TEXTUAL ANALYSIS OF THE PRESIDENT BUHARI'S SPEECHES

Chimdimma Noelyn Onah

Executive Summary

In politics, words are an integral part, with officials and citizens using words to express opinions, make proposals, and defend their actions. This report is motivated by this and the recent advances in text mining and analysis, it aims to identify if the President of Nigeria's campaign policies were consistently reflected on when he delivered speeches. An attempt to classify this speeches into homogenous groups is also made.

Considering these goals, the bag of words text mining method and unsupervised machine learning clustering model was adopted for the analysis. It was found that three key policies were made and reflected on during the President's inauguration and the three key policies were consistently referred on when subsequent speeches were made. This meant that the clustering algorithm did not identify very unique groups, 4 out of the 6 clusters had one speech as a member. The other two had 34 and 9 members, with the smaller group being more homogenous than the larger group but both having similar themes.

Contents

1. Introduction

In March 2015, Nigeria had its fifth election and the results were the first in which a ruling party was elected out of office. It is hailed as the most credible, freest and fairest election in Nigeria's fourth republic and also ranked among the most issue-driven contests in the history of elections in Nigeria (CDD, 2017). An agreement by the then President Jonathan Ebele Goodluck of Peoples Democratic Party (PDP) and now President Mohammed Buhari of All Progressives Party (APC) was to conduct an issue-based campaign which meant that they both set clear policies via their manifesto.

The president was elected on three key priorities; fighting corruption, growing the economy and increasing security. These resonated with voters across ethnicities and age, as these issues were what were needed to tackle the state of the country at the time. With high rate of terrorist attacks by Boko-Haram, an economy grounded by eroding oil prices and institutionalized corruption. Evidenced in the rescheduling of the election, moving from February 2015 was to March 2015 due to insecurity in the north-eastern part of the country where Boko Haram had a stronghold. As Ojo (2015) put it the plummeting oil prices since December, 2014 had presented new challenges for Nigeria's economy which was an insidious deciding factor in the 2015 election. Nigeria has been consistently perceived as a corrupt state, rating of 0 (most corrupt) and 100 (clean), sees Nigeria averaging 20 points between 1999 and 2015 (Trading Economics, 2018).

Though both candidates had manifestos based on tackling these issues and more, Buhari won by playing up his military background as a tool to tackle the insurgency, taking advantage of the high perception of corruption associated with the then ruling party, PDP, and vowing to fight it. With the major campaign quote of 'Change', which resonated with Nigerians who had witnessed the corrupt PDP rule for over 15 years (Zane, 2015). To check Buhari's policy consistency, this analysis would compare and contrast his inauguration speech with those he made over the course of his presidency. The research questions are

1. Are the key campaign promises reflected on in the speeches made?
2. Can these speeches be grouped into unique groups/themes?

To tackle these questions, the text mining and analysis approach which typically sits within the larger Natural Language Processing (NLP) area would be applied. Liddy (2001) defines NLP as

> "…is a theoretically motivated range of computational techniques for analysing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications."

The application of this approach would be introduced in Section 2, with Section 3 used to discuss the data used. The analysis and results would be discussed in Section 4 and Section 5 of this report would conclude and make recommendations.

2. Methods

Text mining and analysis seeks to utilise methods from NLP to help turn unstructured data into a structured form and thereby allows for the application of analytical methods in conjunction with visualisation to help explore and analyse textual data sources (Clough, 2018). To ensure this application actually mimics human understanding of text, it intersects with other disciplines, including but not limited to sociology, communications, history, literary study, math, logic, cognitive science, and computer science (Froehlich, 2018). It can therefore be seen as a methodological approach which dictates the entire text analysis life cycle. For this analysis, the open source

programming language R would be used to implement algorithms throughout the identified lifecycle to answer the research questions. This lifecycle adopted is summarised in Figure 1 below.

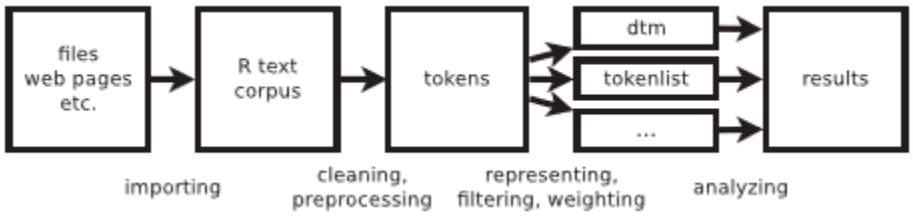


Figure 1: Text analysis lifecycle
Source: Welber et al., 2017

a. **Importing data:** To perform any analysis, data has to be collected and imported for further processing. In text analysis, a corpora (a collection of corpus) is often used. A corpus is a collection of written text especially if it is by an author or about a particular body of knowledge. Without access to any corpus of Buhari's speeches, this report includes a step to identify, extract and turn the speeches into a corpus.
b. **Clean and pre-processing text:** One of the challenges of unstructured data is the effort needed to get it to a form suitable for any subsequent analysis. The cleaning and pre-processing of data is therefore paramount, for example, to ensure valuable inferences are drawn it is common to remove stopwords (and, or, the, my….) and punctuations. The text may contain erroneous characters which do not add value to the analysis and thus needs removing. Unstructured data should be converted into a structured from (most commonly used for text analysis is a document term matrix) before analysis can commence. This stage requires utmost care as final results would be influenced by the pre-processing performed. In summary, text pre-processing includes tokenization (turning body of text into words), filtering (removing unnecessary words), normalization (through stemming or lemmatization which aims to adopt one word for words that have the same meaning) and transformation into a document tem matrix (DTM).
c. **Performing data analysis:** Boumans & Damian Trilling (2016) proposed three approaches of text analysis: counting and dictionary methods, supervised machine learning, and unsupervised machine learning. Counting and dictionary methods includes summary statistics (highest occurring words, co-occurring words), identifying how often certain concepts occur and performing sentiment analysis. Supervised machine learning analysis is a deductive approach which assume that there is apriori knowledge, for example identifying spam emails requires that some spam emails are known and labelled which are then used to train a model to detect other spam emails from a corpus of emails. Unsupervised machine learning is inductive where the computer algorithm itself somehow extracts meaningful codes from texts for example clustering documents into similar themes based on common words. This analysis would perform both the dictionary and unsupervised methods.

3. Data

Text mining in recent years has gained a lot of attention due to high volume of unstructured information produced and the advances made in increasing the capability of processing these information. The volume of production is mainly driven by the wide use of internet, high capacity for electronic and cloud storage systems, this has encouraged the creation of textual data in the health sector, via social media, news outlets and so much more. These textual data are easily

processed and perceived by humans, but is significantly harder for machines to understand. However, the volume of text necessities development of algorithms with increased capability of processing text (Allahyari, M. et al, 2017).In recent years, text mining algorithms have seen rapid advances due to technological advancements, many are open source and thus can be applied to discover patterns.

Text and speeches in particular has always been an important part of political analysis, with the accessibility of machine learning techniques to analyse vast amount of text as a corpora of an author, subject specific, country specific or in which ever context is on interest. In this report, the data analysed are President Buhari's speeches, this would be restricted to 46 speeches found on the official website (http://statehouse.gov.ng/) and the inauguration speech found in Channels Television Website – one of Nigeria's top reputable new channel (https://www.channelstv.com/). These speeches were made in various events, conferences and appearances by The President and are all in active voice, all press releases made on behalf of The President were not included in the analysis. Though he was inaugurated into office in May, 2015, aside from the National Executive Council (NEC) inauguration meeting on June 2015 found in the official government website, the site has only speeches from October, 2017. Thus, the speeches used in this analysis are limited in timeframe which are between 20th October, 2017 and 17th July 2018. This may restrict the output of this analysis, in that it does not truly capture the change (or no change) in the language used by The President as the February, 2019 election in which he is seeking re-election draws closer.

3.1. Data Collection

The text used here were published in the official website in form of html web pages which means there is no simple download button and copy-paste would be cumbersome. The download functionality provided was tried but produced text with a lot of redundant information for example, each page had the title of the article repeated, this would increase effort to clean the data and which may increase the chances of incorporation scrupulous text in analysis. To ensure as clean data as possible is extracted, web scraping was implemented to automatically collect the required text from the website and store as text document.

rvest an R package by Wickham (2014) that makes it easy to scrape data from html web pages is used to collect these speeches. This is used as it can handle many internet connection types such as https (which is what is applicable here) which may be problematic for other packages and functions. The *read_html()* function in particular is used to parse the web page and extract only html tags (content) of interest which means less data cleaning effort.

