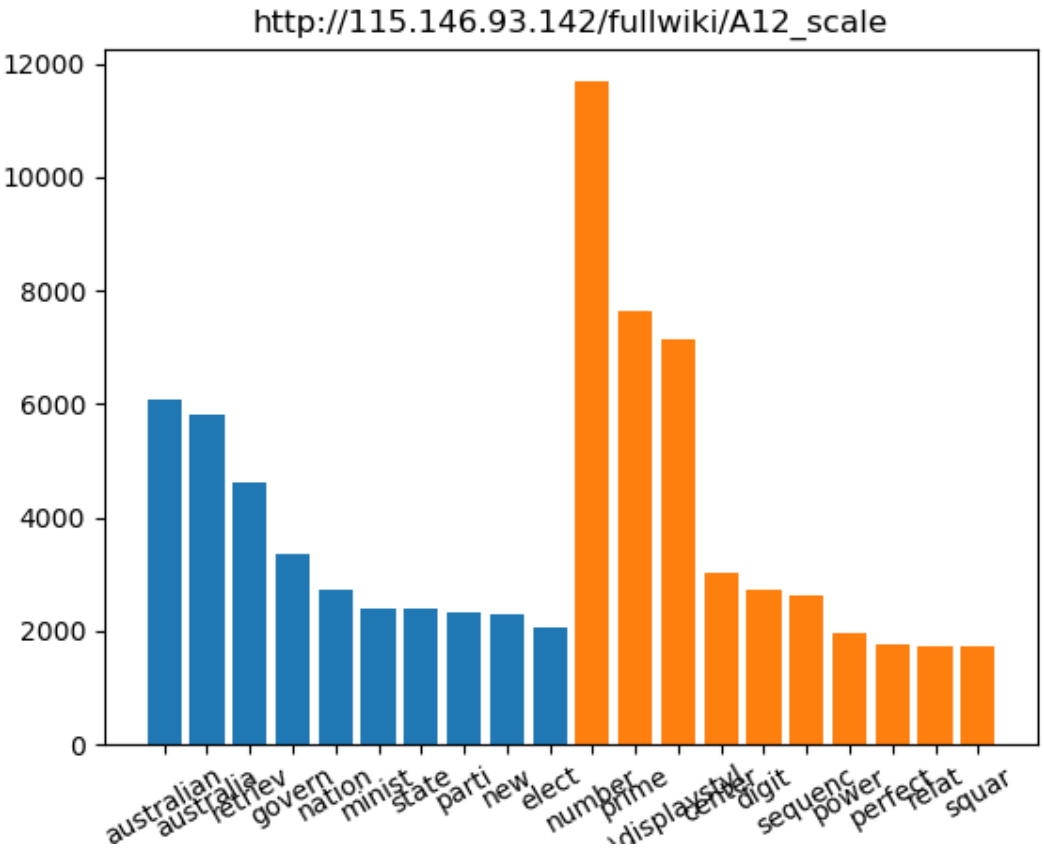
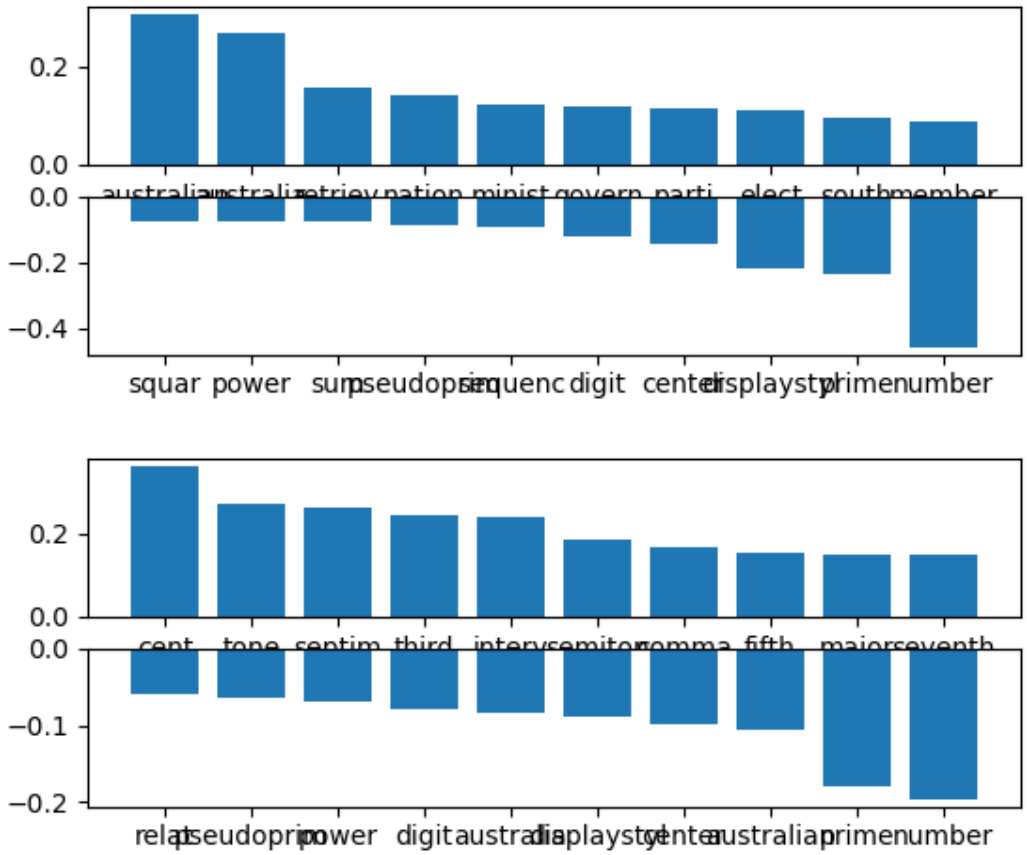
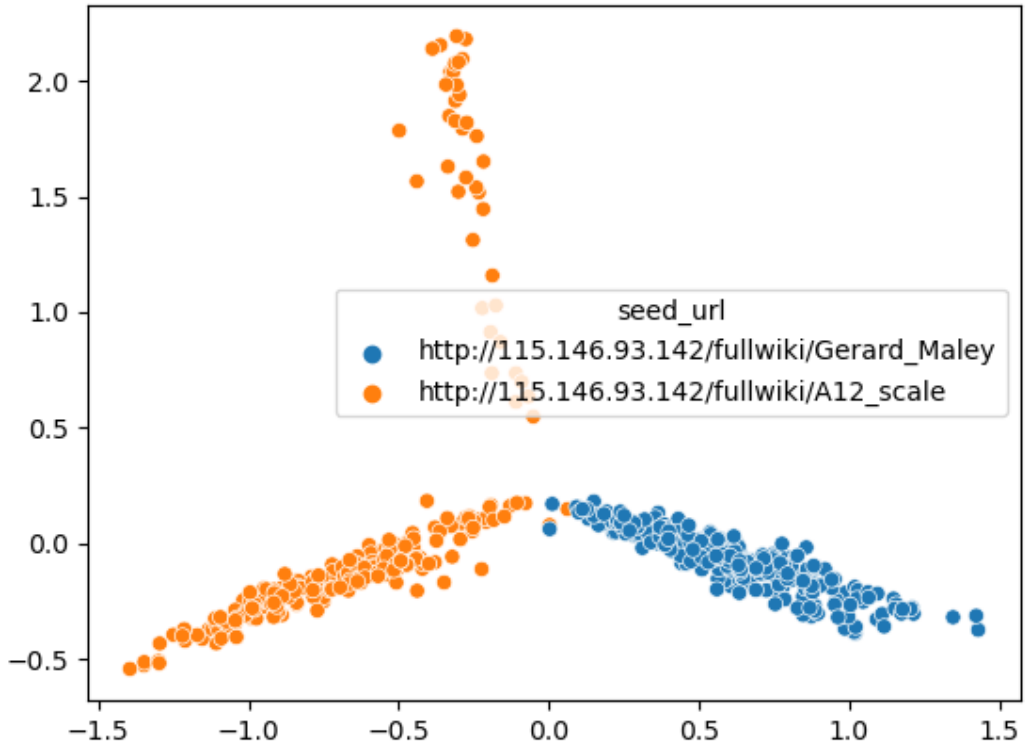
Task 6: Analysis Report



