Data Science for Industry Project 2

Introduction

Exploratory Data Analysis

The data set consists of 30 State of the Nation Address (SONA) speech transcripts from all the presidents from February 1994 til recently in February 2018.

A hypothesis can be made that sentiments of speeches differ depending on the political season.There are 3 political seasons :

1. Pre-Election

2. Post-Election

3. Normal Term

Below is a summary of all the speeches as per the various presidents categorized by the 3 political seasons :

Period	Presidents	Pre-Election	Post-Election	Normal Term	Total
1994	de-Klerk	1			**1**
1994-1999	Mandela	1	2	4	**7**
2000-2008	Mbeki	1	1	8	**10**
2009	Motlante	1			**1**
2009-2017	Zuma	1	2	7	**10**
2018	Ramaphosa			1	**1**
Total		**5**	**5**	**20**	**30**

Table 1: The number of SONA per president and arranged by political seasons

From the table above the following remarks can be made :

- de Klerk,Motlante and Ramaphosa all have one speech each which will make it extremely difficult to accurately predict given the far higher number of speeches from their counterpart presidents.

- There are far more speeches done during the normal season which will inherently bias the training data towards that season

- Mbeki and Mandela dominate the number of speeches with 10 apiece.This will also inherently bias the training data towards them.

- Pre-Election speeches are evenly distributed across 5 of the 6 presidents whilst post election speeches are dominated by Mandela and Zuma.

Word Distribution

Below are the most frequently used words of all the 30 presidential speeches rescaled according to their respective political seasons :
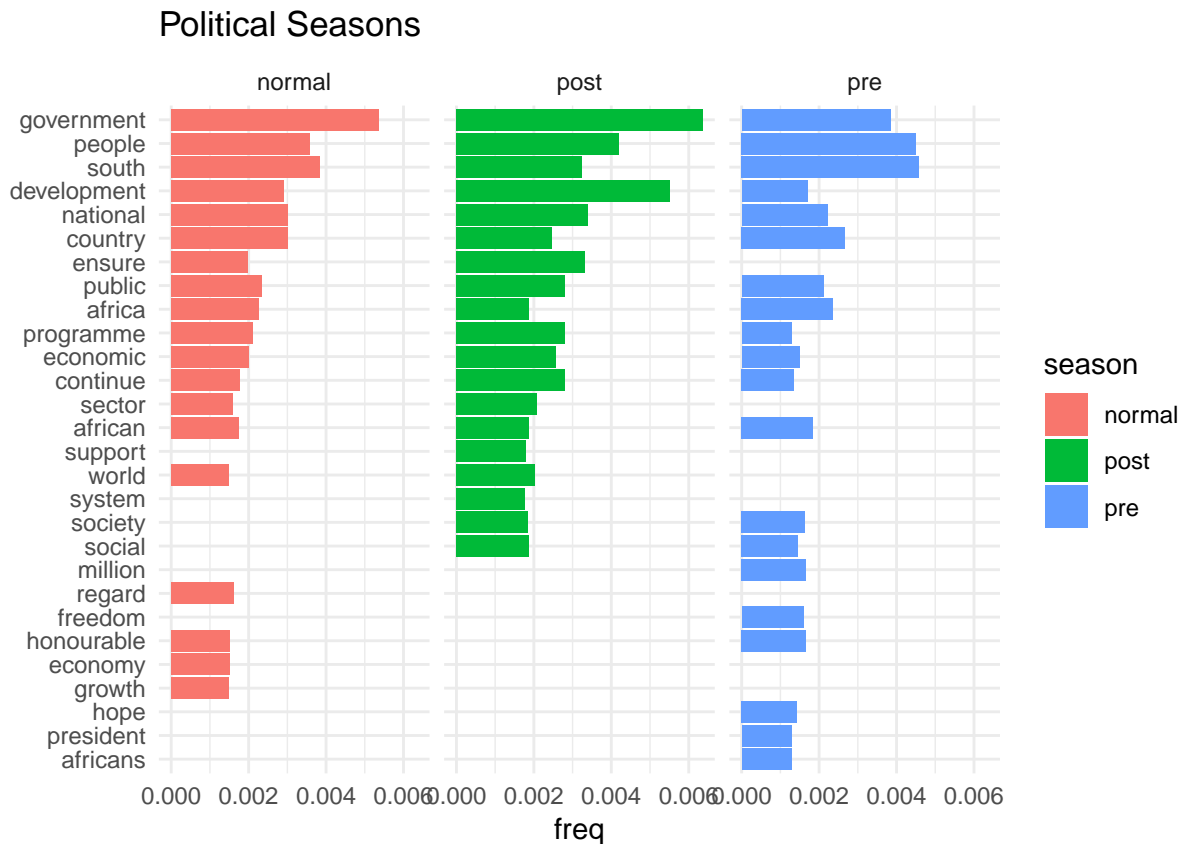

Figure 1: Frequently used words across political seasons

From the illustration above the following remarks can be made :

- Notable words commonly used across all political seasons are **south**, **africa**,**government**,**development**,**people**,**country** words will potentially not assist in distinguishing the respective presidents.