A simple syntax as below can be used to get a speech from a hyperlink. This can then be developed to go through a list of hyperlinks, gathering the text and storing in the local computer in a loop (see Appendix 1 for full code and Appendix 2 for list of speeches).

```
speech <-  read_html('http://statehouse.gov.ng/news/president-muhammadu-buhari-on-the-20th-anniversary-celebration-of-the-international-criminal-court-the-hague-netherlands/') #assigns the url to variable name 'speech'
pageText <- speech%>% #assigns speech to pageText
        html_nodes('p')%>% #identifies only the 'p' tags (paragraphs)
        html_text()  #turns to text
pageText <- paste(pageText, collapse = ' ')  #join all the paragraph into a full document
writeLines(pageText, 'speech46.text')   #saves locally as 'speech46' text file
```

3.2. Data Pre-processing

The extracted text has to be converted to data so as to apply any quantitative analytical technique, this involves removing irrelevant text and the transforming into a suitable form. This would be mostly done using the *tm* and *tidytext* package in R. In this case, the scraped and saved text is turned into a corpus (see Appendix 3 for code). This corpus (an R list) contains the 47 speeches made by President Buhari as elements, the advantage of doing so is that it is an easy way to organise and manage a collection of documents for further analysis. It allows for easy view of all document characteristics or a subset and the addition of document metadata.

With this, it is common practise to remove common words, convert text to lowercase, remove punctuations, sparse terms and other symbols or irrelevant text. Wilkerson and Casas (2017) noted that this should be done with care as standard stopwords such as 'can't' and 'cannot' might be relevant features for a study of presidential address tone.

Transformations applied in this prior to the quantitative analysis are listed below and the code can be found in Appendix 4

1. **Removal of stopwords:** In this work, common English words were removed as this analysis is aimed at identifying the central theme of speeches made and not the semantic tone, no exemptions were made. See Appendix 5 for the list of common words removed, these are those identified by the text mining package *tm*
2. **Removal of numbers, punctuations and trailing whitespaces**
3. **Converting upper to lower case:** In this transformation.
4. **Removal of other symbols and irrelevant text:** while the above transformations depends on an inbuilt R package, this was done by manually inspecting the documents and identifying characters that were irrelevant e.g. <a0>.
5. **Document-term matrix:** With the text cleaning completed, the text needs to be converted into a dataframe which is a 2x2 matrix where the speeches in the corpus are rows and the words appear as columns and the values are the frequency of occurrence as depicted Figure 1 below. This is the form used as input in further analysis performed. The values do not have to be frequencies, it could be weights assigned to depict importance of certain words.

	administration	also	country	development	government	national
inauguration	2	5	0	2	8	3
Speech15	11	8	12	6	17	4
Speech21	0	4	1	1	8	0
Speech28	3	6	3	4	2	3
Speech3	4	6	1	3	5	3
Speech33	0	8	0	3	2	3
Speech36	4	2	9	0	13	1
Speech41	0	5	3	6	2	2
Speech43	13	21	11	18	20	17
Speech9	1	1	2	0	7	19

Figure 2: Snippet of the DTM with rows as speeches and column as top 6 terms

4. Analysis

The bag-of-words representation of texts is memory-efficient and convenient for various types of analyses, and this often outweighs the disadvantage of losing information by dropping the word positions (Welbers et al., 2017). In this report, this approach is adopted to understand the common words used by The President comparing the inauguration speech to other speeches and finally an attempt to identify clusters of these speeches is made.

4.1. Bag of Words

The 47 speeches contain 5123 unique words, with 94% sparsity indicating that most words were not reused in the corpus (see bag of words code in Appendix 6). Looking at all speeches, the words that occur at least 100 times as shown in Figure 3 portrays a theme concentrating on the state and its citizens, which is as expected of a President. A theme which highlights the focus of the Presidency in the economy, on corruption and security (the three identified policy areas) can be seen in Figure 4.

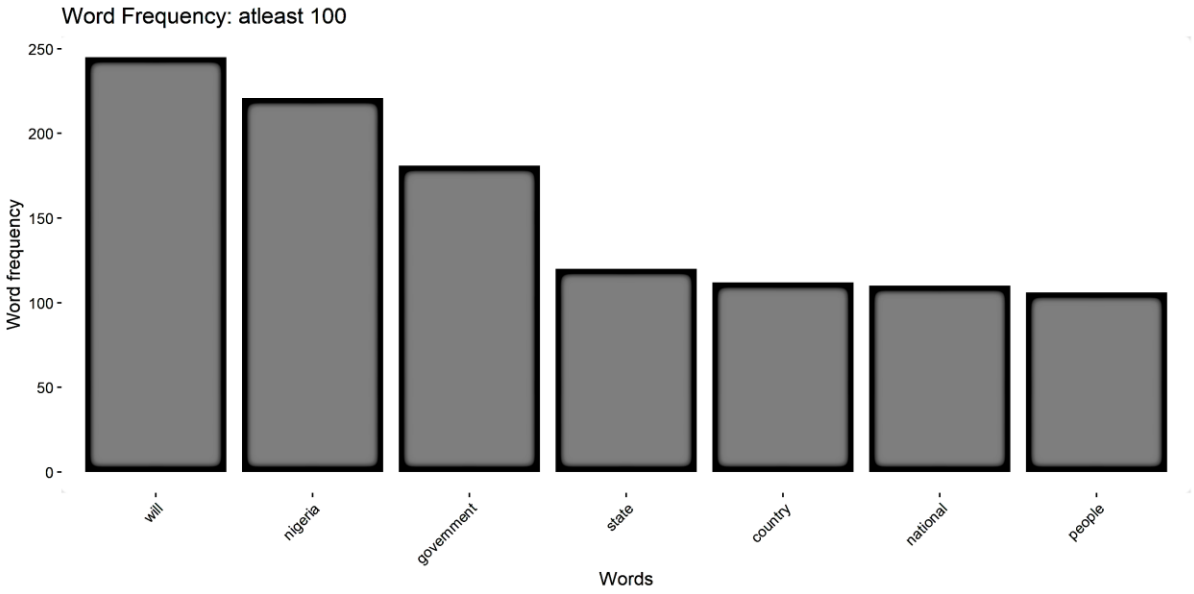


Figure 3: All Speeches word frequency - at least 100 occurrences

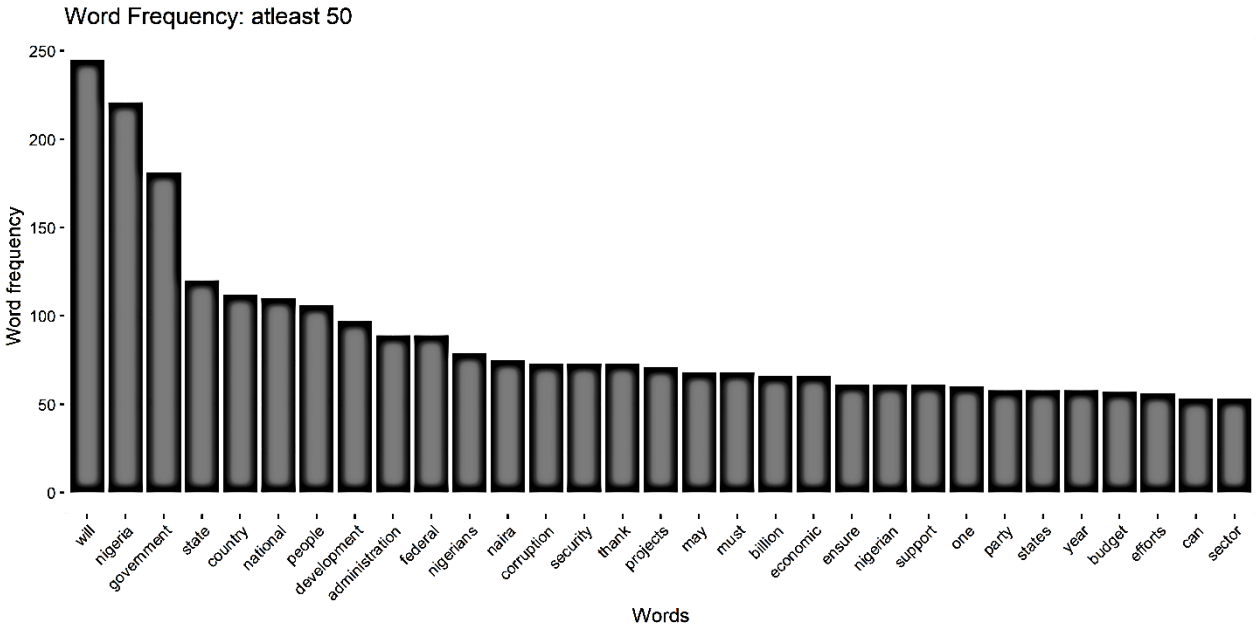


Figure 4: All Speeches word frequency - at least 50 occurrences

To compare the inauguration speech with the others, there were both analysed independently. The inauguration speech contained words of appreciation (e.g. thank, appreciated), addressed to the citizens (e.g. Nigeria, nation, people, state) and highlighting a

commitment to govern in line with the mandate given on equally ranked issues of corruption, terrorism and economy