Task 4 Graph for Full dataset



Task 5 Graph 1 for Full dataset



Task 5 Graph 2 for Full dataset

The top 10 words in the first seed_url consists of (australian, australia, retreiv, govern, nation, minist, state, parti, new, elect) whereas the second seed_url consists of (number, prime, \\displaystyl, center, digit, sequenc, power, perfect, relat, squar). There might be a difference between the two as the websites may be focusing on different subjects or content therefore some words may be used more than others to describe a particular topic, another reason could be the structure of the website as in one website more pictures may be present compared to others which might change the most repeated words of a website. Based on the information present in both the graphs in task 5 we would be least surprized to see the words (australia, australia, retreiv, govern, parti, elect) in the first seed_url and (squar, power, sequenc, digit, center, \\displaystyl, prime, number) as these numbers have the highest weights when we perform normalization. The second graph of task 5 shows the distribution of weightage of each word in each seed_link and by looking at the graph we can conclude that the unseen url belongs to the seed_url http://115.146.93.142/fullwiki/A12_scale. The limitation of this dataset is that it only contains two seed_url which limits the testing of the code so even if the code runs perfectly on the two urls it may cause an error in some other url. Another limitation may be that the seed_url are too similar containing the same protocols making it easy to separate protocol which might not work other url protocols. The limitations of processing technique is that each website may have different structures and different rules which may restrict web crawling, also data crawling may require huge amounts of data processing power which may restrict data processing depending on the user's system. To provide further insights we could improve the data quality of the website and use the websites policies to prevent any breach of agreements or any legal issues.