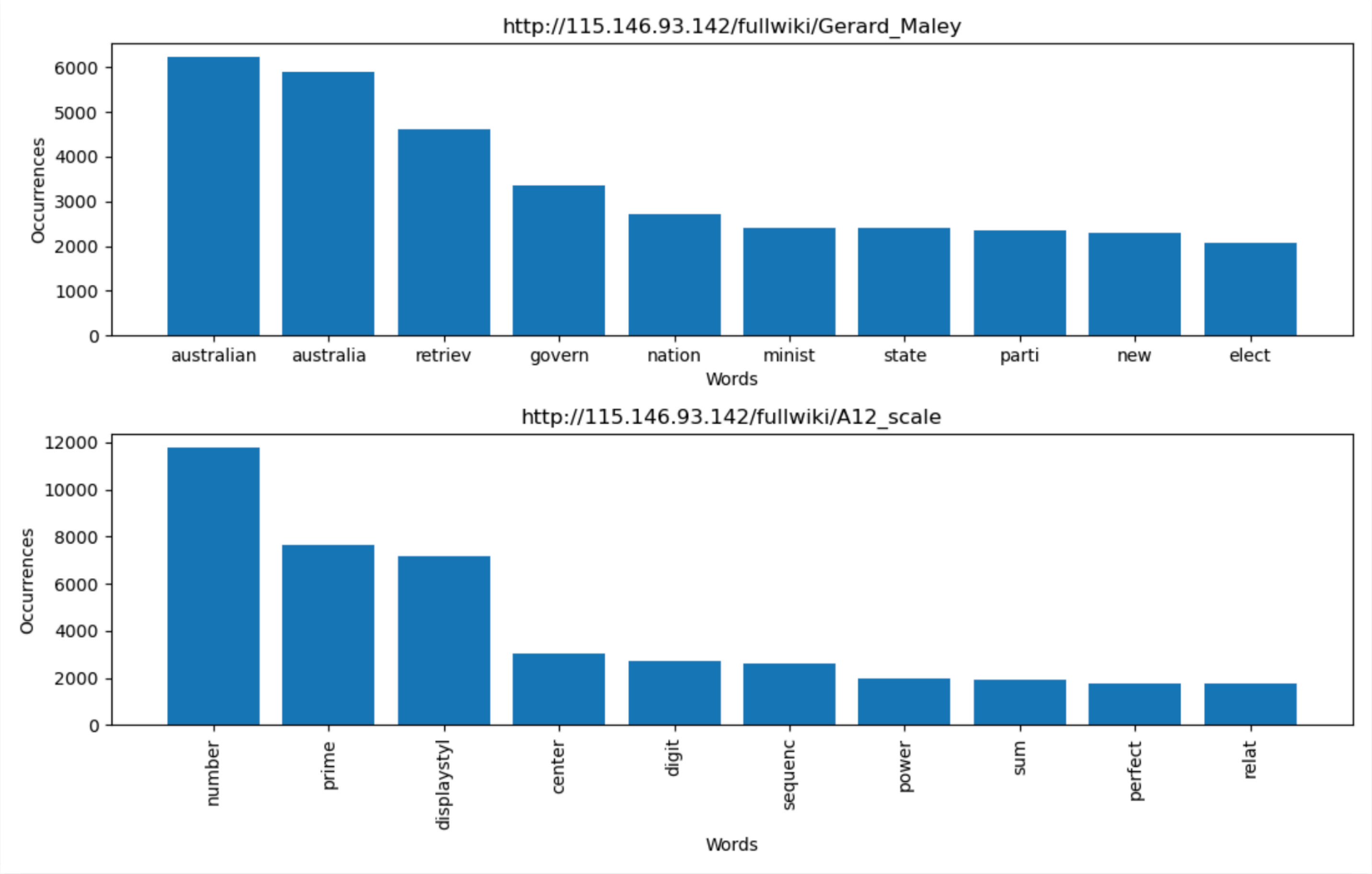

Figure 4

Task 4 takes a pandas dataframe and outputs a bar chart representing the most common words across each webpage "http://115.146.93.142/fullwiki/A12_scale" and "http://115.146.93.142/fullwiki/Gerard_Maley" and the most common words found were “number”, “prime”, “displaystyl”, “centre”, “digit”, “sequence”, “power”, “sum”, “perfect”, and relat” for the former and “australian”, “australia”, “retrieve”, “govern”, “nation”, “mint”, “state”, “ parti”, “new”, “elect” for the latter. In terms of repetition, the most common word for the first seed URL ("http://115.146.93.142/fullwiki/A12_scale"), exceeded the top word for the second URL ("http://115.146.93.142/fullwiki/Gerard_Maley") by a large margin. Although these 10 words have been repeated over a thousand times in each of the articles, there seems to be no common word, that made it to the top 10, between the two seed_urls. The differences in words could be because these two URLs are of different web pages, with different domains. After analyzing the words it could be deduced that the first URL is about politics as it contains the words “govern”, “nation”, “parti” while the second URL could be about music since it possesses the words “sequence”, “digit”, “perfect”. This gives rise to another reason why there is a difference in the common words between the two URLs. Nonetheless, these common words give the readers a hint about the page's subject.

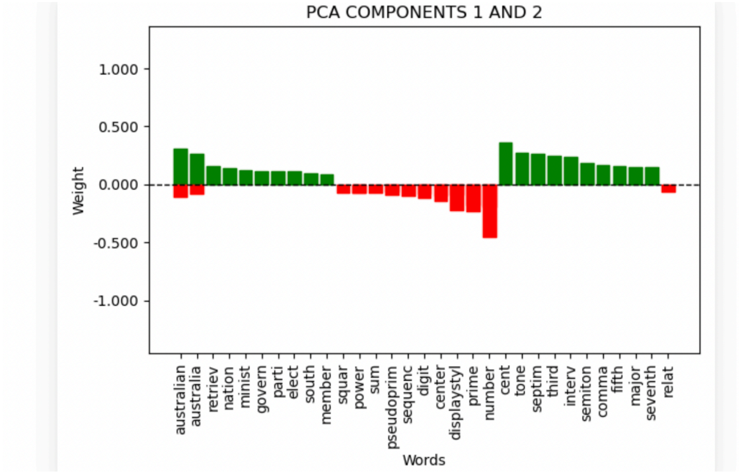

Figure 5a

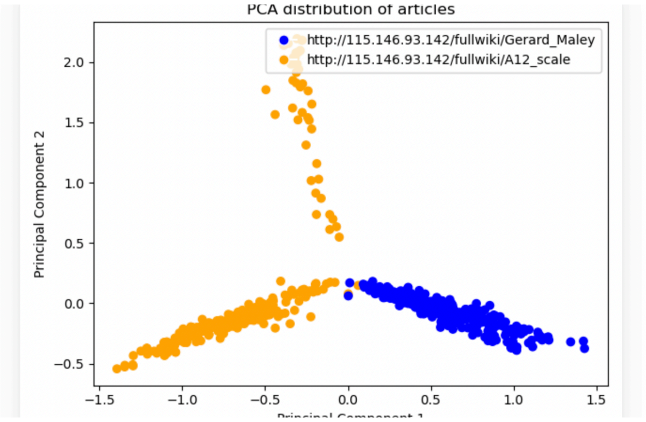

Figure 5b

Figure 5a is a bar chart of the weight of the words and figure 5b is a scatter plot of the article. After looking at 5a we can see that the words with the highest weights are “australian”, “australia”, “govern”, “parti”, “retrieve”, “tone”, “septim” , etc. This means that these words are the most important words in separating the article. After looking at Figure 5b, we understand that there are very few clusters between the articles from different seed URLs, in contrast to large clusters between the articles from the same URL. This shows that the articles from the same URL are similar to each other whereas there is an enormous dissimilarity between articles of different seed URLs. So all in all, after analysing figures 5a and 5b, we can interpret that political terms such as “govern”, “parti”, and as well as musical terms such as “tone” would not be difficult to find.

As mentioned earlier, it can be seen in Figure 5b that there is a major overlap between articles from the same seed URL and very little clustering between articles from different URLs. The articles from different seed URLs are mostly scattered throughout while clustering among articles of the same seed URL. It is very difficult to determine which seed URL a new unseen link originated from.

One of the limitations is that only the articles with the same domain as the seed URL is collected. This reduces the sample size and hence limits our ability to understand the data and comprehend what the content of the seed URL is about. The sample size of the data can be increased, not only by collecting articles of the same domain but of a different domain as well. This will increase our sample size and give us a more accurate result.