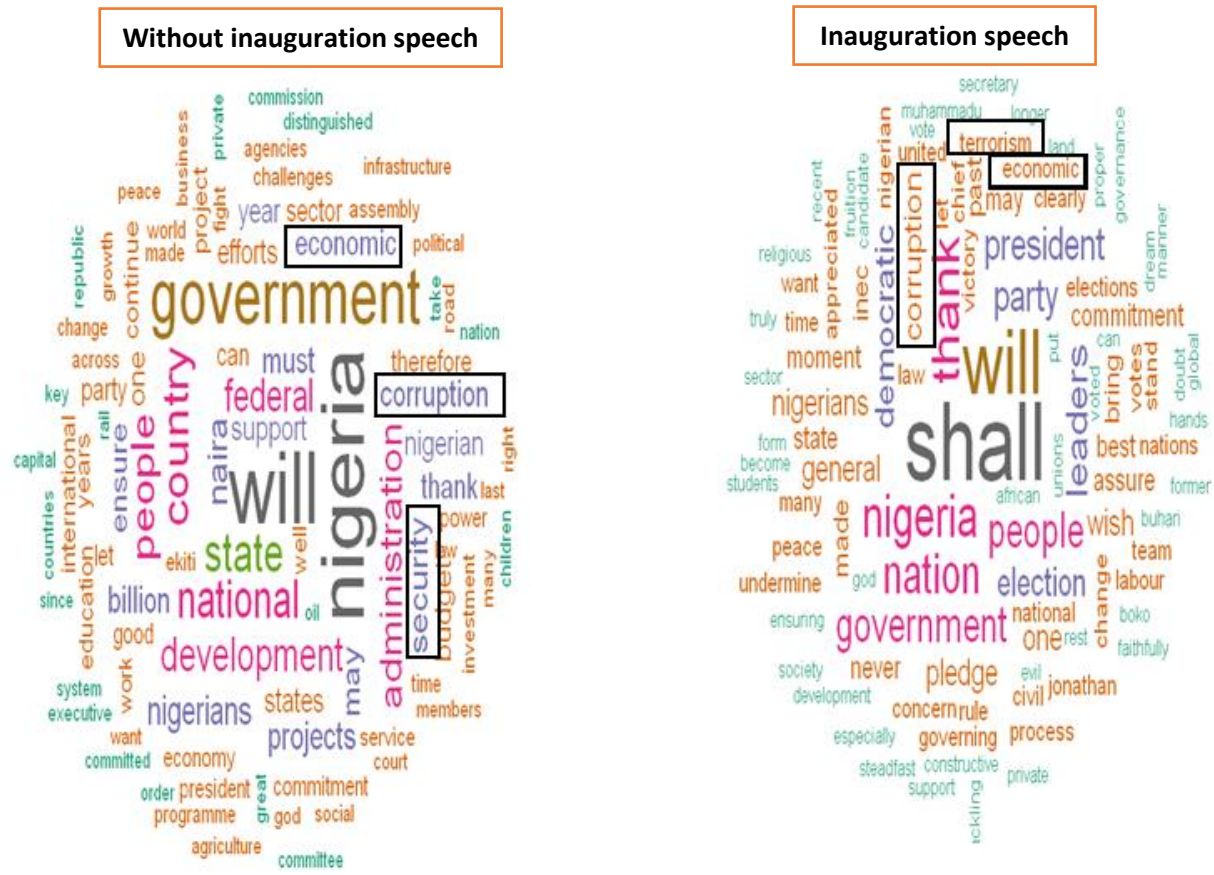


Figure 5: Word-cloud of terms which occur at least 25 times
The size indicate total frequency and the colour scheme represents frequency group.

This same rhetoric except for appreciation was found in the other 46 speeches. The term 'terrorism' is replaced with 'security' which highlights a change in events, prior to office the security menace was Boko-Haram but during his term other forms of insecurity arose, such as Biafra separatist agitation (IPOB calling to break away from Nigeria), Niger-delta militants and Fulani herdsmen killings[1].

With the insecurity identified to be state/ethnic related, an analysis to check which states are being talked about using frequency count was done.

[1] Indigenous People of Biafra (IPOB) is an organisation which is calling for a referendum to break away the south-eastern Nigeria into a new country called Biafra. Their activities were proscribed. The Niger delta militants arose to fight against perceived exploitation from the oil wealth found in their lands. Fulani herdsmen are semi-nomadic, pastoralist ethnic group and mostly Muslim who living in the central regions of Nigeria and clashes with indigenous tribes and local, mainly Christian, farmers over grazing land which has recently led to a large spike in deaths.

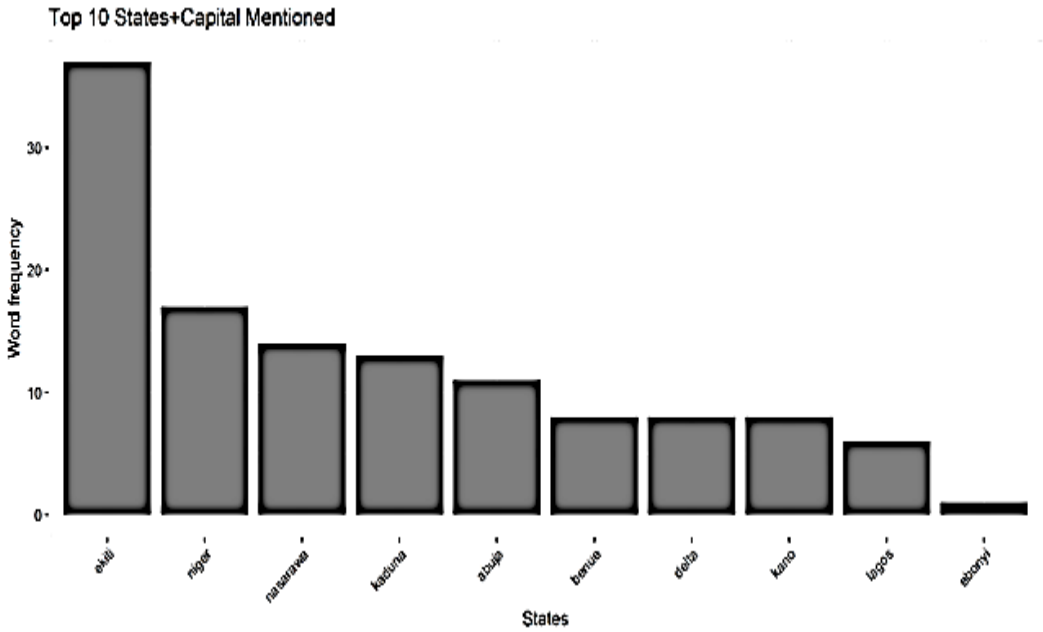


Figure 6: All speeches: occurrence of states and capital names.

With Ekiti being the most talked about state which is due to a governorship polls which took place on 14th July, 2018. This poll if won by the APC party means they take a seat off their opposition PDP as the incumbent Governor is a PDP member. With a belief that the results would act as a referendum on Buhari's performance, it is clear why high importance was placed on it by The President and thus the high difference in total occurrence when compared to other state mentioned.

With the highest mention made in Speech 3 during the APC grand finale of the governorship campaign, other occurrence were in Speech 15 and 18 (see Appendix 7). Buhari made campaign pitches focused on a fair election centred on the people, which should bring in a new APC Governor to produce jobs and deal with insurgency (Figure 7).

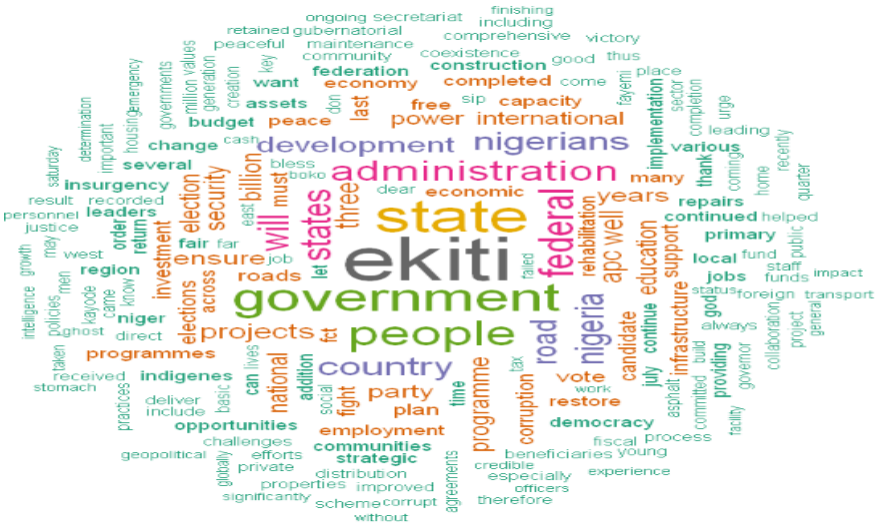


Figure 7: Word cloud of Speech 3, 15 and 18

Indicating word frequency of speeches in which 'ekiti' is present. Appendix 8 shows the co-occurrence of ekiti and other words at 0.7 correlation limit

The high occurrence of Ekiti could be said to reflect the importance placed by Buhari and his party on winning elections. When he talked about the Niger delta, as can be seen in Figure 8, he focused on economic recovery, ogoni clean up (needed due to oil spillages), amnesty and payment due salaries.

