chunguang liu 2230873

DS6501 Assignment 1

Analysis Report of Tweets from Three Political Parties

2024

**Executive Summary**

This report aims to do sentiment analysis towards three New Zealand political parties. All the necessary sentiment analysis stages are shown in detail. All related outputs are also included in this report, such as pie charts, bar chats, tables, and word clouds. Interpretation about the results is generated for better understanding. Among the three parties, Labour Party has the highest positive sentiment score, while National Party has most negative count.

**Stages & Visual Results & Analysis**

**Preparations:**

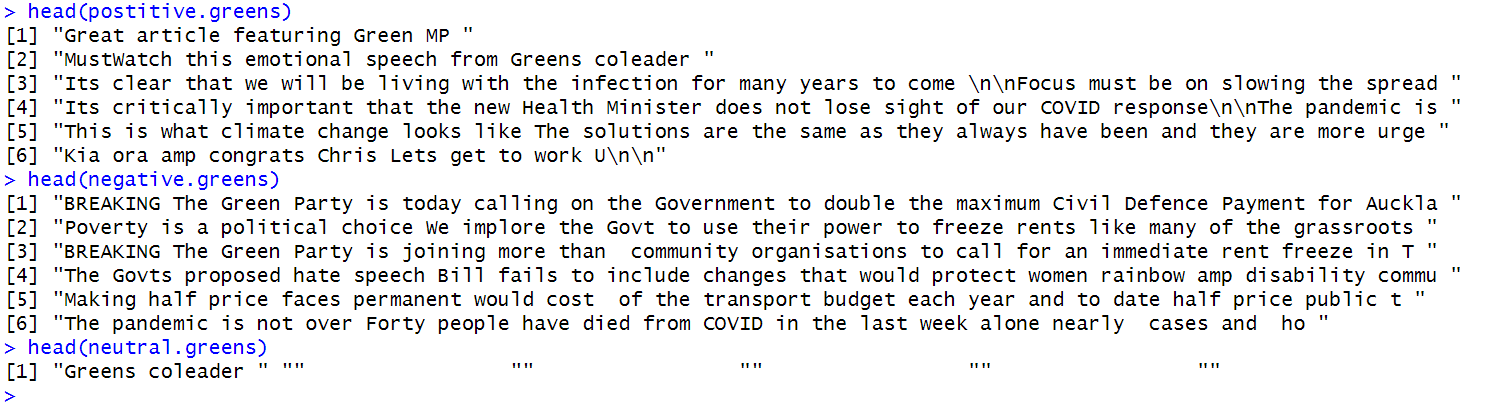
To do the tweets analysis, there are several preparation steps needed to be done with RStudio. The first preparation step is installing relative packages, load, and use them by using library(). The packages include SnowballC, NLP, tm, syuzhet, wordcloud, RColorBrewer, ggplot2, RSentiment, and DT. Then check the current working directory and set it to the preferable directory path. After these preparations, it is the time to load the dataset to analyze using read.csv() function. Verifying the data is always a good habit, so head() and tail() functions are used to check whether the data was loaded successfully.

**Stage 1: Data Pre-processing**

The aim of this part is to clean the data for analysis by removing the URLs, hashtags, punctuation, etc. Only the text column is in use in the analysis. Gsub function is used to remove URLs in the text. the content that needs to be removed is replaced with nothing. Punctuations and digits are also removed. All the clean data is stored in variables named green.df, labour.df, and national.df separately. The new variables will be used in next step.

**Stage 2: Set Sentiment Scores**

In this stage, the sentiment scores to each cleaned tweet text is assigned. The processed clean data needs to be converted to a vector first, then use get\_nrc\_sentiment() function to create score for each tweet, and combine it with original cleaned tweets. After doing this, we can get different score for each sentiment for every tweet content. Then Get\_sentiment() function is used to set the values of positive, negative, and neutral tweets. Thus, all the positive, negative and neutral tweets of the three accounts can be gotten. For example, Figure 1 below is the positive, negative, and neutral tweets of the Greens Party.

