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 :

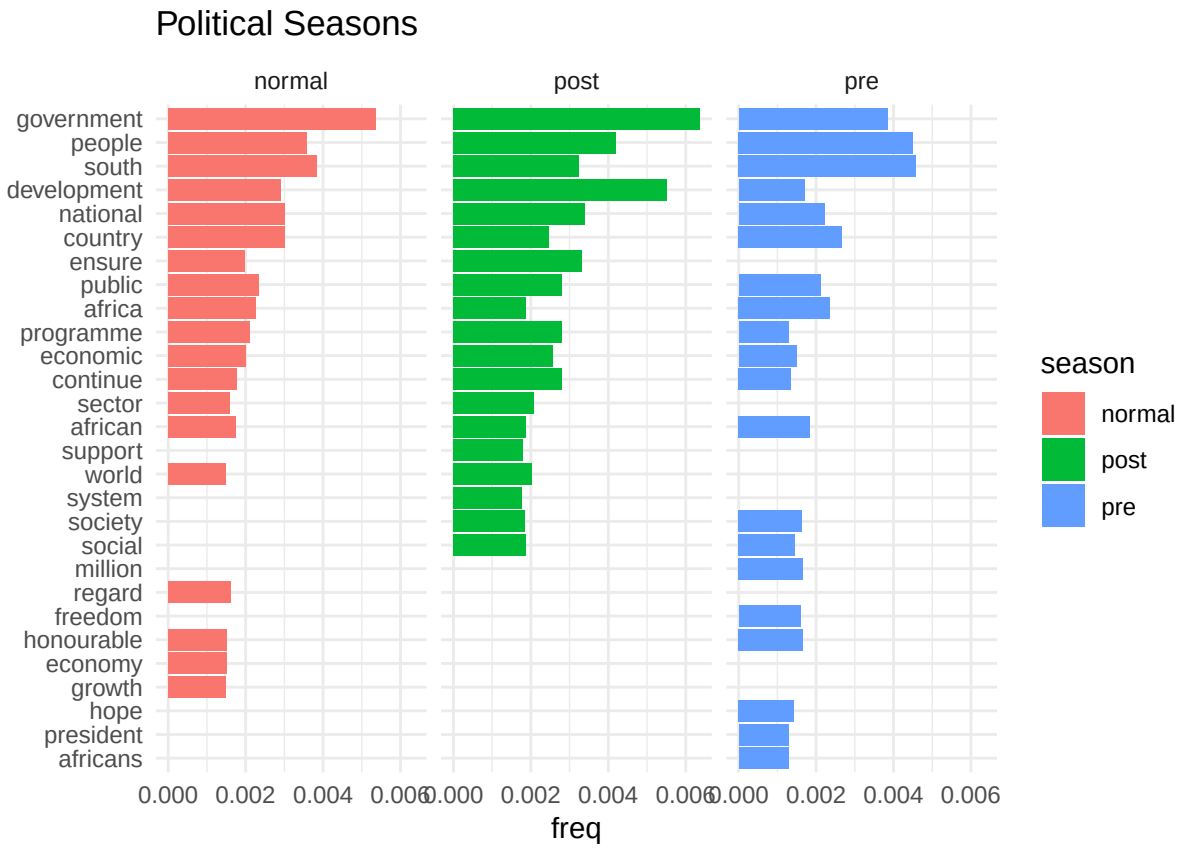


Figure 1: Frequently used words across political seasons

From the illustration above the following remarks can be made :

- Notable words commonly used across all political seasons are **south**, **africa**,**government**,**develpment**,**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 utilisation 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 :