- Comparing pre and post to normal political seasons ,notable words like **economy** and **growth** are introduced into their speeches .The utilization of these words (which could imply economic growth) are understandable given that these are typical themes that need to be addressed constantly throughout the normal period of presidential terms.

- Comparing pre to post political seasons,uplifting words like **freedom**,**hope**,**africans** are used before and not after elections.

- Comparing pre to post political seasons,notable words introduced are **support**,**system**,**ensure**.These words convey a theme of action and execution which is expected after coming from an election.

Below are the most frequently used words of all the 30 presidential speeches rescaled according to the respective presidents :
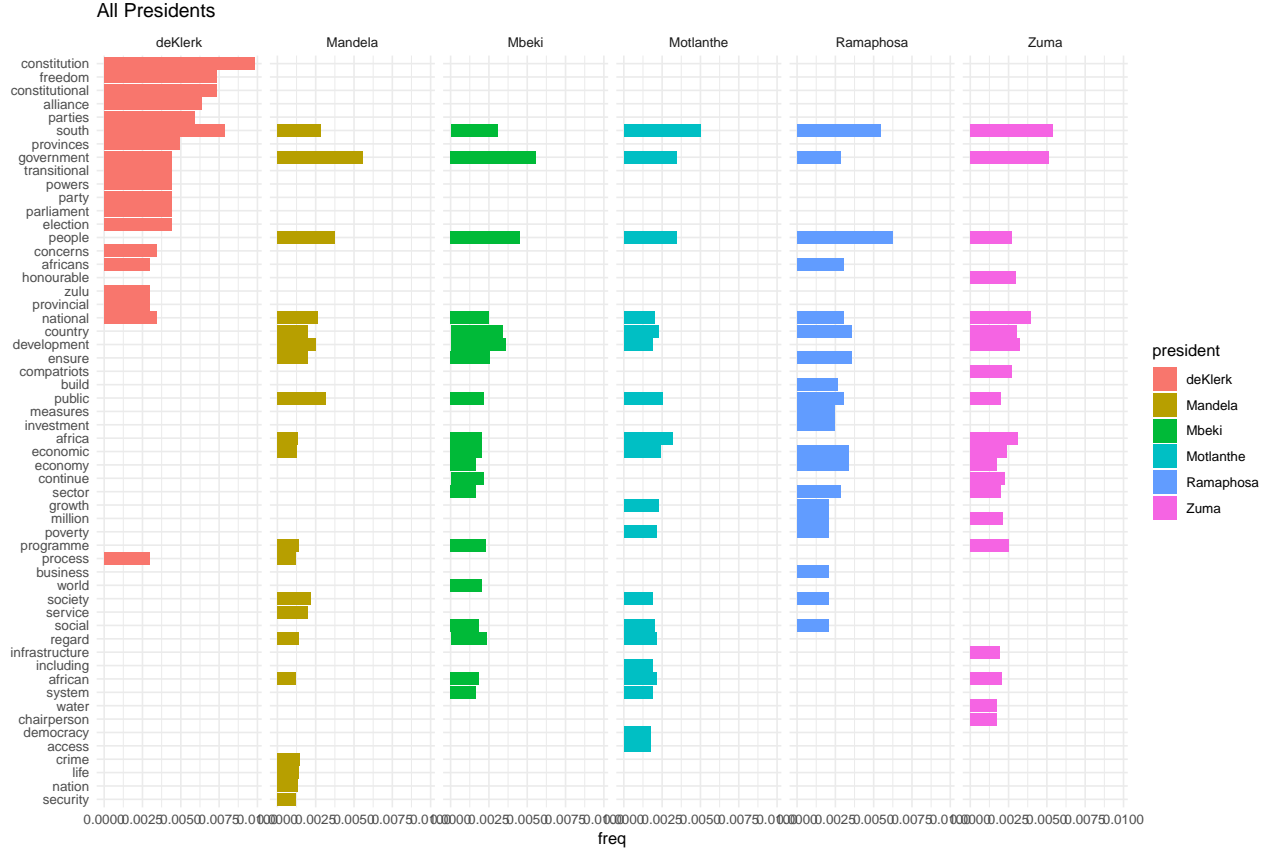
Figure 2: Top words used by all presidents in SONA

From the illustration above the following remarks can be made :

- Looking at all the presidents frequently used words ,de Klerk has the least common words.This is due to the fact that firstly there is only one speech in the dataset and secondly given that the speech was before the first democratic elections,the content will be far different to the speeches made by the presidents that preceded in the democratic era of South Africa.