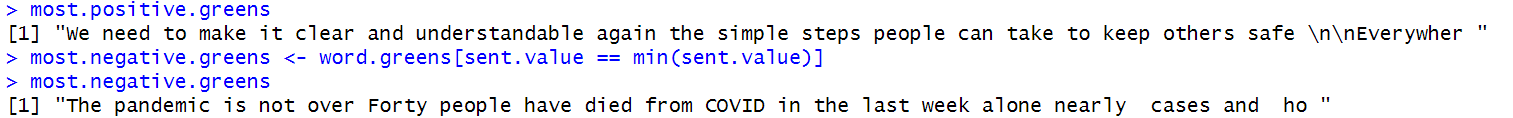


*Figure 2.1 the Positive, Negative, and Neutral Tweets of the Greens Party*

**Stage 3: Display the Most Positive and Negative Tweets**

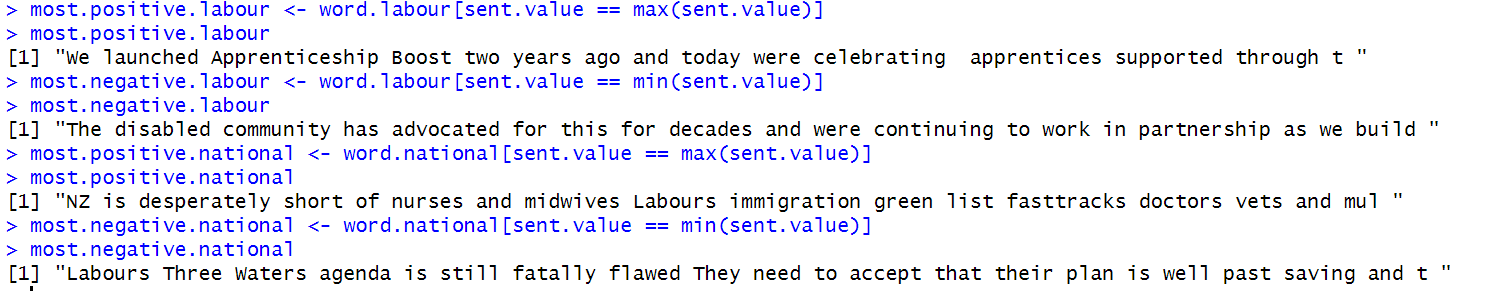
This step is to identify and display the most positive and negative tweets by using the max(sent.value) and min (sent.value) functions on the sorted data. The Figure 3.1, Figure 3.2, and Figure 3.3 show the most positive and negative tweets of the three political parties.

For the Greens Party (Figure 3.1), the most positive tweet mentions clarity, understanding, and safety. All these can be considered as active response. In the most negative tweet, the word “not over” and “died” indicate passive emotion.



*Figure 3.1 the Most Positive and Negative Tweets of the Greens Party*

For the Labour Party (Figure 3.2), the most positive tweet emphasizes the boost indicating the success. The negative one potentially expresses that the progress of disabled community is not fast, even though the effort lasting decades of years.

****

*Figure 3.2 the Most Positive and Negative Tweets of the Labour Party*

For the National Party (Figure 3.3), the most positive tweet has the meaning that fast-tracking immigration helping to address an important problem, which is desperate shortage of nurses and midwives. The most negative one is a kind of dissatisfaction because the agenda has fatal flaws and beyond repair.