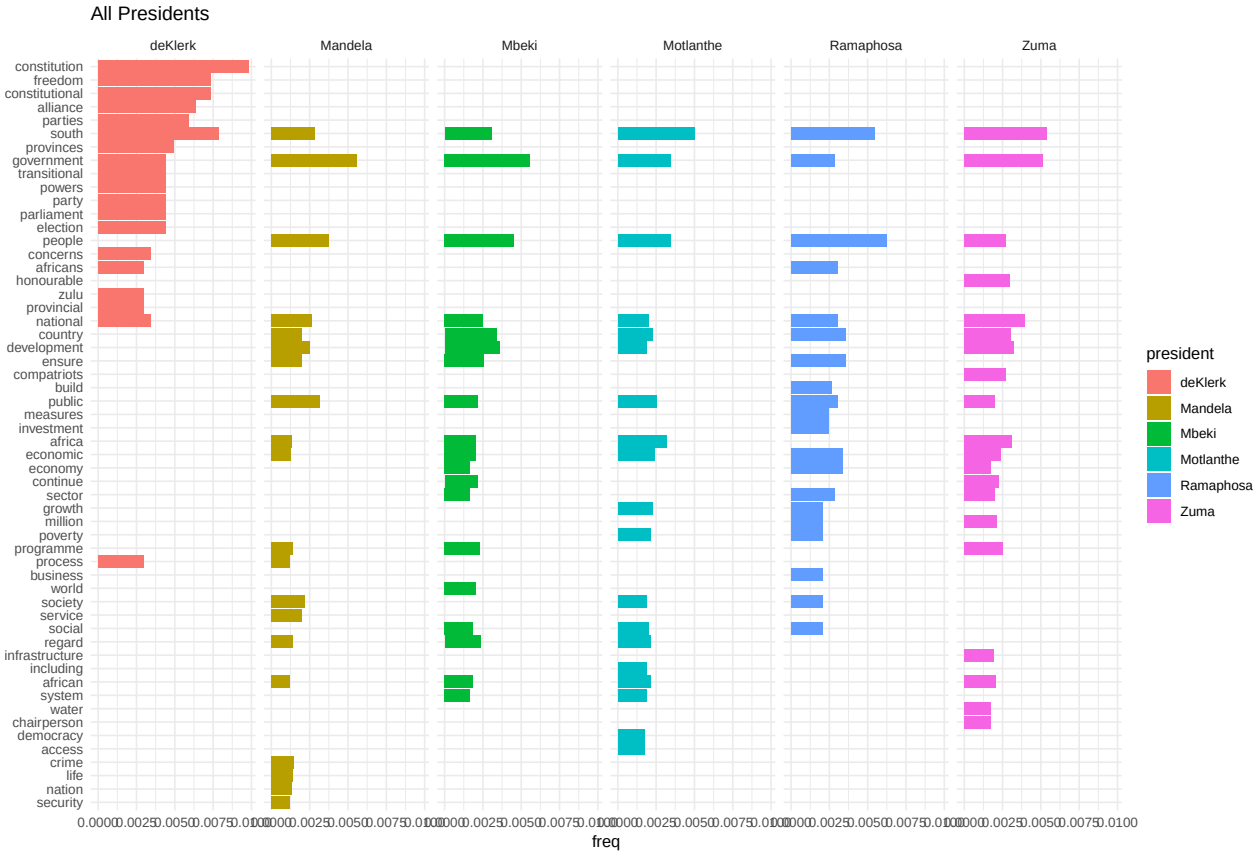


Figure 2: Top words used by all presidents in SONA

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.The resultant word counts were then normalised to avoid biases of presidents with more speeches .

K means clustering was then conducted and a $k = 3$ was selected based on the 'elbow rule'.