- Commonly used words across all presidents are **south**,**government**,**national** which are also common words across political seasons.

- Notably words commonly used across all presidents excluding deKlerk are **people**,**country**,**public**,**ensure**,**development**

Clustering by Term Similarity

Words from the respective speeches we aggregated by the respective presidents.Words greater than 4 letters were considered so as to focus primarily on descriptive words.Sparse elements that did not reflect at least 40% of the time in the resultant document matrix were removed. The resultant word counts were then normalized to avoid biases of presidents with more speeches .

K means clustering was then conducted and a $k = 4$ was selected based on the 'elbow rule'.

The objective is to see what frequent common words do the presidents use and what potential themes to these similar words posses.The resultant visualization can be viewed below:

Figure 3: K means clustering of speech words of the 5 presidents

From the illustration above the following remarks can be made :

- Approximately 84 percent of the variability resides in the first two principal components.

- Cluster 2 has words like **south africa** ,**country**, synonymous with SONA introduction/concluding paragraphs that all presidents will generally mention.

- Cluster 1 portrays a very serious tone.The theme that comes to mind is pertinent issues that need to be addressed.Words include **concerns**,**security**,**education**,**address**.

- Cluster 3 portrays a more warm theme with a slight patriotism undertone.Notable words include **africa**,**growth**,**freedom**,**economic**,**service**.

Sentiment Analysis Across Political Seasons

The SONA of before and after the elections over the years are compared to determine whether there is an inherent tone difference.The *bing* lexicon was utilized in this analysis.The sentiment of the words in the lexicon were summed up to determine the net sentiment which will be referred to as polarity.Below are the results
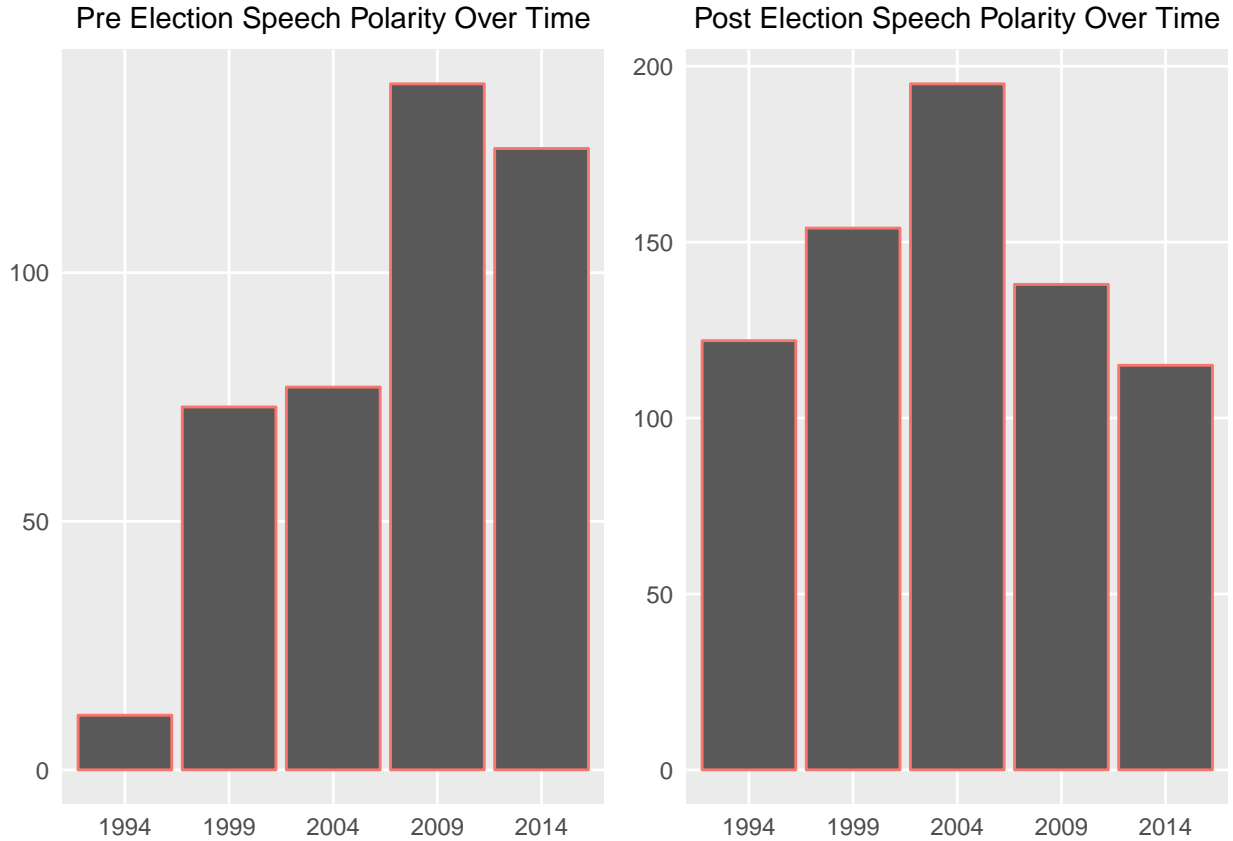
Figure 4: Pre vs Post Election Speech Sentiment over time