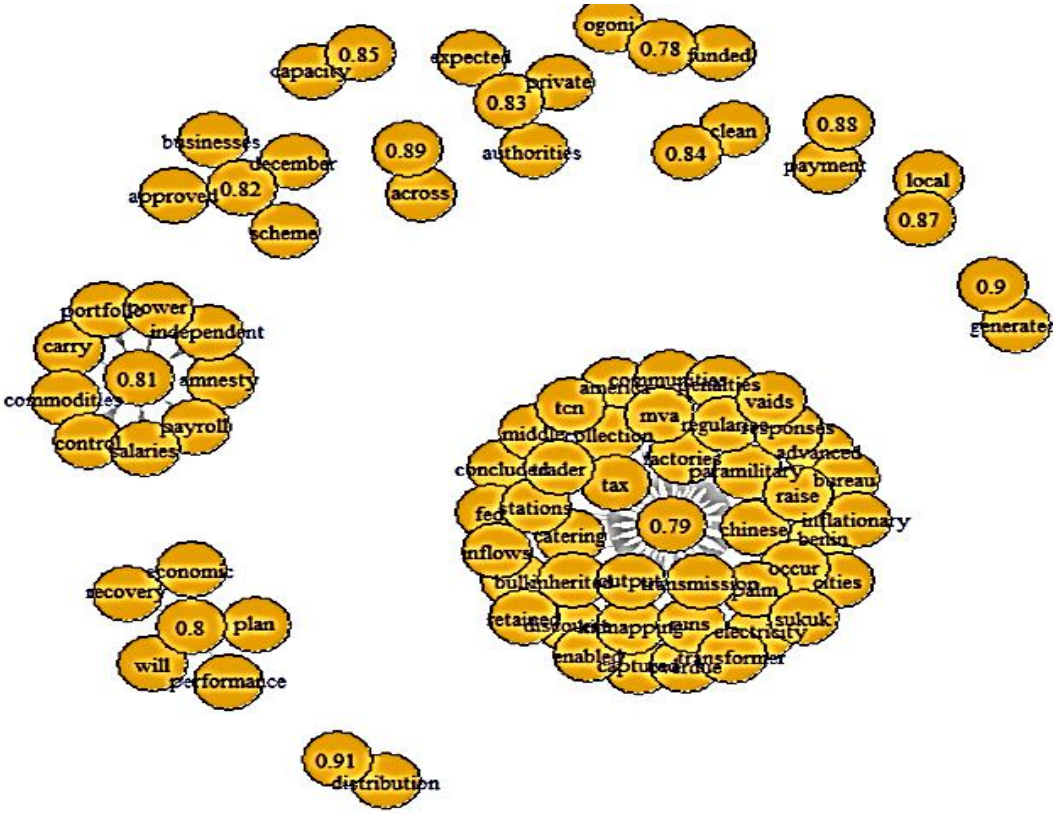


Figure 8: Co-occurrence of 'niger delta' with other words
With a correlation limit of 0.7, the figure depicts words that are mostly likely to occur with 'niger delta'

When talking about the northern states (specifically nasarawa, kaduna and kano) the focus are disability and with use of words such as curbing, barbaric, philosophy, suspected, this may be references to the terrorist surge that is dominant in the Northern state (Figure 9). However, given the IPOB agitation, south-eastern states are rarely mentioned while the common words related to the movement (such as Biafra, ipob) are never mentioned. This may be perceived as not placing importance on the concerns of the people of these states.

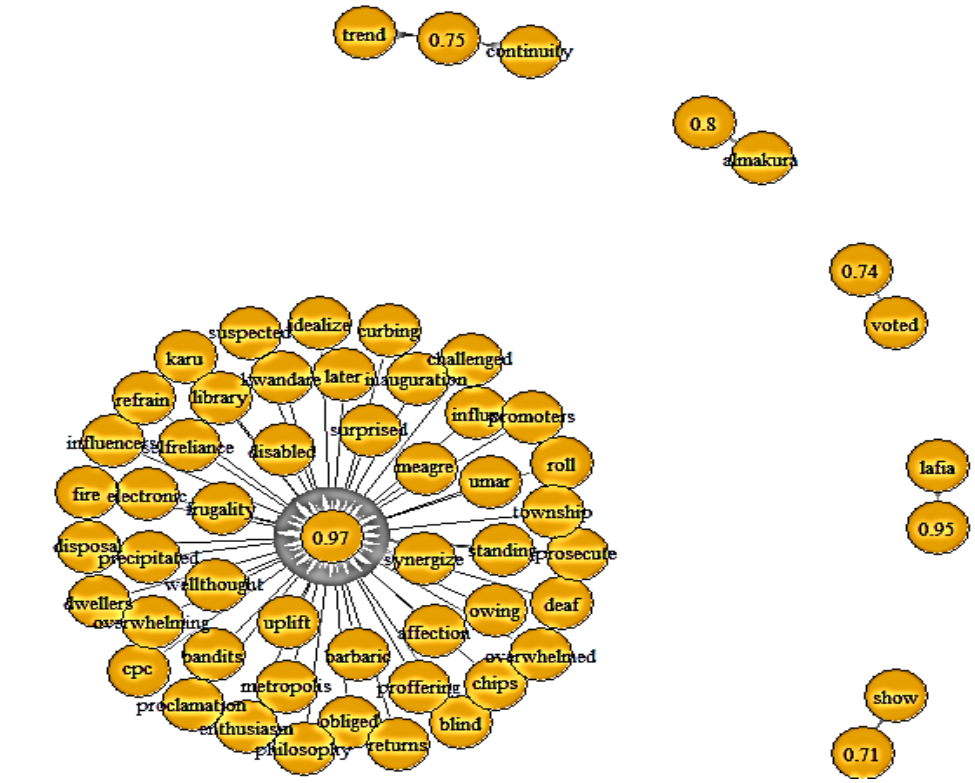


Figure 9: Co-occurrence of 'niger delta' with other words

With a correlation limit of 0.7, the figure depicts words that are mostly likely to occur with 'niger delta'

Looking at the three policy focus (economic, corruption, security) identified, there is a consistent mention of these across speeches as seen in Figure 10 below. With Speech 21 (ranked second) delivered during a state visit to the United States, where The President would want to appear to be tackling these issues. The occurrence of these themes can be said to have been consistently varied across the speeches, there is no clear pattern.

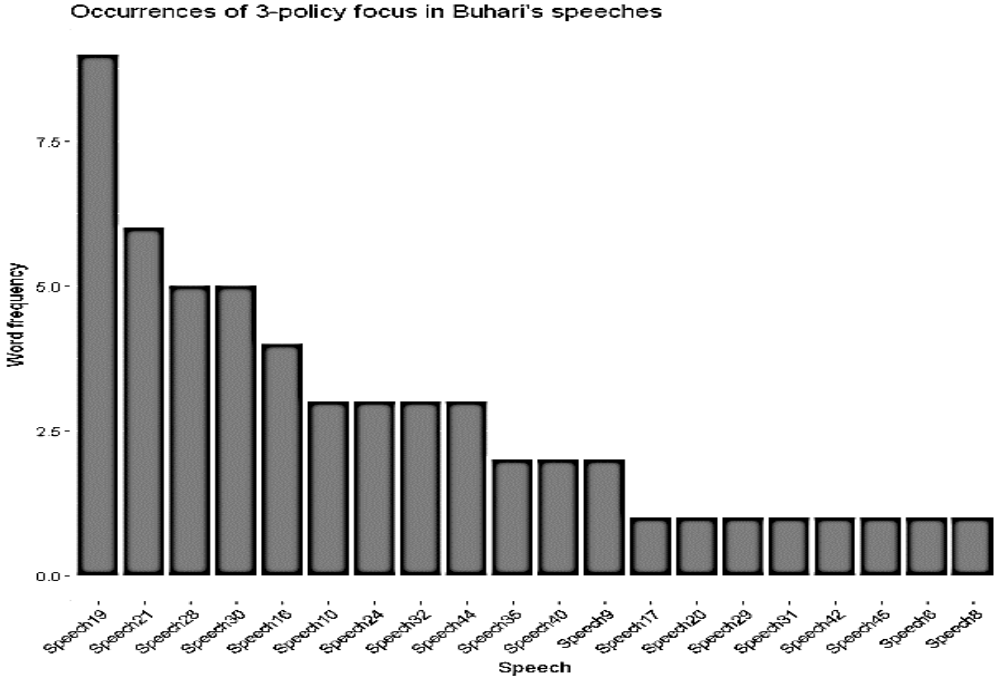


Figure 10: Occurrence of 'economic, corruption, security' in the speeches

The occurrence of these themes in these specific speeches makes sense when the speech title (shown in Appendix 2) is reviewed. With this in mind, using an unsupervised machine learning approach, an attempt to group these speeches into homogenous classes (ideally, speeches focused on the same theme would be grouped together) is made.

4.2. Clustering

The model adopted is unsupervised as it is not trained to know which theme the speech is focused on, the model thus learns this accessing the word frequency and occurrence between speeches. In particular, the hierarchical agglomerative clustering (HAC) algorithm using the Wards method is used to identify words with similar themes (see code in Appendix 9). The data used for the analysis is a matrix of the word count across speeches, using this a dissimilarity matrix is constructed. This dissimilarity matrix as well as a linkage criterion (in this analysis Wards is applied) is used to decide which speeches (initially as clusters) should be combined. Wards criterion aims to minimise the total within-cluster variance. To evaluate the success of this model, the groups of speeches identified would be looked at to see if it those have the same theme.