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

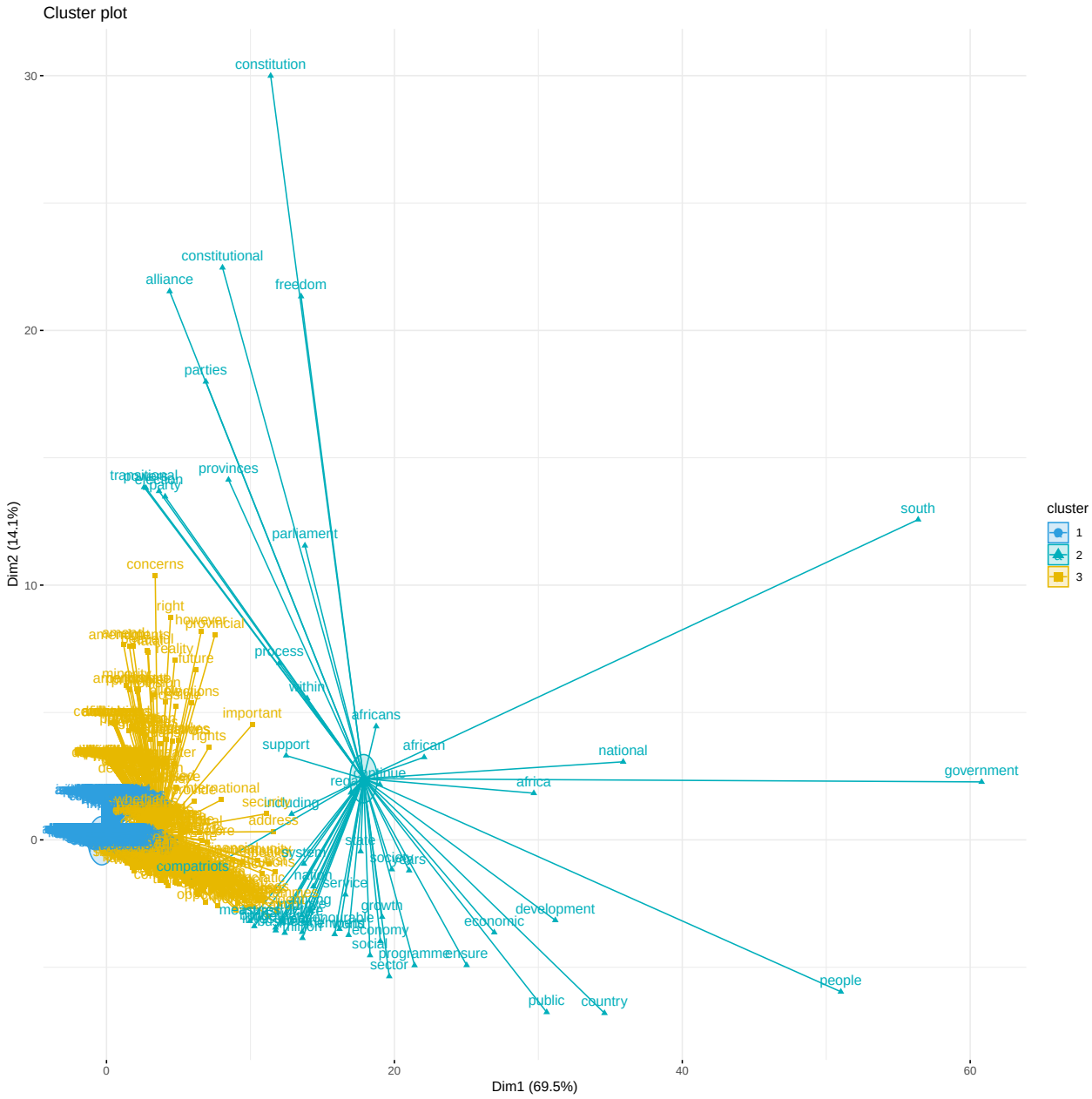
