TASK 6: Analysis Report

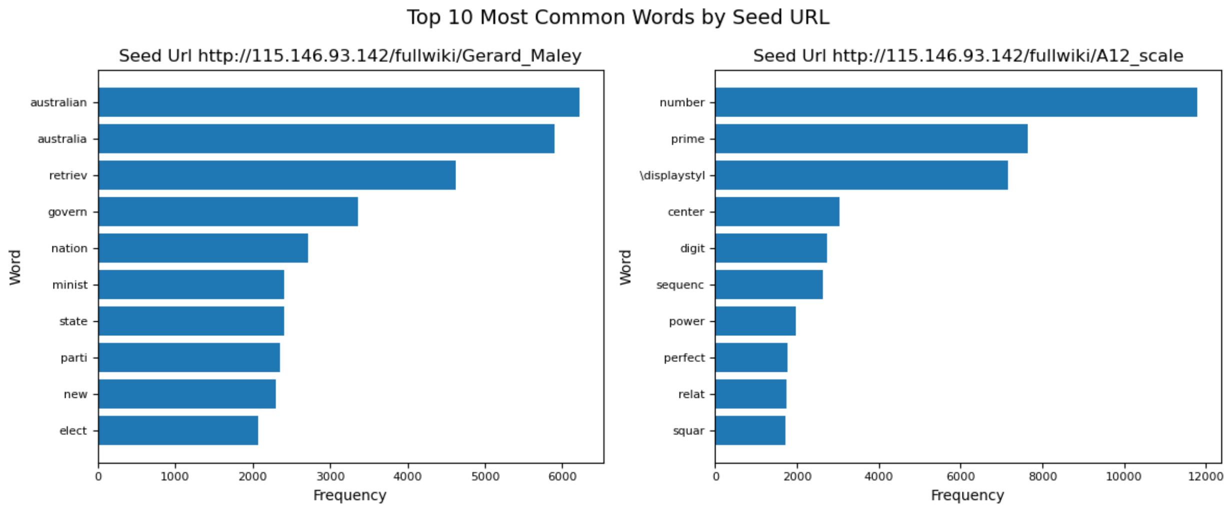

Figure 1: task4_my_full.png

Looking at the first graph from figure 1, Seed URL http://115.146.93.142/samplewiki/Gerard_Maley TOP 10 words are **["senat", "vote", "australian", "australia", "elect", "constitut", "parti", "state", "hous", "territori"]**, which suggest a theme related to Australian politics or government.

Whereas seed URL “http://115.146.93.142/fullwiki/A12_scale”, words like **[“number”, “prime”, “sequence”, “digit”, “power”, “perfect”, and “square”]** made it to top 10 words. This suggests that content in this URL might be related to mathematics or numbers.

These differences are present here since the content, topic, theme on each of these seed URL are different; thus, the most frequently used words on each page are also different.