**文本

中度可信度描述已自动生成**

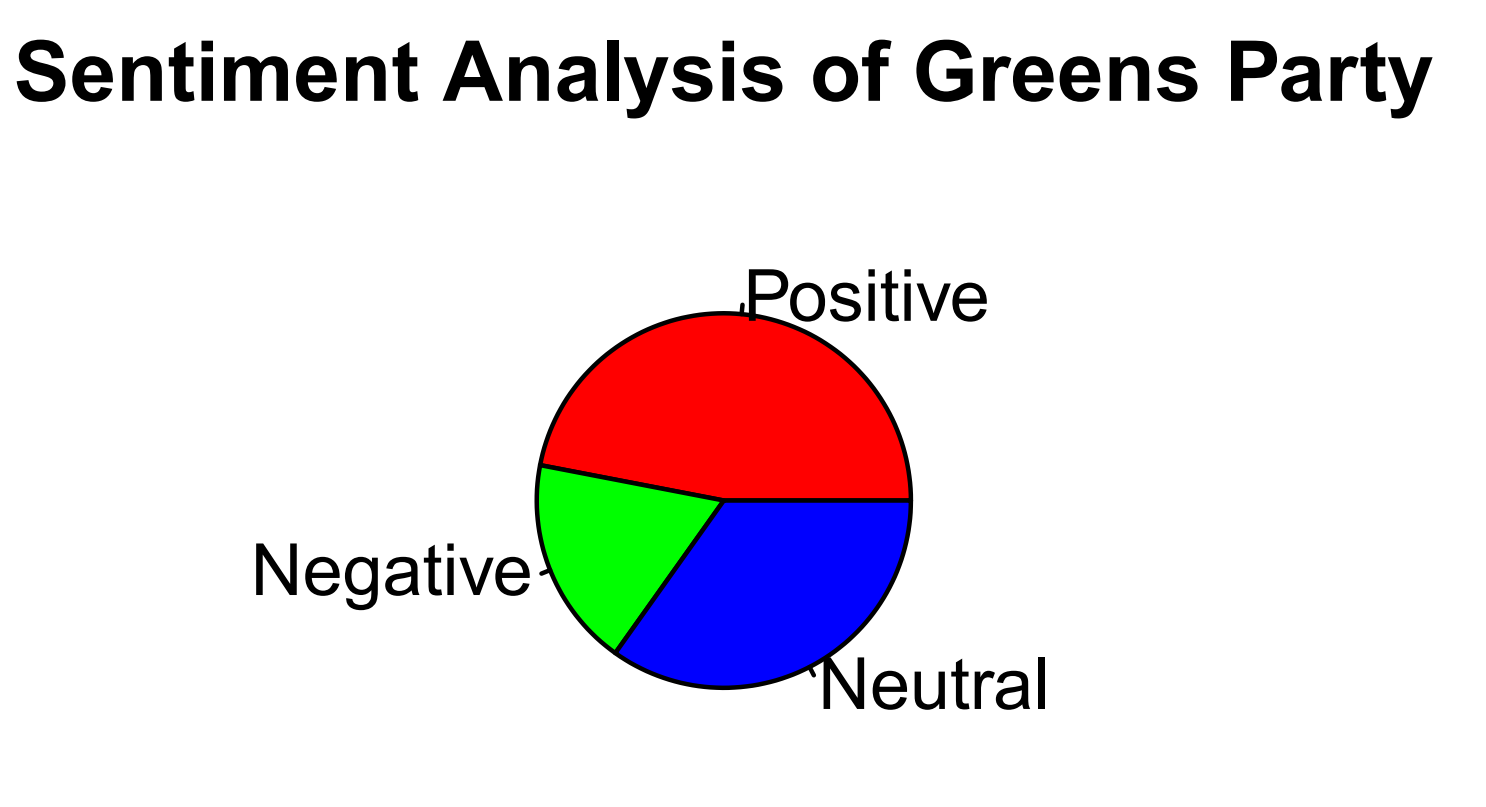
*Figure 3.3 the Most Positive and Negative Tweets of the National Party*

All these most positive and negative tweets from the three parties explain that while algorithm determines positivity and negativity, not only the meaning of words will be considered, but also the tone and context also play a critical role.

**Stage 4: Plot Pie, Bar Charts, and Tables of Sentiment Scores**

A pie chart of positive, negative, and neutral sentiment cores is generated to visualize the proportion of each sentiment. In order to plot a pie chart, the length() function is used to count the number of positive, negative, and neutral tweets. Then wen can combine the count number, define the label, and create the pie.