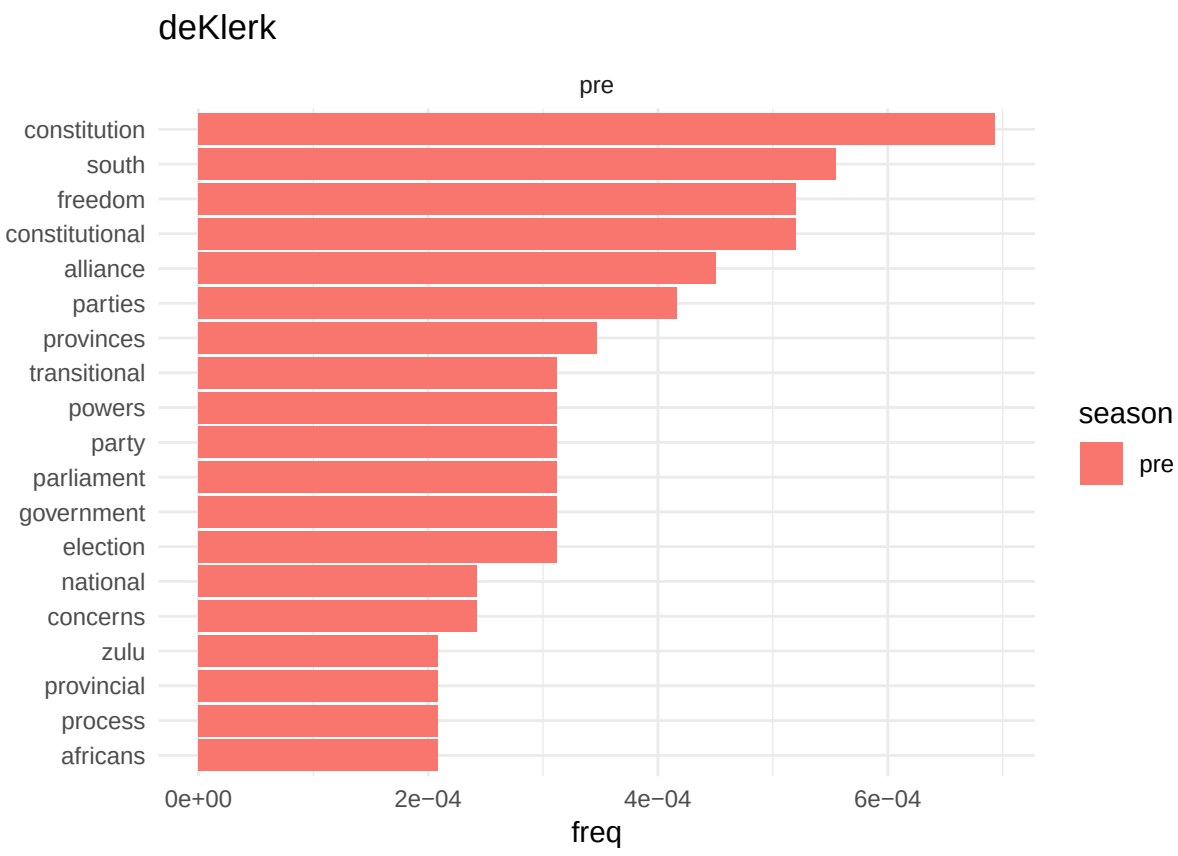
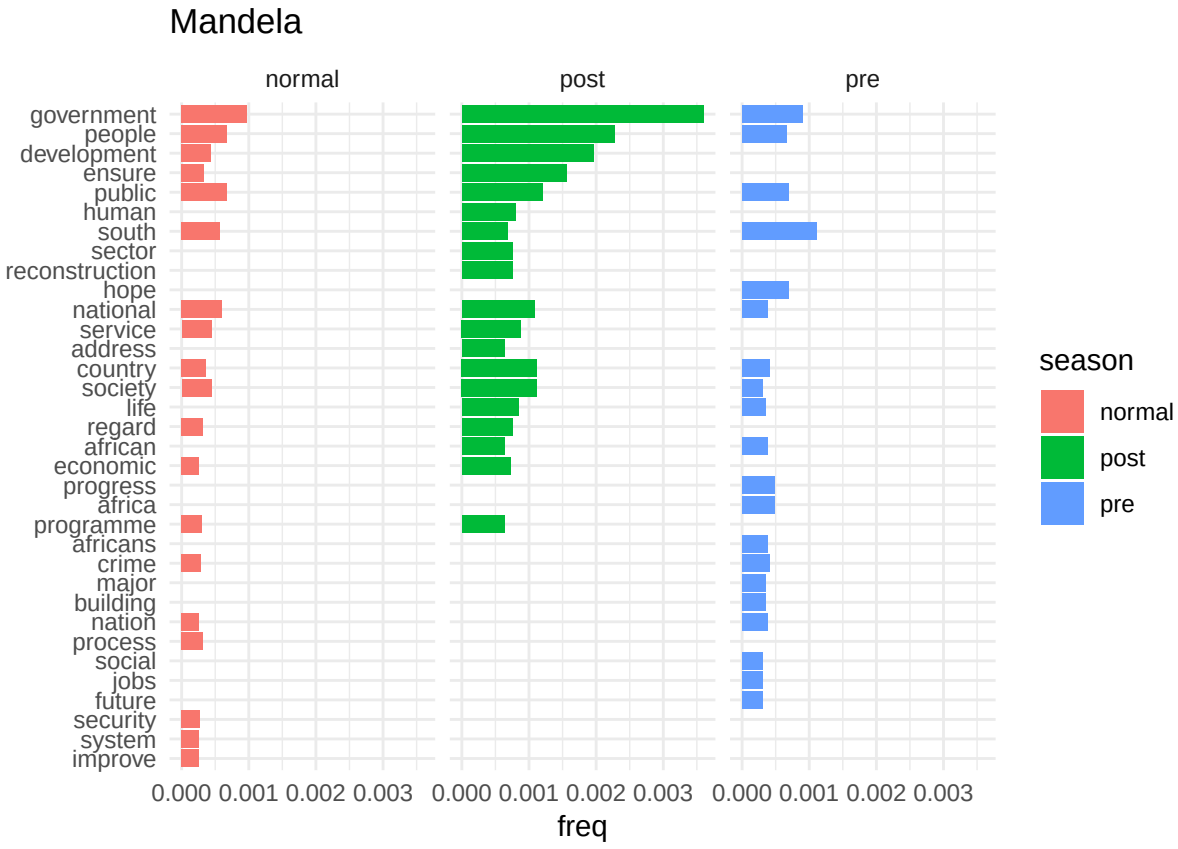