One of the advantages of using a HAC, is that number of groups is not to be decided prior to the analysis and a dendogram produced from the analysis can be visually analysed to identify the suitable number of groups.

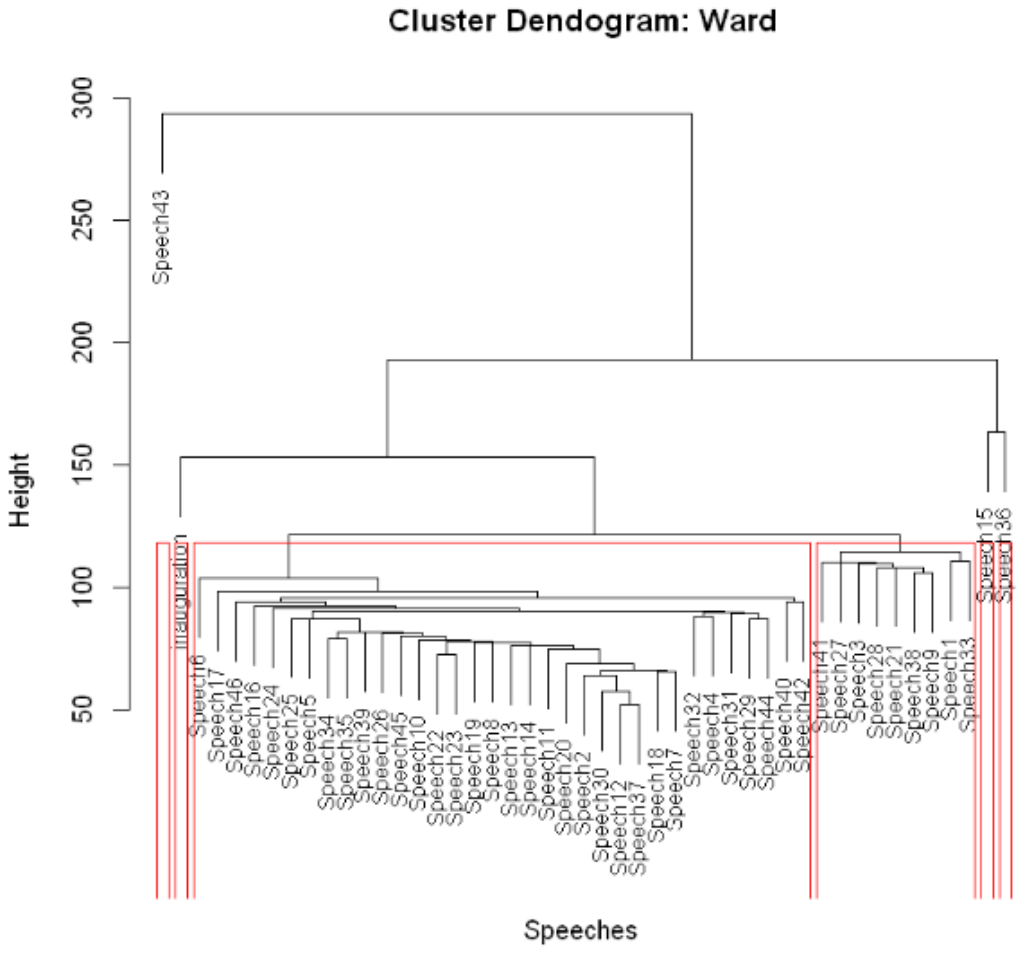


Figure 11: Dendogram of speeches - 5 groups

From Figure 11, six groups were identified with total class membership as shown in Table 1 below. Cluster's 1, 2, 5 and 6 had one speech each which are all distinct when compared to other speeches (See Appendix 10 for bar chart of top terms).

Table 1 : Cluster membership

Cluster	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Size	1	1	34	9	1	1
Speeches	43: 2018 Budget Address	inauguration	See Figure 10 theme:**Development**	See Figure 10 theme:**Corruption**	15: 2018 democracy day celebration	36: Commissioning of the Kaduna inland dry port

For cluster 3 and 4, there are 34 and 9 speeches as members respectively, using a word cloud as shown in Figure 12, there are some differences in both clusters. The common theme of being the president of the people can be elicited. Inspecting Cluster 4, there is evidence of a corruption/money theme (pink coloured words) whereas Cluster 3 has high frequency of words 'development', 'administration', 'party' (pink coloured) which can be labelled as evidence of the party's desire to work ensure development.

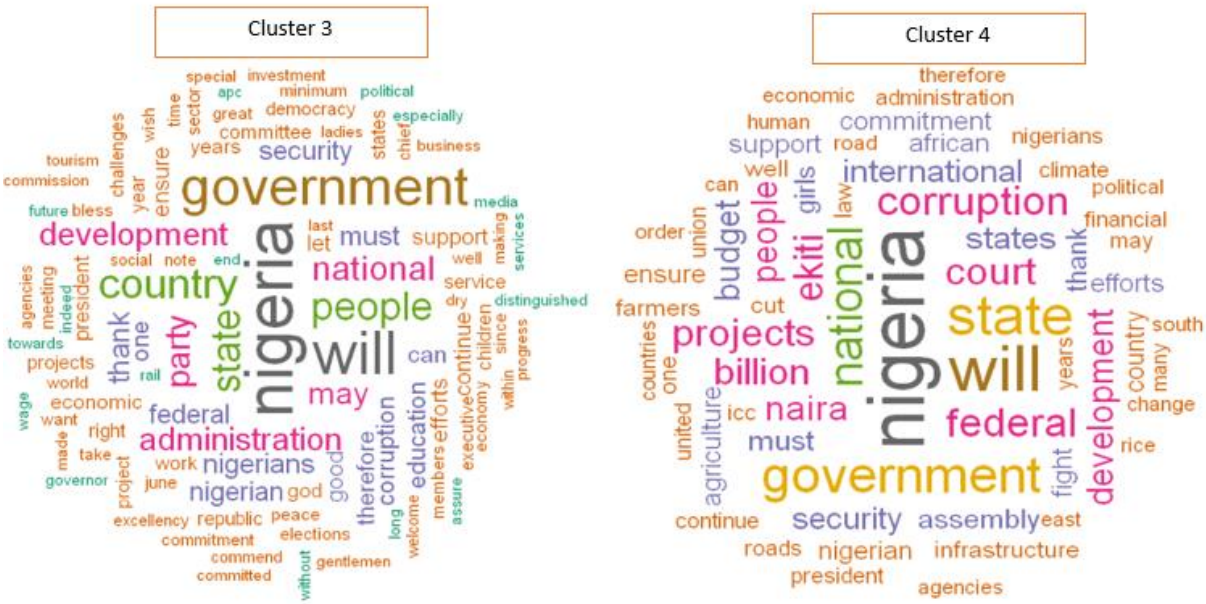


Figure 12: Word cloud of cluster 3 and 4.

5. Conclusion

This work has shown how text mining techniques can be applied to a body of text so as to gain insights that might take longer if done manually. In particular, The President of Nigeria's speeches has been analysed and results show consistency in tone. The 3-policy focus was focused on across

majority of the speeches. It was however found that developing events e.g. the IPOB agitation was not addressed by The President. A focus on Ekiti state (1 out of 37 states and capital) was identified which was due to the by-election conducted in Ekiti in which the APC candidate subsequently won. The clustering performed cannot be said to have performed well, as the 2 large clusters were not very distinct when the word frequency was compared. This may be due to the fact that the President showed consistency by keeping focus on the campaign promises.

It may also be that the particular algorithm used or number of clusters selected is not fit for purpose, for number of cluster selection, future research could use evaluation metrics such as intra-cluster variance to evaluate results of different number of clusters and select the better performer. Supervised algorithms could be used, when a percentage of speeches are handle labelled, a model trained and then cluster membership predicted for the unlabelled speeches. The bag of words approach which ignored the positive of terms and used term frequency as weight could be improved on by adopting algorithms which process a corpus as close to possible as a human would.