The pie of Greens Party (*Figure 4.1*) shows that the percentage of positive and neutral sentiment is similar. The percentage of negative sentiment is lower than the other two sentiments. The pie of Labour Party (*Figure 4.2*) indicates that the proportion of positive sentiment is significantly higher than negative and neutral sentiment. From the National Party’s pie chart (*Figure 4.3*), the proportion of positive and negative sentiment is similar and higher than neutral sentiment.



*Figure 4.1 Greens Party’s Sentiment Pie Chart to Show the Proportion of Positive, Negative, and Neutral*

图表, 饼图

描述已自动生成

图表, 饼图

描述已自动生成

*Figure 4.2 Labour Party’s Sentiment Pie Chart to Show the Proportion of Positive, Negative, and Neutral*

图表, 饼图

描述已自动生成

图表, 饼图

描述已自动生成

*Figure 4.3 National Party’s Sentiment Pie Chart to Show the Proportion of Positive, Negative, and Neutral*

Besides the simplified and overview sentiment analysis like the pie charts above. Some more complex sentiment analysis can be done, like bar charts including more emotions. These bar charts are created to display the exact number of each sentiment by using the barplot() function and text() function. By using DT::datatable(), related tables can also be generated.

From the bar chart and table, we can see that in Greens Party the highest count is positive sentiment. The score of “trust” is also relatively high. The positive sentiment of Labour Party is significantly higher than the other two parties, which is 116. The trust sentiment of Labour Party is also higher. The negative sentiment of National Party is the highest among the three parties, which is 69. Also, the fear and sadness sentiments are higher than the other two parties.

图表

描述已自动生成

*Figure 4.4 Bar Chart of Greens Party’s Sentiment Score to Show the Total Scores of Each Sentiment*

表格

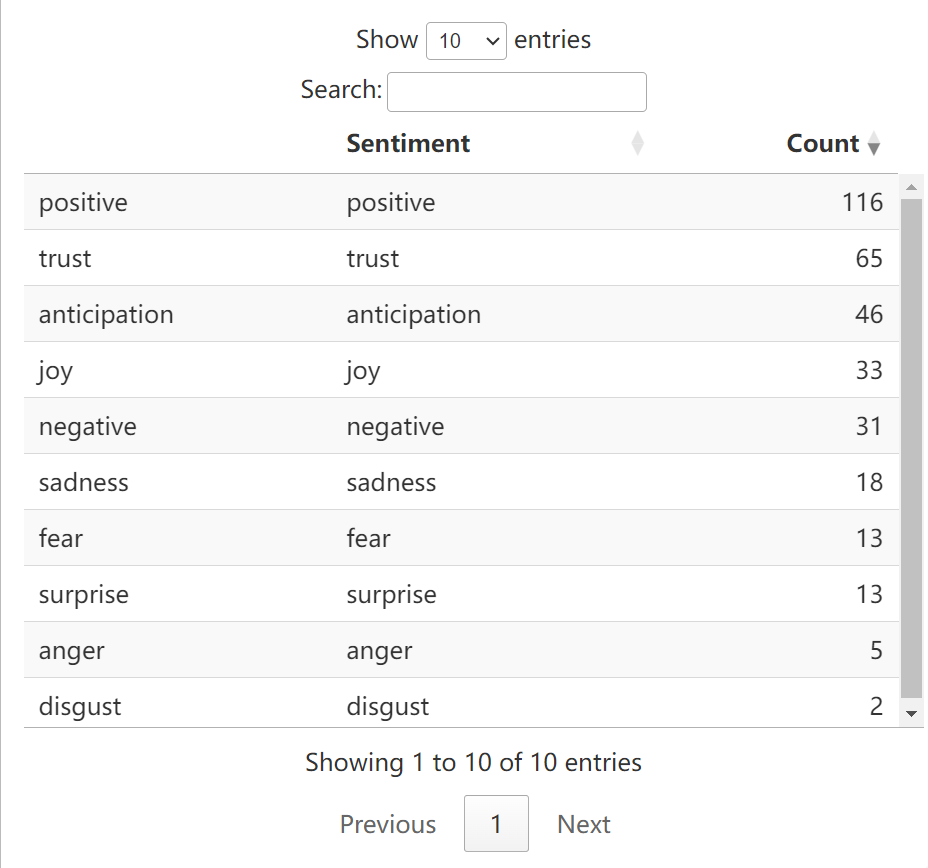
描述已自动生成

*Figure 4.5 Table of Greens Party’s Sentiment Score to Show the Total Scores of Each Sentiment*