From the illustration above the following remarks can be made :

- There is an interesting behavior in the first 3 election years (1994,1999,2004) there was an improvement in overall sentiment after the elections,however in the more recent 2 election years (2009,2014) there has been a more cautionary tone after elections

- The biggest overall sentiment disparities before and after elections occur in 2004 and 2009.The 2004 pre/post comparison is interesting - both speeches were done by Mbeki and it showed the highest increase in net sentiment in any given year.This could be due to the fact that Mbeki had just won his second election and wanted to reassure the country with a positive speech.The 2009 pre/post comparison is counter intuitive- one would expect that given that when Zuma was elected into power for the first time that he would have a higher net sentiment then that of the his predecessor who was there on a temporary basis.

Sentiment Analysis Across Presidential Terms

Sentiment analysis was conducted on presidents with at least a full presidential term to observe where there is seasonality on with their net sentiment over their respective terms.The presidents were:

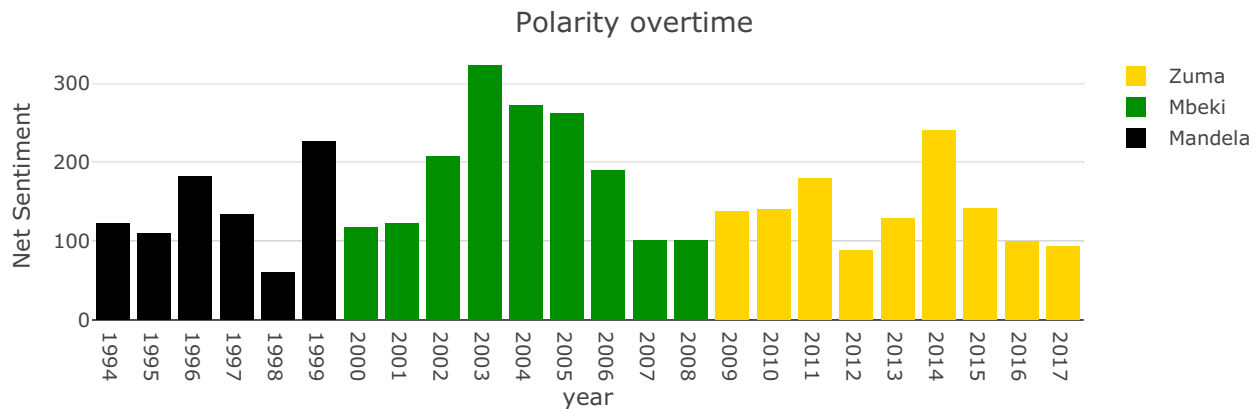1. Mandela
2. Mbeki
3. Zuma

Figure 5: Polarity of Presidents across their terms

From the illustration above the following remarks can be made :

- Net Sentiment across the 3 presidents tend to trend up in their first 3 years of service followed by a downward trend .

- The most startling change in sentiment was Mandela in his last year of service who finished off on a vary high sentiment note.