References

1. Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E.D., Gutierrez, J.B. and Kochut, K., 2017 (2017) *A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques* https://arxiv.org/pdf/1707.02919.pdf
2. arc12 (2012) *Standard set of english stopwords*. Available at: https://github.com/arc12/Text-Mining-Weak-Signals/wiki/Standard-set-of-english-stopwords (Accessed: 9th July, 2018)
3. Centre for Democracy and Development (2017) *Buharimeter Mid-term Report*. Available at: http://cddwestafrica.org/wp-content/uploads/2018/01/CDD-PROFILE-CURRENT-1.pdf (Accessed: 15th July, 2018)
4. Damian Zane (2015) *Nigeria's Goodluck Jonathan: Five reasons why he lost*, BBC News, 31 March. Available at: https://www.bbc.co.uk/news/world-africa-32136295 (Accessed: 15th July, 2018)
5. Godwin Uyi Ojo (2015*) 'The fall in oil price could be a turning point for Nigeria's economy'*, The Guardian, 5 March [Online]. Available at: https://www.theguardian.com/world/2015/mar/05/the-fall-in-oil-price-could-be-a-turning-point-for-nigerians-economy (Accessed: 13th July, 2018)
6. Hadley Wickham (2014) *rvest: easy web scraping with R*, R Studio Blog, 24 November. Available at: https://blog.rstudio.com/2014/11/24/rvest-easy-web-scraping-with-r/ (Accessed: 9th July, 2018)
7. Heather Froehlich (2018) *Intro to Text Analysis,* PennState University Libraries, 8 August. Available at: https://guides.libraries.psu.edu/textanalysis (Accessed: 15th July, 2018)
8. Jelle W. Boumans & Damian Trilling (2016) Taking Stock of the Toolkit, Digital Journalism, 4:1, 8-23, DOI: 10.1080/21670811.2015.1096598
9. Kasper Welbers, Wouter Van Atteveldt & Kenneth Benoit (2017) Text Analysis in R, Communication Methods and Measures, 11:4, 245-265, DOI: 10.1080/19312458.2017.1387238
10. Liddy, E.D. (2001) *Natural Language Processing*. In Encyclopedia of Library and Information Science, 2nd Ed. NY. Marcel Decker, Inc.
11. Trading Economics (2018) *Nigeria Corruption Index.* Available at: https://tradingeconomics.com/nigeria/corruption-index (Accessed: 13th July, 2018)

Appendix

Packages to install

In [1]:
```
install.packages("tm", repos = "http://cran.us.r-project.org")
library(tm)


install.packages('tidyverse', repos ='http://cran.us.r-project.org')
library(tidyverse)

install.packages("wordcloud", repos ='http://cran.us.r-project.org')
library(wordcloud)

install.packages("tidytext", repos ='http://cran.us.r-project.org') #download this to use tidy
library(tidytext)

install.packages("proxy", repos = "http://cran.us.r-project.org")
library(proxy)

install.packages("fpc", repos = "http://cran.us.r-project.org")
library(fpc)

install.packages("cluster", repos = "http://cran.us.r-project.org")
library(cluster)

install.packages(c("dplyr", "stringr"), repos ='http://cran.us.r-project.org')
library(dplyr)
library(stringr)

install.packages(c("tidyr", "igraph", "ggraph"), repos ='http://cran.us.r-project.org')
library(tidyr)
library(igraph)
library(ggraph)
```

```
url <- c('https://statehouse.gov.ng/news/president-muhammadu-buhari-on-the-20th-anniversary-celebration-of-the-international-crim
         'https://statehouse.gov.ng/news/president-muhammadu-buharis-address-at-the-commissioning-of-the-abuja-light-rail-system/
         'https://statehouse.gov.ng/news/president-muhammadu-buharis-remarks-at-the-apc-grand-finale-of-the-ekiti-guber-national-
         'https://statehouse.gov.ng/news/president-muhammadu-buharis-remarks-at-his-meeting-with-the-christian-association-of-nig
         'https://statehouse.gov.ng/news/president-muhammadu-buharis-speech-at-the-commissioning-of-calabar-rice-seedling-plant-c
         'https://statehouse.gov.ng/news/president-muhammadu-buharis-address-at-the-commissioning-ceremony-of-nigerian-navy-refer
         'https://statehouse.gov.ng/news/president-buharis-closing-remarks-at-the-apc-convention-in-abuja/',
         'https://statehouse.gov.ng/news/president-buharis-speech-at-the-opening-ceremony-of-the-67th-international-press-institu
         'https://statehouse.gov.ng/news/president-buharis-speech-at-the-signing-of-the-2018-appropriation-bill-into-law-at-presi
         'https://statehouse.gov.ng/news/president-buharis-statement-at-the-hosting-of-members-of-the-diplomatic-corps-for-ramada
         'https://statehouse.gov.ng/news/president-buharis-remarks-at-the-investiture-honouring-the-heroes-of-june-12-1993/',
         'https://statehouse.gov.ng/news/president-buhari-declares-june-12-the-new-democracy-day/',
         'https://statehouse.gov.ng/news/president-buharis-address-at-the-61st-meeting-of-the-unwto-caf/',
         'https://statehouse.gov.ng/news/president-buharis-remarks-at-the-signing-of-the-not-too-young-to-run-bill-in-abuja/',
         'https://statehouse.gov.ng/news/president-buharis-address-in-commemoration-of-the-2018-democracy-day-celebration/',
         'https://statehouse.gov.ng/news/president-buharis-remarks-at-the-2018-nigeria-democracy-day-lecture-at-icc-abuja/',
         'https://statehouse.gov.ng/news/president-buharis-message-on-national-childrens-day-may-27-2018/',
         'https://statehouse.gov.ng/news/president-buharis-remarks-at-dinner-with-apc-south-west-leaders-in-abuja/',
         'https://statehouse.gov.ng/news/president-buharis-address-at-the-commissioning-of-the-efcc-headquarters-in-abuja/',
         'https://statehouse.gov.ng/news/press-release-nigerias-agricultural-revolution-on-course-president-buhari/',
         'https://statehouse.gov.ng/news/press-release-press-statement-by-president-buhari-during-his-visit-to-the-united-states-
         'https://statehouse.gov.ng/news/president-buharis-address-on-the-report-submitted-by-the-all-progressives-congress-natio
         'https://statehouse.gov.ng/news/president-buharis-address-at-the-apc-nec-meeting-in-abuja/',
         'https://statehouse.gov.ng/news/president-buharis-address-at-the-inauguration-of-the-national-food-security-council-in-a
         'https://statehouse.gov.ng/news/president-buharis-address-at-the-occasion-of-receiving-the-release-of-dapchi-school-girl
         'https://statehouse.gov.ng/news/president-buharis-address-at-the-commissioning-of-sunti-golden-sugar-estate-in-niger-sta
         'https://statehouse.gov.ng/news/president-buharis-address-on-his-official-visit-to-yobe-state/',
         'https://statehouse.gov.ng/news/president-buharis-speech-at-a-meeting-with-stakeholders-in-the-rice-value-chain/',
         'https://statehouse.gov.ng/news/speech-by-president-buhari-at-the-61st-independence-day-anniversary-of-ghana/',
         'https://statehouse.gov.ng/news/speech-president-buharis-address-at-the-first-adamawa-state-anti-corruption-summit/',
         'https://statehouse.gov.ng/news/speech-president-buharis-remarks-at-the-commissioning-of-comprehensive-special-school-la
         'https://statehouse.gov.ng/news/press-release-president-buharis-letter-to-senate-president-on-benue-killings/',
         'https://statehouse.gov.ng/news/speech-president-buharis-address-at-the-30th-ordinary-session-of-assembly-of-heads-of-st
         'https://statehouse.gov.ng/news/speech-president-buhari-at-the-launch-of-new-locomotives-and-coaches-for-the-kaduna-abuj
         'https://statehouse.gov.ng/news/speech-president-buhari-at-the-commissioning-of-the-kaduna-inland-dry-port-january-4-201
         'https://statehouse.gov.ng/news/new-year-address-by-president-buhari-jan-1-2018/',
         'https://statehouse.gov.ng/news/statement-president-buhari-on-fuel-scarcity/',
         'https://statehouse.gov.ng/news/speech-president-buharis-address-at-the-international-climate-change-summit-in-paris/',
         'https://statehouse.gov.ng/news/speech-president-buharis-address-at-the-inauguration-of-the-tripartite-minimum-wage-comm
         'https://statehouse.gov.ng/news/speech-president-buharis-remarks-at-the-2017-all-nigeria-judges-conference-of-the-suprem
         'https://statehouse.gov.ng/news/speech-president-buharis-address-at-his-state-visit-to-ebonyi-state/',
         'https://statehouse.gov.ng/news/speech-president-buharis-remarks-at-fecs-special-retreat-on-education-held-in-abuja/',
         'https://statehouse.gov.ng/news/speech-president-buharis-2018-budget-address/',
```

1. Web scraping using rvest

Web scraping

```
for (i in 1:length(url))
{


  speech <- read_html(url[i])
  str(speech)
  h2 <- speech%>%
  html_nodes('section')
  pageText <- speech%>%
    html_nodes('p')%>%
    html_text()
  pageText <- paste(pageText, collapse = ' ') #join everything together

  writeLines(as.character(pageText), paste0(gsub(" ", "_", url[i]), ".text"))
}
```

2. List of Speeches

SPEECH	TITLE	Date
1	President Buhari at the International Criminal Court 20th Anniversary celebration	17/7/2018
2	Commissioning of Abuja Light rail system	12/7/2018
3	APC grand finale of the Ekiti gubernational campaign rally	11/7/2018
4	Meeting with Christian association from 19 Northern state	5/7/2018
5	Commissioning of Calabar rice seedling plant	1/7/2018
6	Commissioning of Nigerian navy reference hospital Calabar	26/6/2018
7	Remarks at APC Convention in Abuja	24/6/2018
8	Opening ceremony of the 67th international press institute, Abuja	22/6/2018
9	Signing of the 2018 appropriation bill into law	20/6/2018
10	Hosting of members of the diplomatic corps for Ramadan affair	16/6/2018
11	Remark at the investiture honouring the heroes of June 12, 1993	12/6/2018
12	Declaration of June 12 as new democracy day	6/6/2018
13	Address at the 61st meeting of the UNWTO CAF	4/6/2018
14	Remarks at the signing of the not too young to run bill	31/5/2018
15	Address in commemoration of the 2018 democracy day celebration	29/5/2015
16	Remarks at the 2018 Nigeria democracy day lecture	28/5/2018
17	Message at the 2018 national children's day	27/5/2018
18	Remarks at dinner with APC South West leaders	17/5/2018
19	Address at commissioning of the EFCC headquarters in Abuja	15/5/2018