图表, 条形图

描述已自动生成

*Figure 4.6 Bar Chart of Labour Party’s Sentiment Score to Show the Total Scores of Each Sentiment*

****

*Figure 4.7 Table of Labour Party’s Sentiment Score to Show the Total Scores of Each Sentiment*

图表, 条形图

描述已自动生成

*Figure 4.8 Bar Chart of National Party’s Sentiment Score to Show the Total Scores of Each Sentiment*

**表格

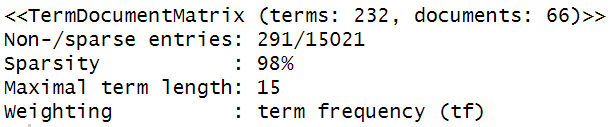
描述已自动生成**

*Figure 4.9 Table of National Party’s Sentiment Score to Show the Total Scores of Each Sentiment*

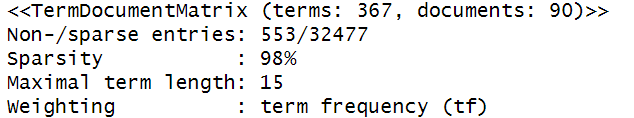
**Stage 5: Create TDMs**

The aim of this stage is to transfer text data into a matrix for further analysis. A TDM is generated to extract words between 5 and 15 characters in length, strip stop words, and set all the data to lower case by using TermDocumentMatrix() function. The TDM is a matrix in which the rows represent words and columns represent documents. Each cell shows the frequency of occurrence of a word in a document. TDMs of the three parties are generated separately.

The TDM of Greens Party (*Figure 5.1*) shows it contains 232 words and 66 documents. These words are the most frequent words. The information of Non-/Sparse Entries means 291 unique term-document pairs have non-zero frequency counts. The sparsity is a ratio. High sparsity means that most documents have zero frequency counts, suggesting that many words do not appear in many documents. The maximal term length means the longest word in the corpus has a length of 15 characters. The weighting means the frequency of each word in each document is the weight considered. The similar interpretation can be applied to the TDMs of Labour Party and National Party shown in Figure 5.2 and Figure 5.3. Since same control requirements are use the TermDocumentMatrix() function, the values of sparsity and maximal term length are same.

****

*5.1 TDM of Greens Party*

****

*5.2 TDM of Labour Party*

**文本

描述已自动生成**

*5.3 TDM of National Party*

**Stage 6: Create Word Clouds**

After creasing TMDs, word clouds can be created, in which the minimum frequency is 50. By using word clouds, the most frequently occurring word can be displayed. While generating the word cloud, there are warnings. We can use warnings() function to show the details. The detailed information can help to make stop words. The figure 6.1, figure 6.2, and figure 6.3 below are the word clouds of the three political parties.