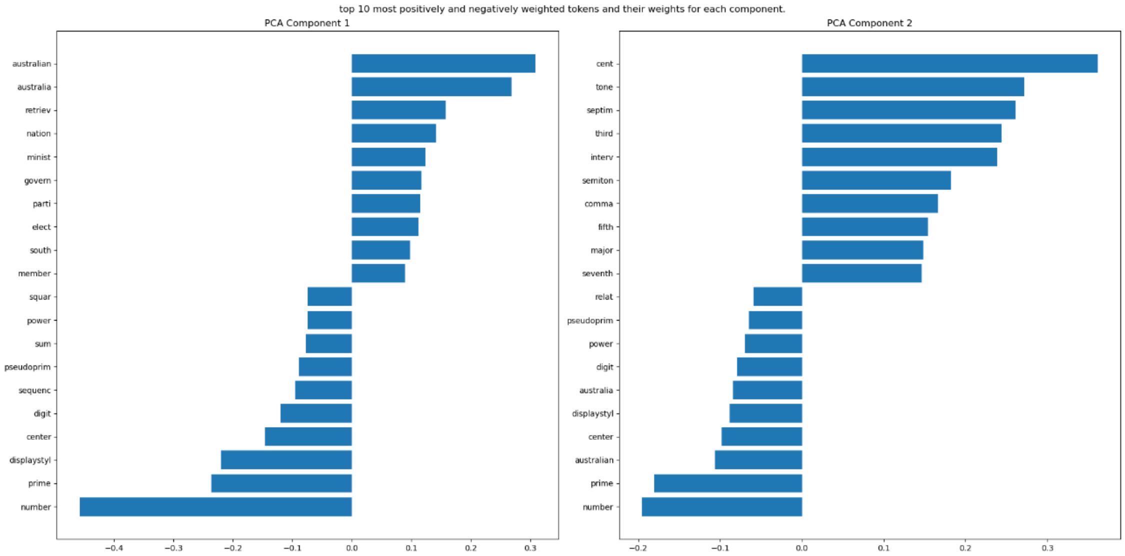

Figure 2: task5_my_full_5a.png

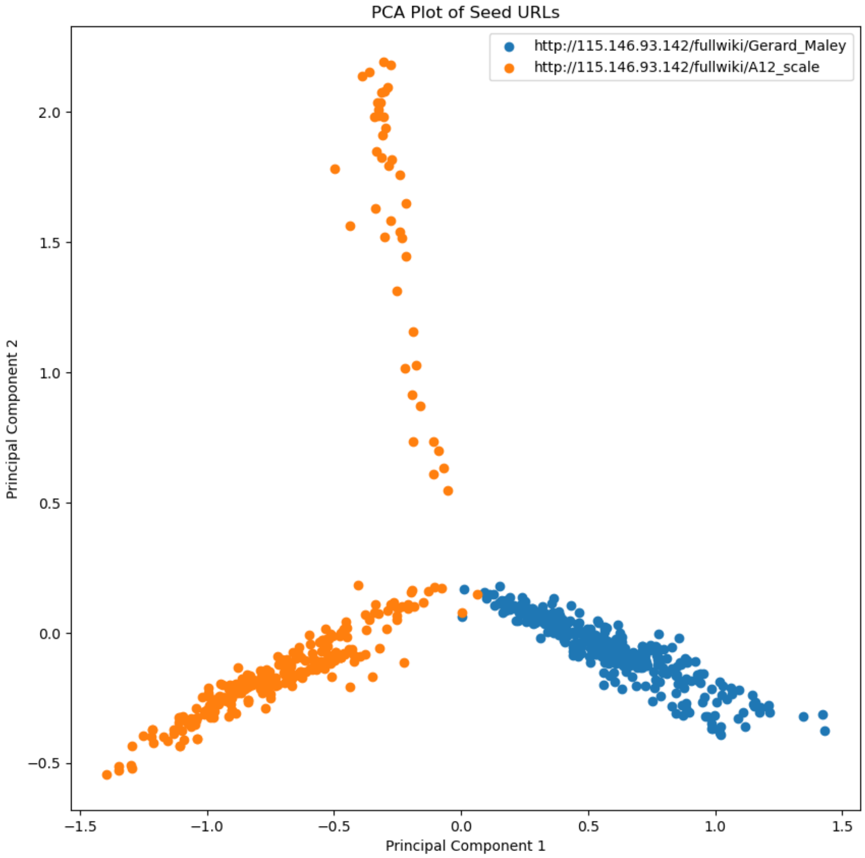

Figure 3: task5_my_full_5b.png

Looking at figure 2 and figure 3, we can see Principal components 1 and 2 assign positive weight to and negative weight to words from the seed URL http://115.146.93.142/fullwiki/Gerard_Maley. Thus, we would not be surprised to find out most of the words in this URL are related to Australian politics and government theme, i.e. “Australian”, “Australia”, “retrieve”, “nation”, “minist”, “govern”, “parti” and “elect”.

On the other hand, the seed URL "http://115.146.93.142/fullwiki/A12_scale" appears to contain the top words with negative weight assigned by both principal component 1 and principal component 2. Since this URL is related to mathematics or numbers, we can easily predict words like “number”, “prime”, “power”, “digit”, “sequence”, “sum” and “pseudoprime” will appear.

Additionally, looking at principal component 2, we can see words like “cent”, “tone”, “septim”, “third”, “interv”, “semiton”, “comma”, “fifth”, “major” and “seventh” are assigned. We should not be surprised to see these words since URL “http://115.146.93.142/fullwiki/A12_scale” is related to music theory and composition.

Having another look at figure 3, the PCA scatter plot depicts somewhat three separated clusters, meaning the URLs grouped in these clusters share same patterns or similarities. These clusters might contain URLs pertaining to music theory, Australian politics and governance, and mathematics and numbers.

However, some overlap still exists between these clusters. When plotted in 2D space after PCA, it becomes reasonable to determine the origin of a new undetected link based on point distribution, depending on how well it fits into the existing clusters; and this dependent on how precisely the new link can be assigned to a certain seed URL, particularly if it is close to the boundary of two clusters. Furthermore, other factors or other topics that are not effectively represented by PCA components may make it difficult to accurately categorising new links.

There are 2 main limitations to this dataset.

Firstly, there are only 2 seed URLs and their respective web pages in the dataset, which makes it a relatively small dataset. This restricts the result's capacity to be generalized and might not adequately capture the trends found in larger or more varied datasets.

Secondly, because the web pages were scrapped all at once, any temporal changes in content over time are not recorded. This could potentially skew the results as the content may no longer be current or accurate, and could affect the dataset's ability to reflect the current state of the web pages.

The processing techniques used have some restrictions as well.

For starters, this technique is that it may not be suitable for all languages, as it assumes English stop words and uses the Porter stemming algorithm which is optimized for English language. For languages with different grammatical rules or syntax, this technique may not produce accurate results.

Additionally, tokenization method used splits words based on whitespace, this can result in incorrect word counts and incorrect tokenization of some words. And stemming algorithms can sometimes produce stemmed forms that are not actual words, which could also result in incorrect interpretations of the text. PCA also assumes that the data is linearly separable, which may not hold true to all datasets.

Furthermore, the analysis only considers the top 10 words, which might not fully reflect the complexity of the data.

Taking all these limitations to consideration, for future work, we could consider adding more or use larger data with more seed URLs and their respective webpages. We should also scrap the web pages at various times to record any temporal trends or content change. Additionally, more advanced processing techniques could be used, such as those that take context into consideration while processing natural language. Lastly, rather than relying exclusively on visual analysis of the PCA plots, machine learning algorithms could be used to catagorise web pages according to their content.