Conclusion

The objective of this section was to conduct an exploratory text analysis of the SONA speeches.This was conducted with the additional of looking at political seasons and how that influences sentinemt. 5/6 Presidents utilize the same common words which are generally topical themes in South Africa. It has been observed that there is an improvement in net sentiment of speeches after elections. Presidents with at lease one term tend to exhibit an increase in sentiment in their first 3 years followed by a decrease in sentiment.
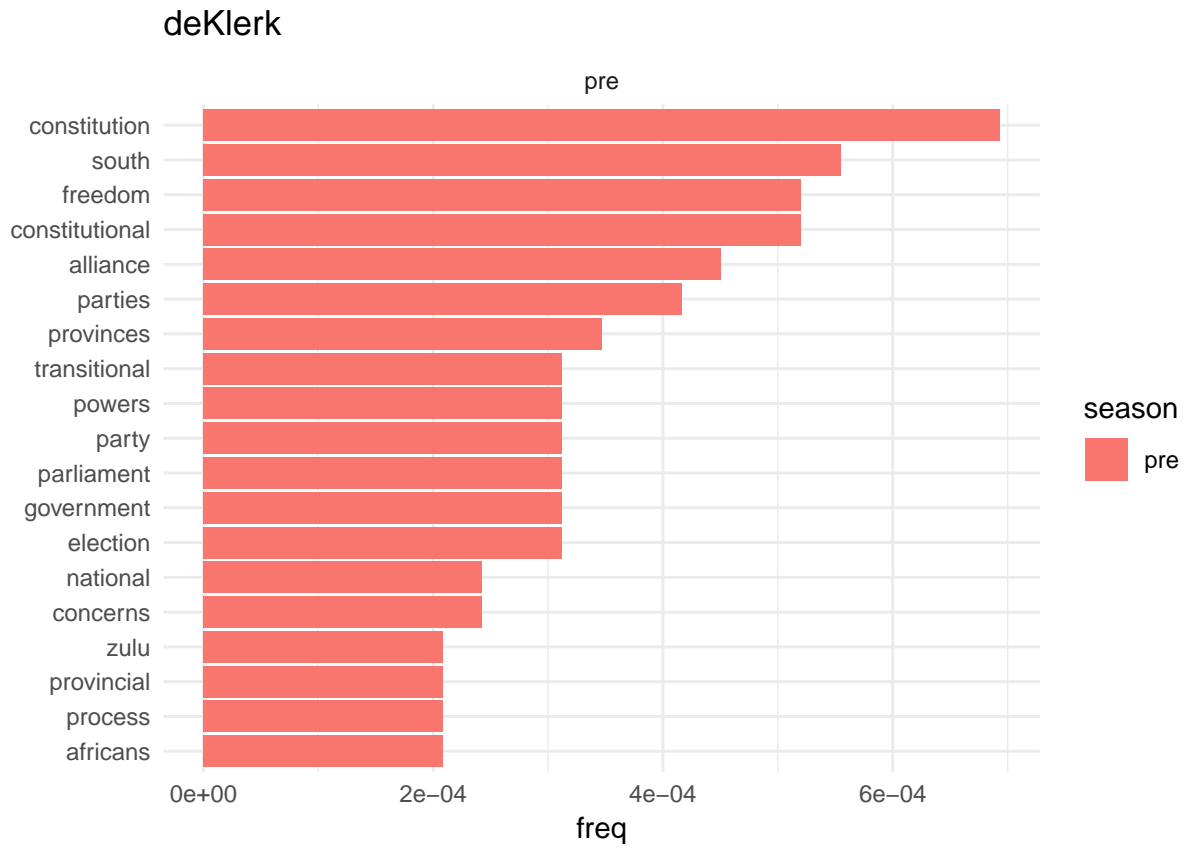
Appendix