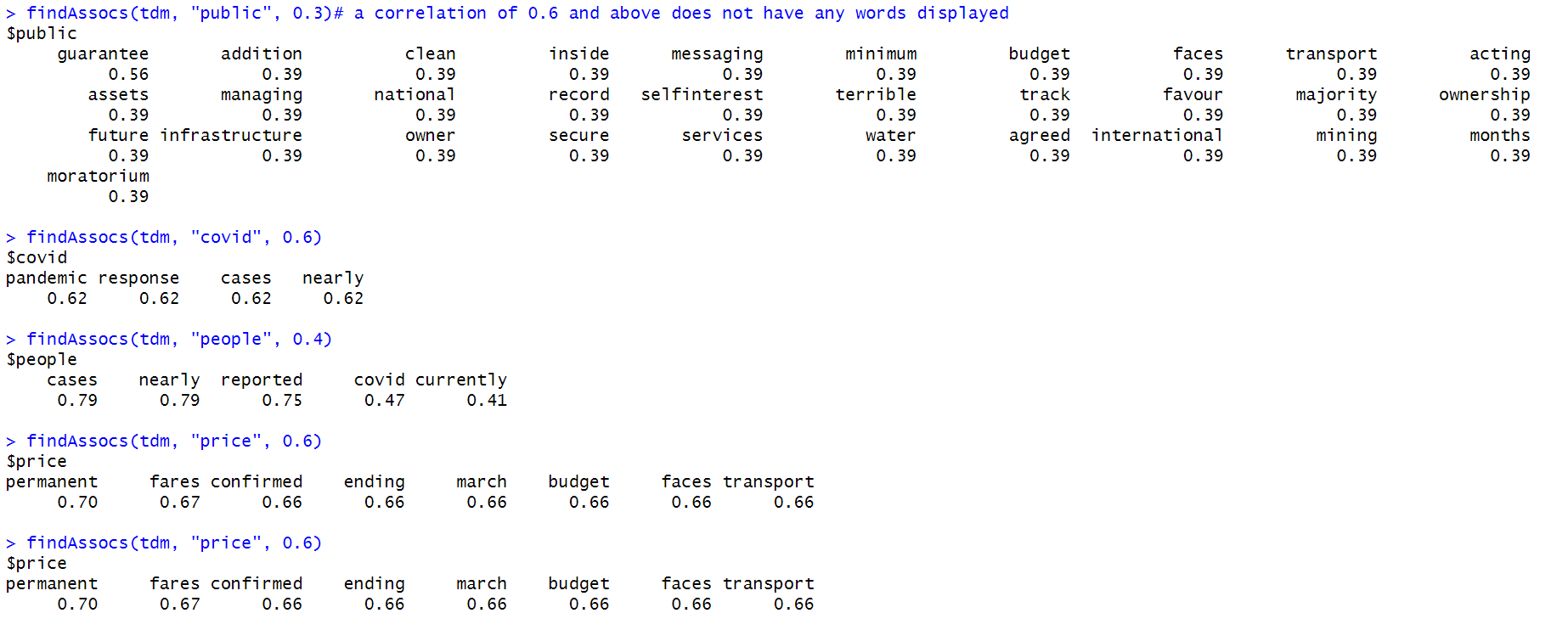
From the word cloud of Greens Party (*Figure 6.1*), the highest frequently occurring words are “public”, “covid”, and “people”. From the word cloud of Labour Party (*Figure 6.2*), the highest frequently occurring words are “people”, “support”, and “covid”.

From the word cloud of National Party (*Figure 6.3*), labour and labours are same words. So stemming should be used to trim these words. Word cloud after stemming is generated as shown in Figure 6.4. The highest frequently occurring words are “labour”, “living”, “emission” and “inflation”.

|  |  |
| --- | --- |
| 图片包含 图示  描述已自动生成  *Figure 6.1 Word Cloud of Greens Party* | 图示  低可信度描述已自动生成  *Figure 6.2 Word Cloud of Labour Party* |
| *Figure 6.3 Word Cloud of National Party*  *(without stemming)* | *Figure 6.4 Word Cloud of National Party*  *(after stemming)* |

