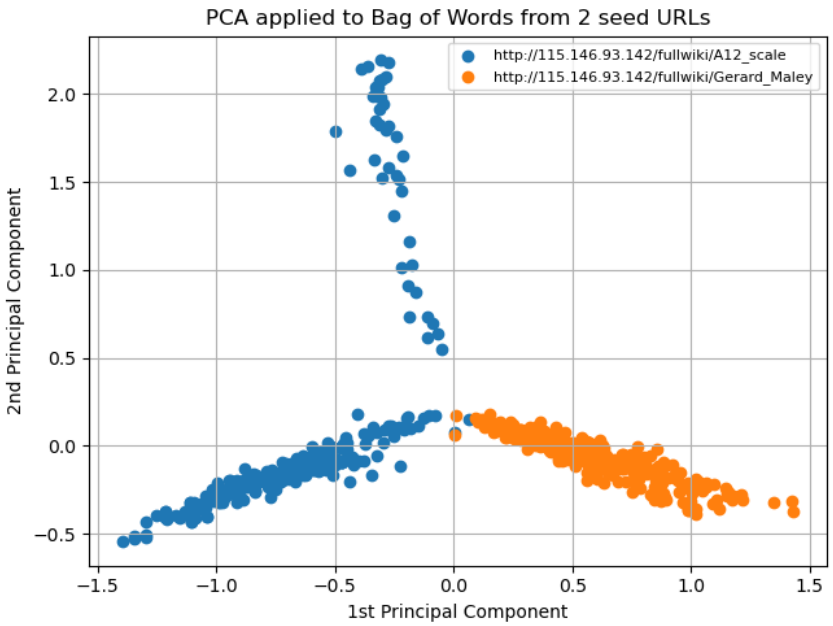

Figure 2.b Task5b.png

In task 5 the scatterplot was produced from applying PCA to the bag of words representation from the two seed URLs. Interpreting the distribution of the plot it can be seen that about three clusters have formed. Two belong to "http://115.146.93.142/fullwiki/A12_scale" and are seen on the left and upper areas of the plot and one belongs to "http://115.146.93.142/fullwiki/Gerard_Maley", seen on the right of the plot. In the middle of the plot a small area of overlap can be seen. If an unseen link originating from the seed URLs was plotted into the plane it would be relatively easy to determine which seed URL it originates from as long as it is not plotted in middle of the plot; the area of some overlap.

The dataset used in this project only allowed the crawling of wikipedia pages associated to the seed URLs which may not be fully representative of the actual branches that stem from the seed URLs. Furthermore, the process of obtaining tokens and comparing bag of words may have affected the representations of the seed URLs. In this project, vector similarity from BoW was looked at in detail, however, implementation of techniques like TF.IDF may have provided further insights into contents of seed URLs. Another limitation would be PCA, as it is unknown exactly what each of the principal components entail, affecting our judgement of what the scatterplot exactly represents. Furthermore, cluster analysis could also help in further solidifying the boundaries between the articles produced in turn helping in providing further insights.