deKlerk

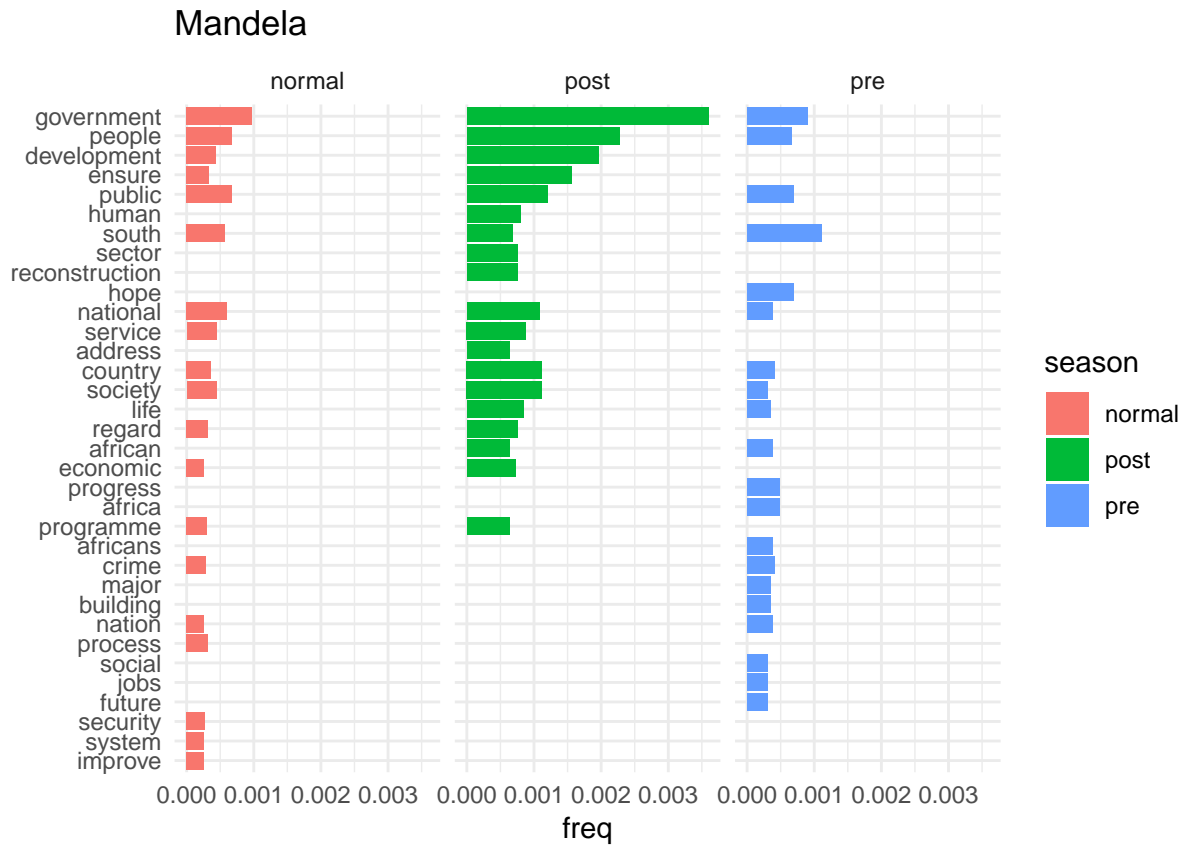
Figure 6: Top Words used by de Klerk
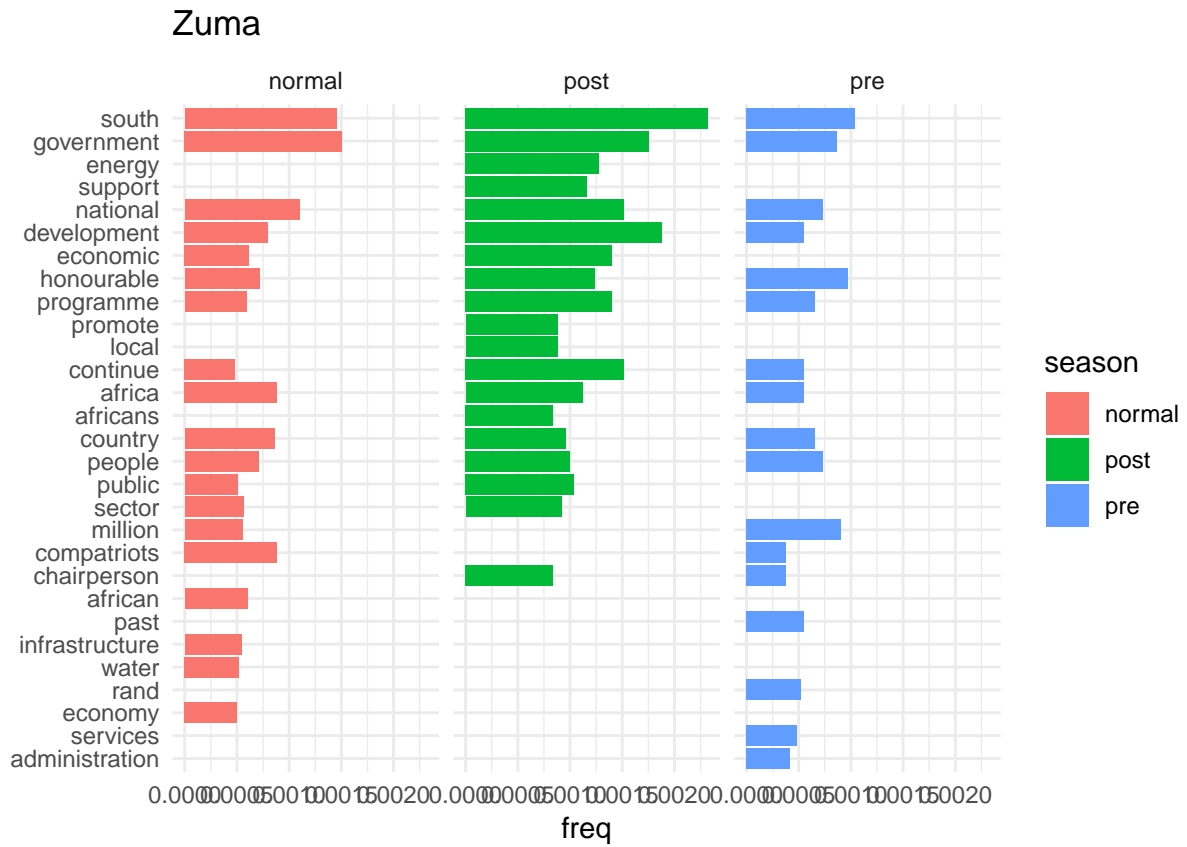