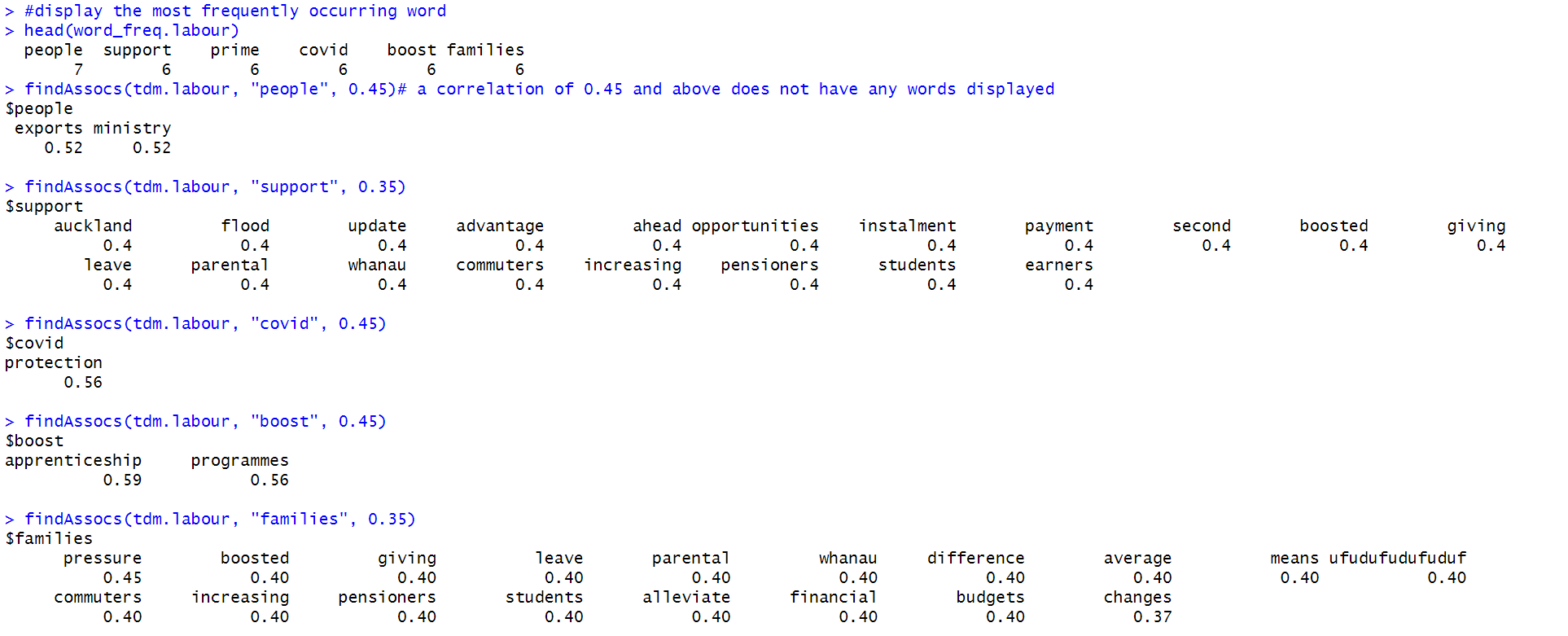
Besides the word clouds, the correlation among different words can also be displayed based on the desired threshold for association strength using findAssocs() function. Normally the number 0.6 will be chosen to measure the minimum association. However, in Figure 6.5, Figure 6.6, and Figure 6.7, lower thresholds are in use. The reason is no value returns when applying 0.6. One possible explanation to this situation could be the data size is too small.

From Figure 6.5, we can see that, the high frequency word “public” associates with "guarantee" "addition" "clean" "inside" and "messaging" and others, each with a correlation coefficient of 0.39 or higher. However, no strong correlation of 0.6 or above is returned. The word "covid" shows stronger associations with terms related to pandemic. The word "people" also shows associations with content related to pandemic reports.