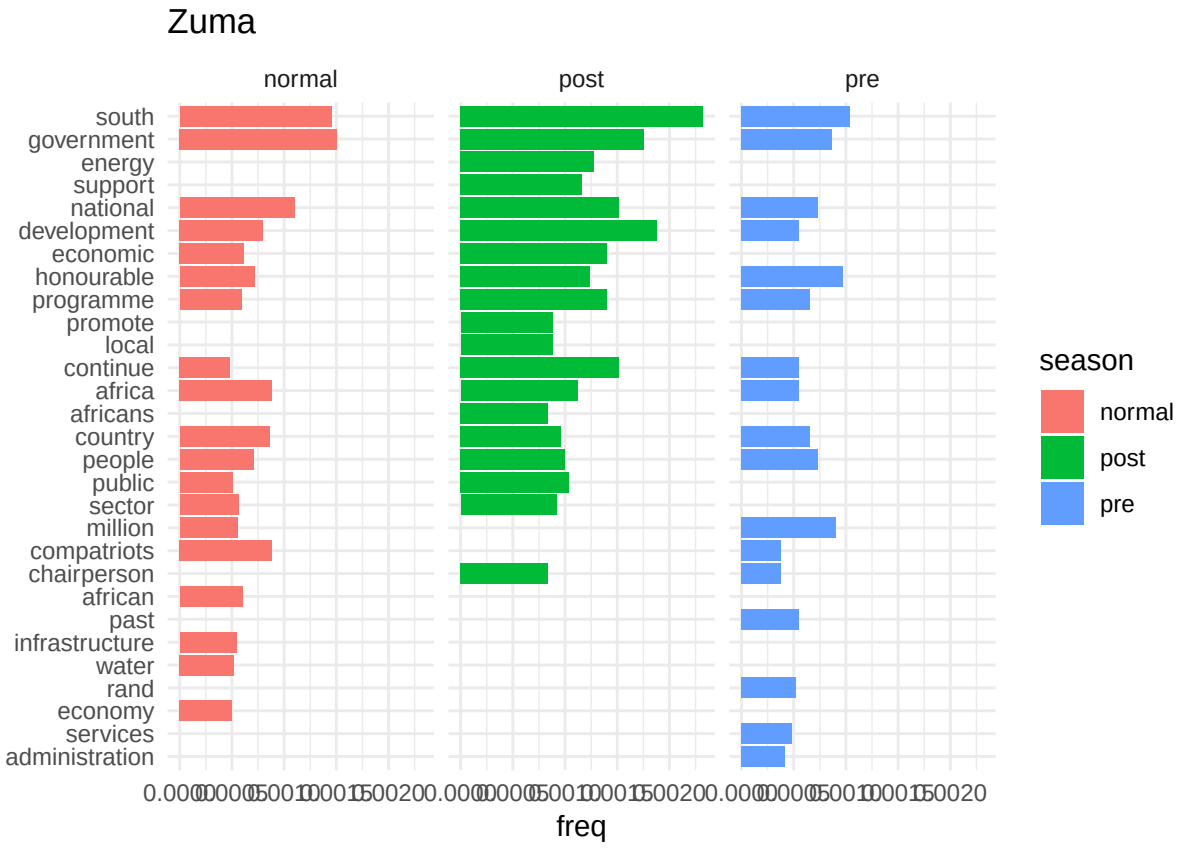
Figure 3: K means clustering of speech words of the 5 presidents