Figure 7: Top Words used by Mandela
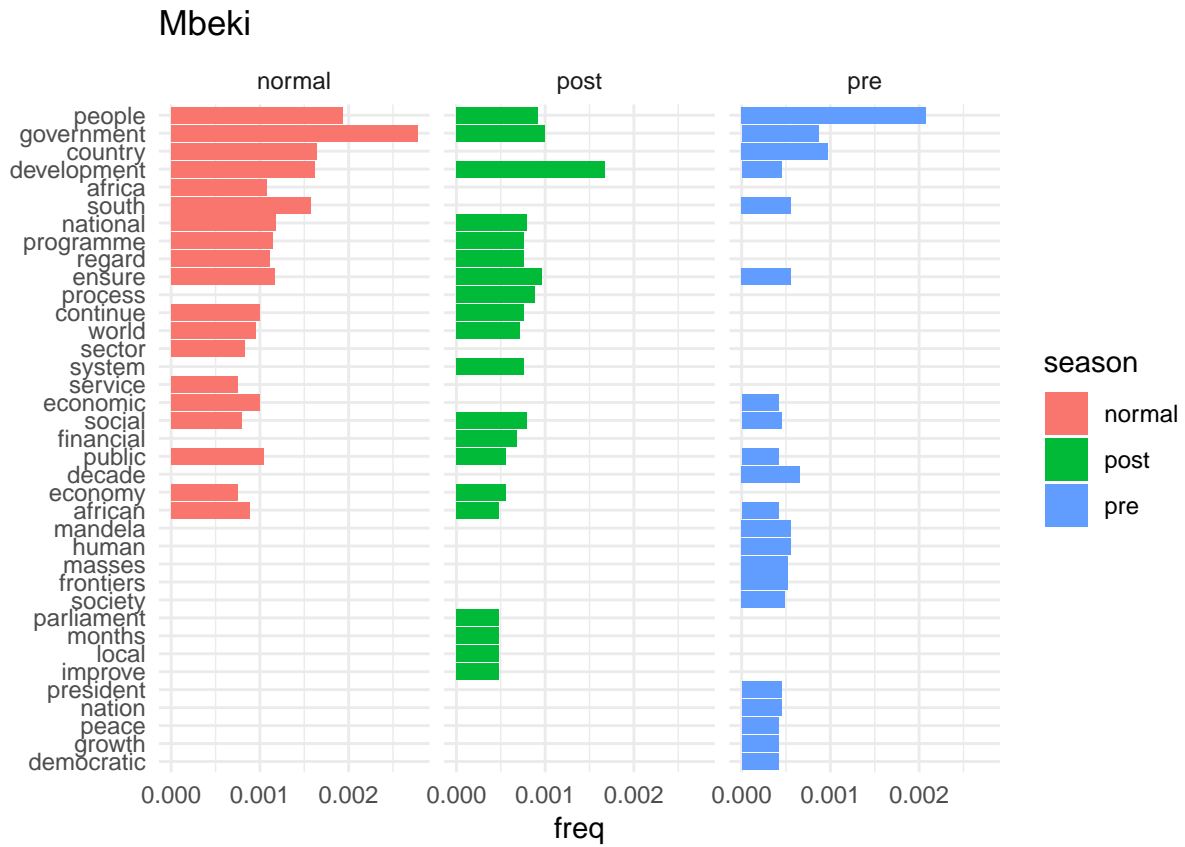
Figure 8: Top Words used by Zuma