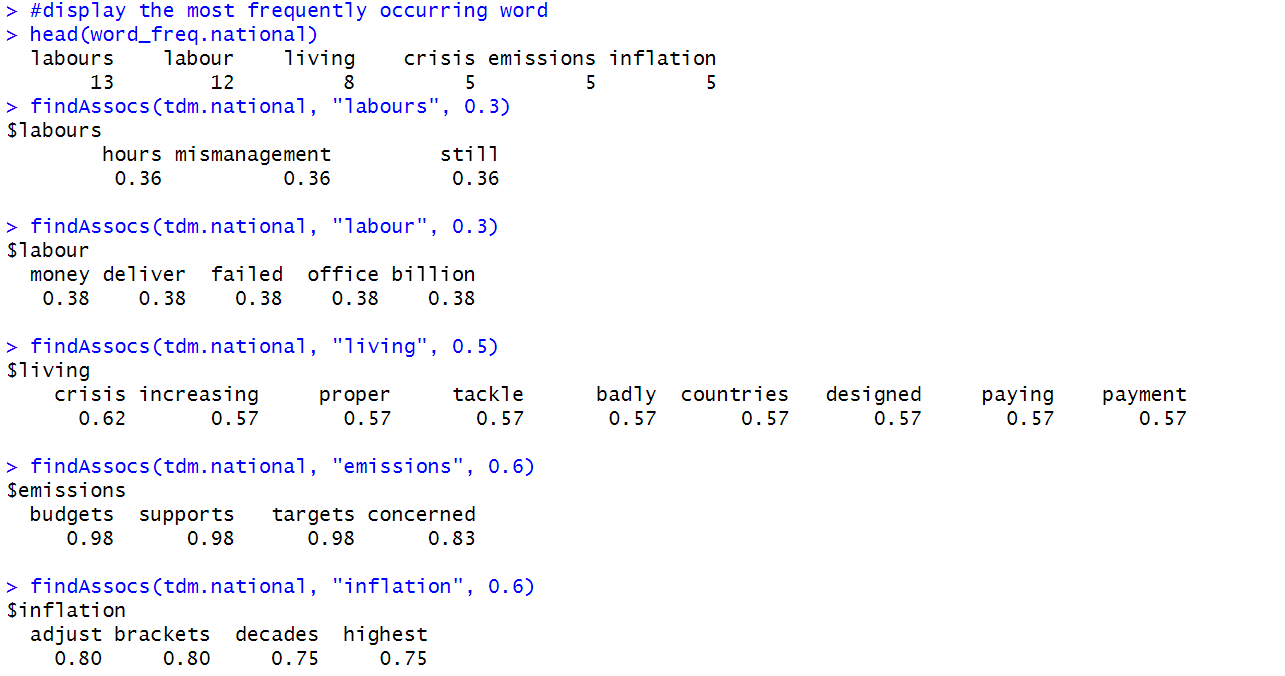
*6.5 The Correlation among High Frequency Words of Greens Party*

From Figure 6.6, the word “people” demonstrates a moderate association with "exports" and "ministry". The word “support” shows associations with various topics, such as "Auckland" "flood" "update" "advantage" and "opportunities". The word "covid" is strongly associated with the word "protection", the correlation is 0.56.



*6.6 The Correlation among High Frequency Words of Labour Party*

From Figure 6.7, the word “labour” exhibits relationship to money and mismanagement. The word “living” shows associations to words like "crisis" "increasing" "proper" and "tackle", which indicates the discussions about dealing with living crisis. The word “emissions” shows strong associations to "budgets" "supports" "targets" and "concerned", indicating the topics about emission concerns and budget problems. The word “inflation” shows strong associations to words "adjust" "brackets" "decades" and "highest", indicating issues about adjusting inflation and concerns about high inflation.



*6.7 The Correlation among High Frequency Words of National Party*

**Conclusion**

In this report, sentiment analysis is placed on twitter data for the three parties in New Zealand, which are Greens Party, Labour Party, and National Party. Based on the analysis, we can see that the Labour Party receives the most positive tweets, the amount exceeds 116. The National Party has the most negative tweets, which is 69. The highest frequent words related to the Greens Party is “public”, “covid”, and “people”. The strong associated discussions are about the pandemic. The highest frequent words related to the Labour Party is “people”, “support”, and “covid”. The strong associated discussions are about topics such as the pandemic and Auckland flood. The highest frequent words related to the National Party is “labour”, “living”, “emission” and “inflation”. The strong associated topics are about living crisis, emission concerns, economy issues, and high inflation issues.