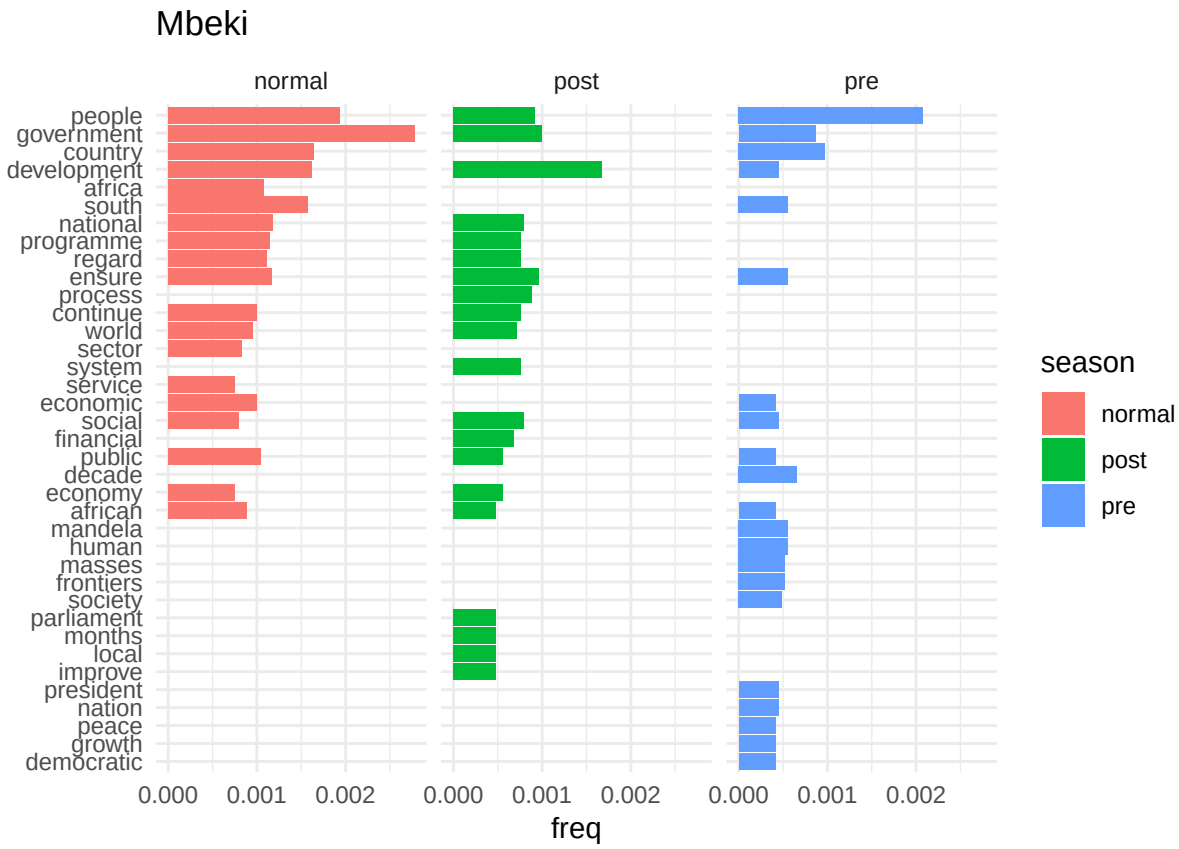
From the illustration above the following remarks can be made :