20	Press release on Nigeria's agricultural revolution on course	15/5/2018
21	Press release during visit to the United States	30/04/2018
22	Address on the report submitted to the APC national executive technical committee	9/4/2018
23	Address at the APC national executive council meeting	27/3/2018
24	Address at the inauguration of the national food security council	26/3/2018
25	Address at the occasion of receiving the release of Dapchi school girls	23/3/2018
26	Address at the commissioning of sunti golden sugar estate in Niger state	15/3/2018
27	Address on his official visit to Yobe State	14/3/2018
28	Speech at a meeting with stakeholders in the rice value chain	13/3/2018
29	Speech at the 61st independence day anniversary of Ghana	6/3/2018
30	Address at the first Adamawa State anti-corruption summit	20/2/2018
31	Remarks at commissioning of comprehensive special school Lafia	6/2/2018
32	Letter to senate president on Benue killings	1/2/2018
33	Address at the 30th ordinary session of assembly heads of state and government of the African Union	27/1/2018
34	Launch of new locomotives and coaches fir the Kaduna-Abuja train service	4/1/2018
35	Commissioning of the Kaduna inland dry port	4/1/2018
36	New year address	1/1/2018
37	Statement on fuel scarcity	24/12/2017
38	Address at the international climate change summit	13/12/2017
39	Inauguration of the tripartite minimum wage committee	27/11/2017
40	Remarks at the 2017 all Nigeria judges conference of the supreme high courts	20/11/2017
41	Address at state visit to Ebonyi state	14/11/2017
42	Remarks at FECS special retreat on education	13/11/2017
43	2018 Budget Address	8/11/2017
44	Launch of 2018 armed forces emblem	1/11/2017
45	Address at the NEC meeting in Abuja	31/10/2017
46	Speech at the 9th summit of the D8 organisation for economic cooperation	20/10/2017
Inauguration	Speech made on the inauguration as President	29/05/2015

3. Code for making a corpus using tm

Create a corpus

```
#define the path of the document collection
cname <- file.path(".", "speech")
cname #check it's done it write
length(dir(cname)) #are the documnet number as expected?
dir(cname) #what are the names of these documents?
#now, turn this into a Corpus - collection of document. we use a simple corpus instead of a virtual one
buhariCorpus <- Corpus(DirSource(cname, encoding="UTF-8")) #use DirSource as documents are stored in file directories
#check data structure
str(buhariCorpus)
#summary of documents loaded
summary(buhariCorpus)
```

4. Code for removing irrelevant characters

Clean up

```
buhariCorpus <- tm_map(buhariCorpus, content_transformer(gsub), pattern = "<91>", replacement=" ")
buhariCorpus <- tm_map(buhariCorpus, content_transformer(gsub), pattern = "<92>", replacement=" ")
buhariCorpus <- tm_map(buhariCorpus, content_transformer(gsub), pattern = "<93>", replacement=" ")
buhariCorpus <- tm_map(buhariCorpus, content_transformer(gsub), pattern = "<94>", replacement=" ")
buhariCorpus <- tm_map(buhariCorpus, content_transformer(gsub), pattern = "<95>", replacement=" ")
buhariCorpus <- tm_map(buhariCorpus, content_transformer(gsub), pattern = "<96>", replacement=" ")
buhariCorpus <- tm_map(buhariCorpus, content_transformer(gsub), pattern = "<97>", replacement=" ")
buhariCorpus <- tm_map(buhariCorpus, content_transformer(gsub), pattern = "<U+200B>", replacement=" ")
buhariCorpus <- tm_map(buhariCorpus, content_transformer(gsub), pattern = "Garba ShehuSenior Special Assistant to the PresidentM
buhariCorpus <- tm_map(buhariCorpus, content_transformer(gsub), pattern = "<a0>", replacement=" ")
buhariCorpus <- tm_map(buhariCorpus, content_transformer(gsub), pattern = "cross river", replacement="crossriver")
buhariCorpus <- tm_map(buhariCorpus, content_transformer(gsub), pattern = "etc", replacement=" ")
```

```
#convert lower cases to upper cases
buhariCorpus <- tm_map(buhariCorpus, content_transformer(tolower))
#remove stopwords
buhariCorpus <- tm_map(buhariCorpus, removeWords, stopwords("english"))
#remove punctuations
buhariCorpus <- tm_map(buhariCorpus, removePunctuation)
#remove numbers
buhariCorpus <- tm_map(buhariCorpus, removeNumbers)
#strip white spaces
buhariCorpus <- tm_map(buhariCorpus, stripWhitespace)
```

```
customised_stopwords <- c("even", "today","also","now","new", "real", "day", "first", "like")
buhariCorpus <- tm_map(buhariCorpus, removeWords, customised_stopwords)
```

Create a document Term matrix

```
#create a term document matrix: this has documents as columns and terms as rows
buhariDTM <- DocumentTermMatrix(buhariCorpus)
buhariDTM
```

```
#check the occurence of the states in Nigeria in the DTM and how often it occurs
stateDTM <- inspect(buhariDTM[, c("abuja", "lagos", "enugu", "ondo", "anambra", "bauchi", "bayelsa","benue", "borno",
                    "delta", "ebonyi", "edo", "ekiti", "gombe", "jigawa", "kaduna", "kano",
                   "katsina", "kebbi", "kogi", "kwara", "nasarawa", "niger", "ogun", "osun", "oyo", "plateau",
                  "rivers", "taraba", "yobe", "zamfara")])
#convert stateDTM to a normal matrix
stateTerms <- colSums(as.matrix(stateDTM))
#convert above to a dtaframe for plotting
stateTermsDF <- data.frame(word=names(stateTerms), freq=stateTerms)
```

5. Stopwords (Standard set of english stopwords, 2012)

"a, about, above, across, after, again, against, all, almost, alone, along, already, also, although, always, am, among, an, and, another, any, anybody, anyone, anything, anywhere, are, area, areas, aren't, around, as, ask, asked, asking, asks, at, away, b, back, backed, backing, backs, be, became, because, become, becomes, been, before, began, behind, being, beings, below, best, better, between, big, both, but, by, c, came, can, cannot, can't, case, cases, certain, certainly, clear, clearly, come, could, couldn't, d, did, didn't, differ, different, differently, do, does, doesn't, doing, done, don't, down,

downed, downing, downs, during, e, each, early, either, end, ended, ending, ends, enough, even, evenly, ever, every, everybody, everyone, everything, everywhere, f, face, faces, fact, facts, far, felt, few, find, finds, first, for, four, from, full, fully, further, furthered, furthering, furthers, g, gave, general, generally, get, gets, give, given, gives, go, going, good, goods, got, great, greater, greatest, group, grouped, grouping, groups, h, had, hadn't, has, hasn't, have, haven't, having, he, he'd, he'll, her, here, here's, hers, herself, he's, high, higher, highest, him, himself, his, how, however, how's, i, i'd, if, i'll, i'm, important, in, interest, interested, interesting, interests, into, is, isn't, it, its, it's, itself, i've, j, just, k, keep, keeps, kind, knew, know, known, knows, l, large, largely, last, later, latest, least, less, let, lets, let's, like, likely, long, longer, longest, m, made, make, making, man, many, may, me, member, members, men, might, more, most, mostly, mr, mrs, much, must, mustn't, my, myself, n, necessary, need, needed, needing, needs, never, new, newer, newest, next, no, nobody, non, noone, nor, not, nothing, now, nowhere, number, numbers, o, of, off, often, old, older, oldest, on, once, one, only, open, opened, opening, opens, or, order, ordered, ordering, orders, other, others, ought, our, ours, ourselves, out, over, own, p, part, parted, parting, parts, per, perhaps, place, places, point, pointed, pointing, points, possible, present, presented, presenting, presents, problem, problems, put, puts, q, quite, r, rather, really, right, room, rooms, s, said, same, saw, say, says, second, seconds, see, seem, seemed, seeming, seems, sees, several, shall, shan't, she, she'd, she'll, she's, should, shouldn't, show, showed, showing, shows, side, sides, since, small, smaller, smallest, so, some, somebody, someone, something, somewhere, state, states, still, such, sure, t, take, taken, than, that, that's, the, their, theirs, them, themselves, then, there, therefore, there's, these, they, they'd, they'll, they're, they've, thing, things, think, thinks, this, those, though, thought, thoughts, three, through, thus, to, today, together, too, took, toward, turn, turned, turning, turns, two, u, under, until, up, upon, us, use, used, uses, v, very, w, want, wanted, wanting, wants, was, wasn't, way, ways, we, we'd, well, we'll, wells, went, were, we're, weren't, we've, what, what's, when, when's, where, where's, whether, which, while, who, whole, whom, who's, whose, why, why's, will, with, within, without, won't, work, worked, working, works, would, wouldn't, x, y, year, years, yes, yet, you, you'd, you'll, young, younger, youngest, your, you're, yours, yourself, yourselves, you've, z"