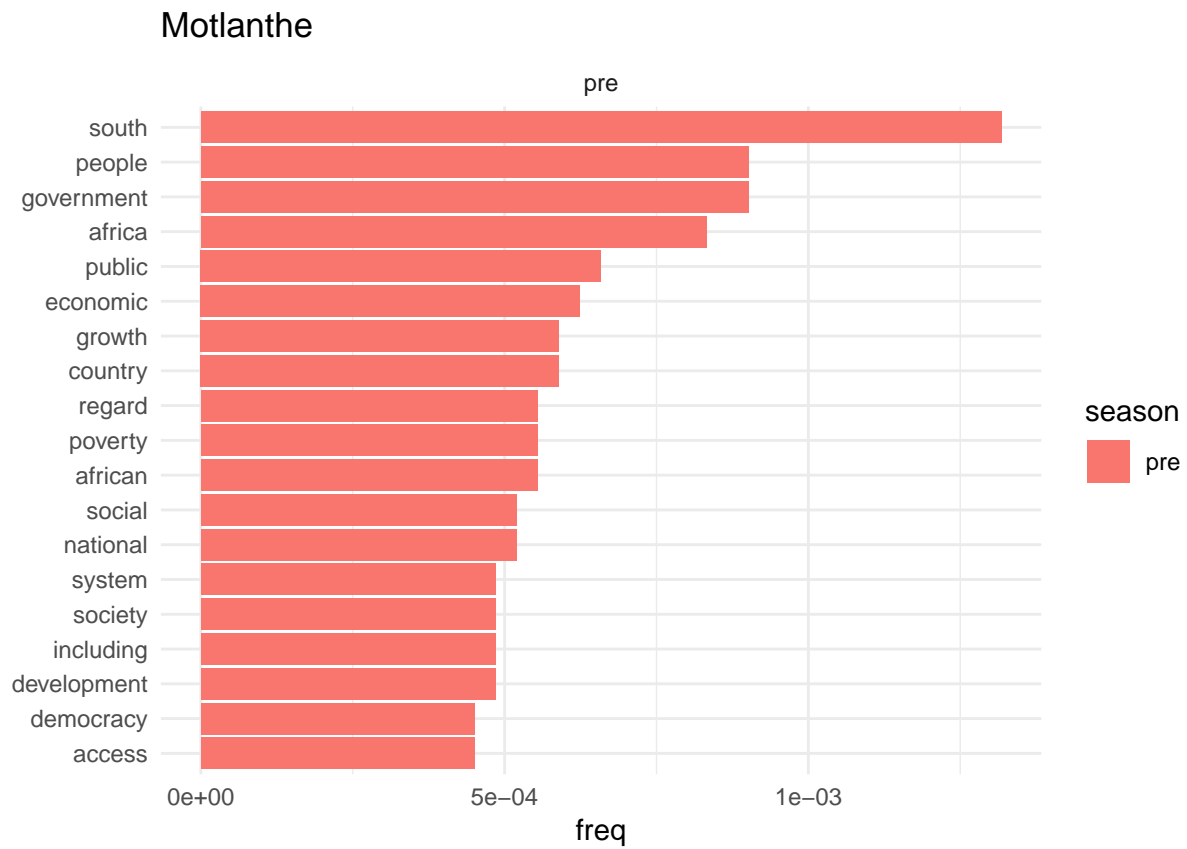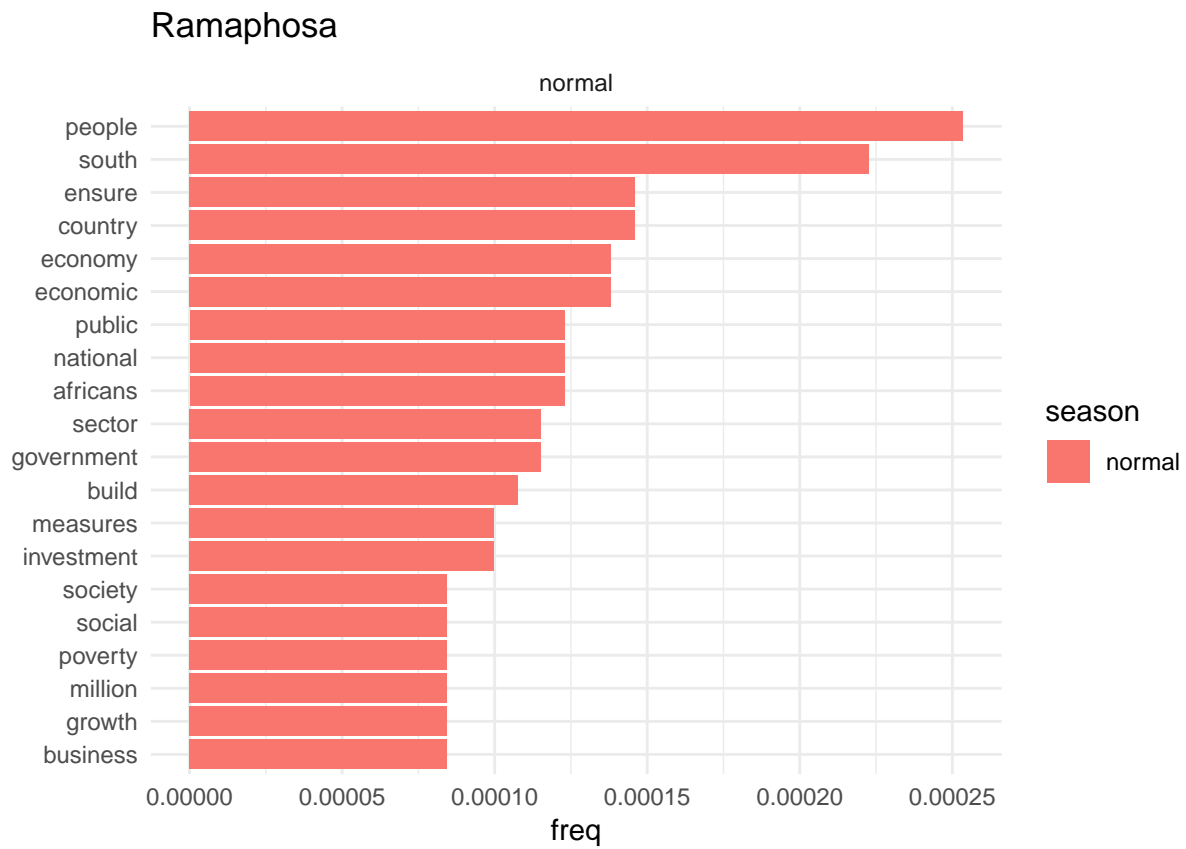Figure 9: Top Words used by Mbeki

Motlanthe

pre

Figure 10: Top Words used by Motlanthe

Figure 11: Top Words used by Ramaphosa