-

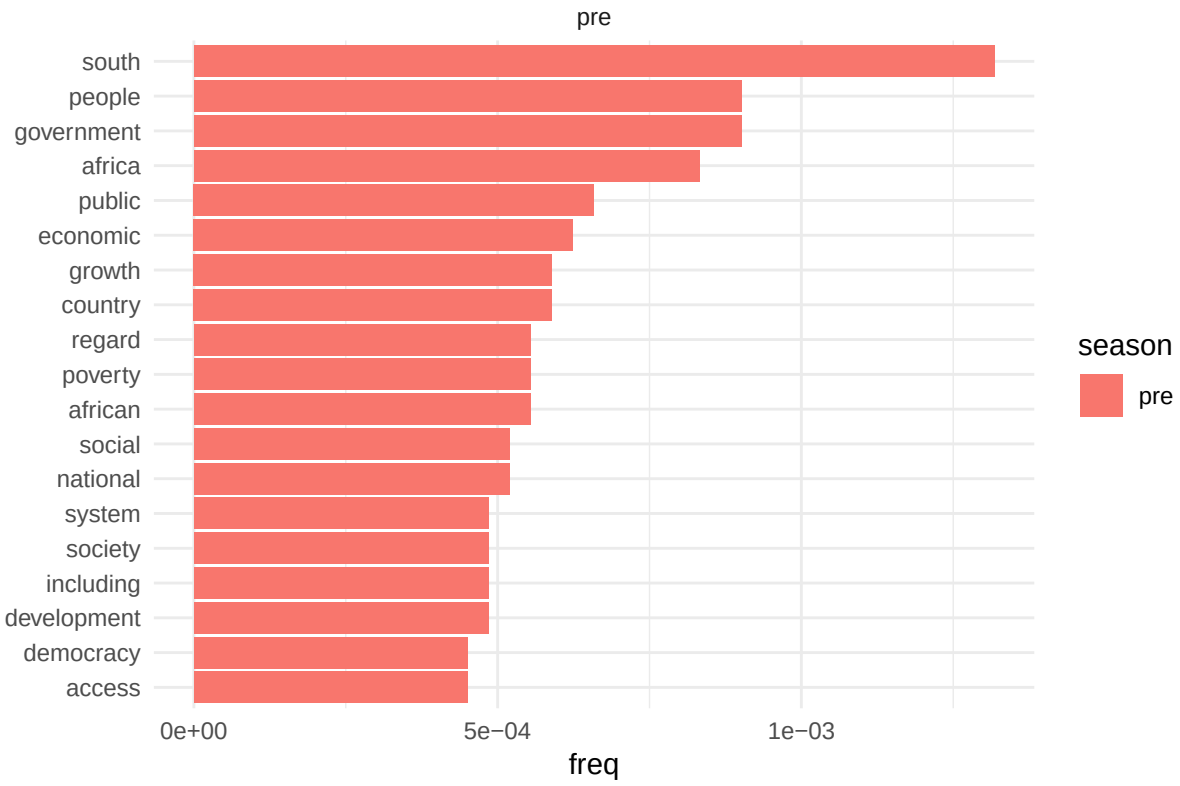
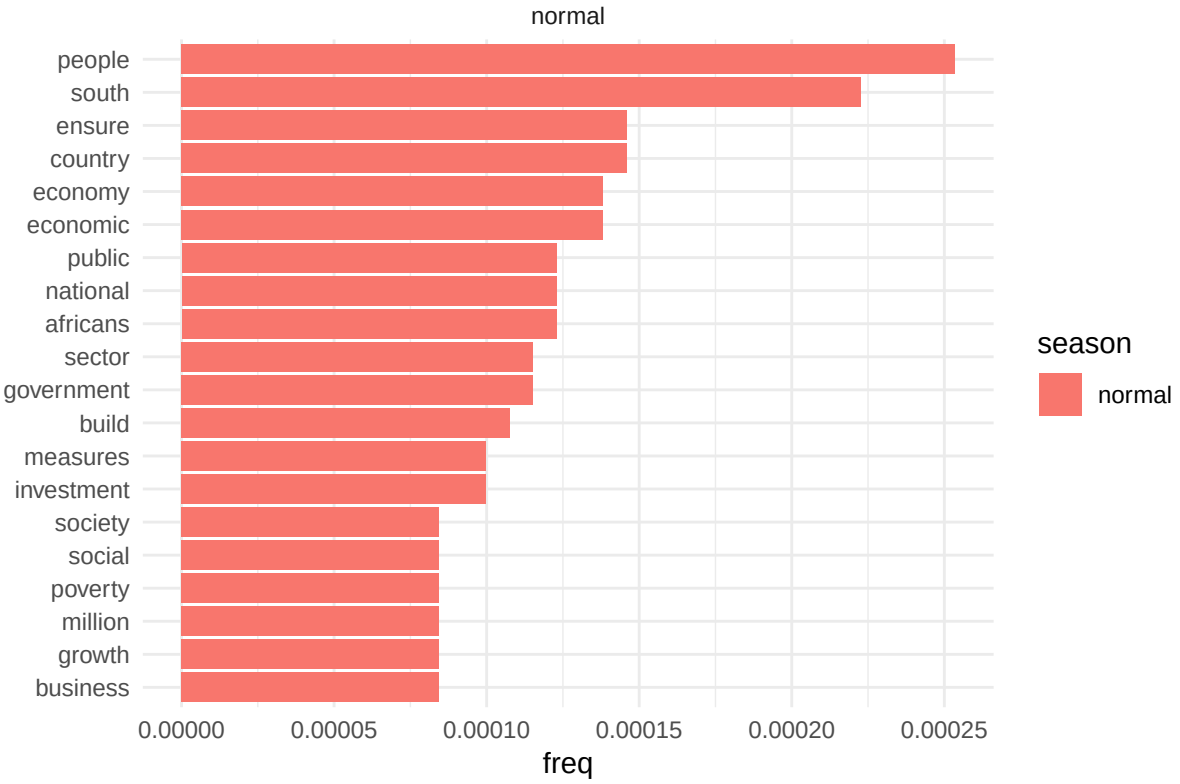

deKlerk
pre
constitution
south
freedom
constitutional
alliance
parties
provinces
transitional
powers
party
parliament
government
election
national
concerns
zulu
provincial
process
africans
0e+00
2e−04
4e−04
6e−04
freq
season
pre


Mandela
normal
post
pre
government
people
development
ensure
public
human
south
sector
reconstruction
hope
national
service
address
country
society
life
regard
african
economic
progress
africa
programme
africans
crime
major
building
nation
process
social
jobs
future
security
system
improve
0.000 0.001 0.002 0.003
freq
season
normal
post
pre


Zuma
normal
post
pre
south
government
energy
support
national
development
economic
honourable
programme
promote
local
continue
africa
africans
country
people
public
sector
million
compatriots
chairperson
african
past
infrastructure
water
rand
economy
services
administration
freq
season
normal
post
pre


Mbeki
normal
post
pre
people
government
country
development
africa
south
national
programme
regard
ensure
process
continue
world
sector
system
service
economic
social
financial
public
decade
economy
african
mandela
human
masses
frontiers
society
parliament
months
local
improve
president
nation
peace
growth
democratic
0.000
0.001
0.002
freq
season
normal
post
pre

Motlanthe



Ramaphosa