6. Code for Bag of Words Analysis

Bag of words analysis

```
#barchart of word frequency
ggplot(subset(buhariTermsDF, freq>50), aes(x = reorder(word, -freq), y = freq)) +
    geom_bar(stat = "identity") +
    ggtitle("Top 50 most occuring words") +
    labs(x="Words", y="Word frequency") +
    theme(axis.text.x=element_text(angle=45, hjust=1))
```

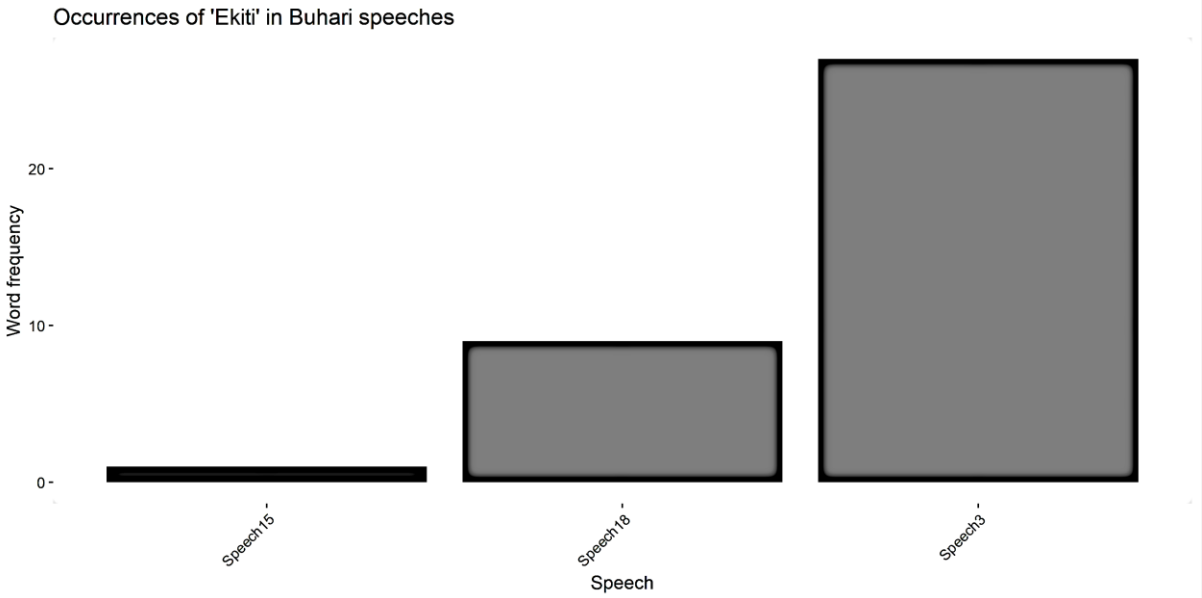
```
#word cloud
set.seed(12222)
#par(mfrow=c(1,2)) #places chart side by side
wordcloud(NoInaugTermsDF$word, NoInaugTermsDF$freq, min.freq = 25,
          max.words=100, random.order=FALSE, rot.per=0.35,
            colors=brewer.pal(8, "Dark2"))
wordcloud(inaugTermsDF$word, inaugTermsDF$freq, min.freq = 25,
          max.words=100, random.order=FALSE, rot.per=0.35,
          colors=brewer.pal(8, "Dark2"))
```

```
#find the occurence of certain words
df_buhari <- tidy(buhariDTM)
df_buhari %>%
filter(term==c("corruption", "security", "economic")) %>%
group_by(document) %>%
ggplot(aes(x = reorder(document, -count), y = count)) +
    geom_bar(stat = "identity") +
    ggtitle("Occurrences of 3-policy focus in Buhari's speeches") +
    labs(x="Speech", y="Word frequency") +
   theme(axis.text.x=element_text(angle=45, hjust=1))
ggsave('policy.png', width=12, height=10)
```

Co-occurence

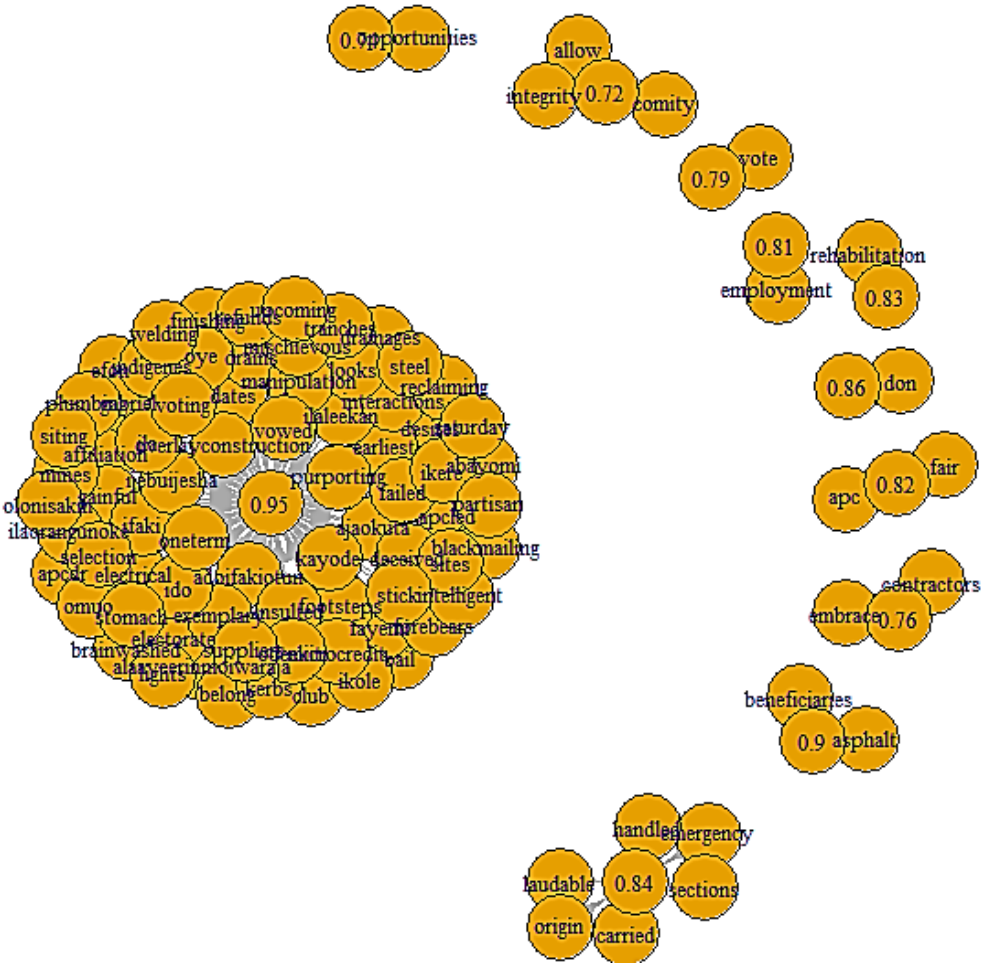
```
delta1 <- findAssocs(buhariDTM, c("niger", "delta"), corlimit=0.78)[[1]]
delta1<-cbind(read.table(text =names(delta1),stringsAsFactors = FALSE), delta1)
g<-graph.data.frame(delta1,directed =TRUE)
png('delta.png', width =1900, height = 700)
plot(g)
dev.off()
```

7. Occurrence of Ekiti in speeches



Speech 3	APC grand finale of the Ekiti gubernational campaign rally	11/7/2018
Speech 15	Address in commemoration of the 2018 democracy day celebration	29/5/2015
Speech 18	Remarks at dinner with APC South West leaders	17/5/2018

8. Co-occurrence of ekiti with other words



9. Code for clustering analysis

CLUSTERING

```
mat <- as.matrix(buhariDTM) #turn DTM into a matrix
docsdissim <- dist(scale(mat)) #calculate distance to create a dissimilarity matrix
h<-hclust(docsdissim, method = 'ward.D') #cluster usisng hiericichal ward.D/ward.D2/complete as the metric
clustering <- cutree(h, 6)
#"ward.D", "ward.D2", "single", "complete", "average" (= UPGMA), "mcquitty" (= WPGMA), "median" (= WPGMC) o
plot(h, cex=0.8, hang=0.1, main = "Cluster Dendogram: Ward", xlab = "Speeches", frame.plot=FALSE)
rect.hclust(h,6, border = "red")
```

```
#word cloud of cluster
cluster3 <- mat[clustering ==3, ]
cluster3Terms <- sort(colSums(as.matrix(cluster3)), decreasing=TRUE)
cluster3TermsDF <- data.frame(word=names(cluster3Terms), freq=cluster3Terms)
wordcloud(cluster3TermsDF$word, cluster3TermsDF$freq, min.freq = 10,
          max.words=100, random.order=FALSE, rot.per=0.35,
          colors=brewer.pal(8, "Dark2"))
```

Tidy Version of Corpus

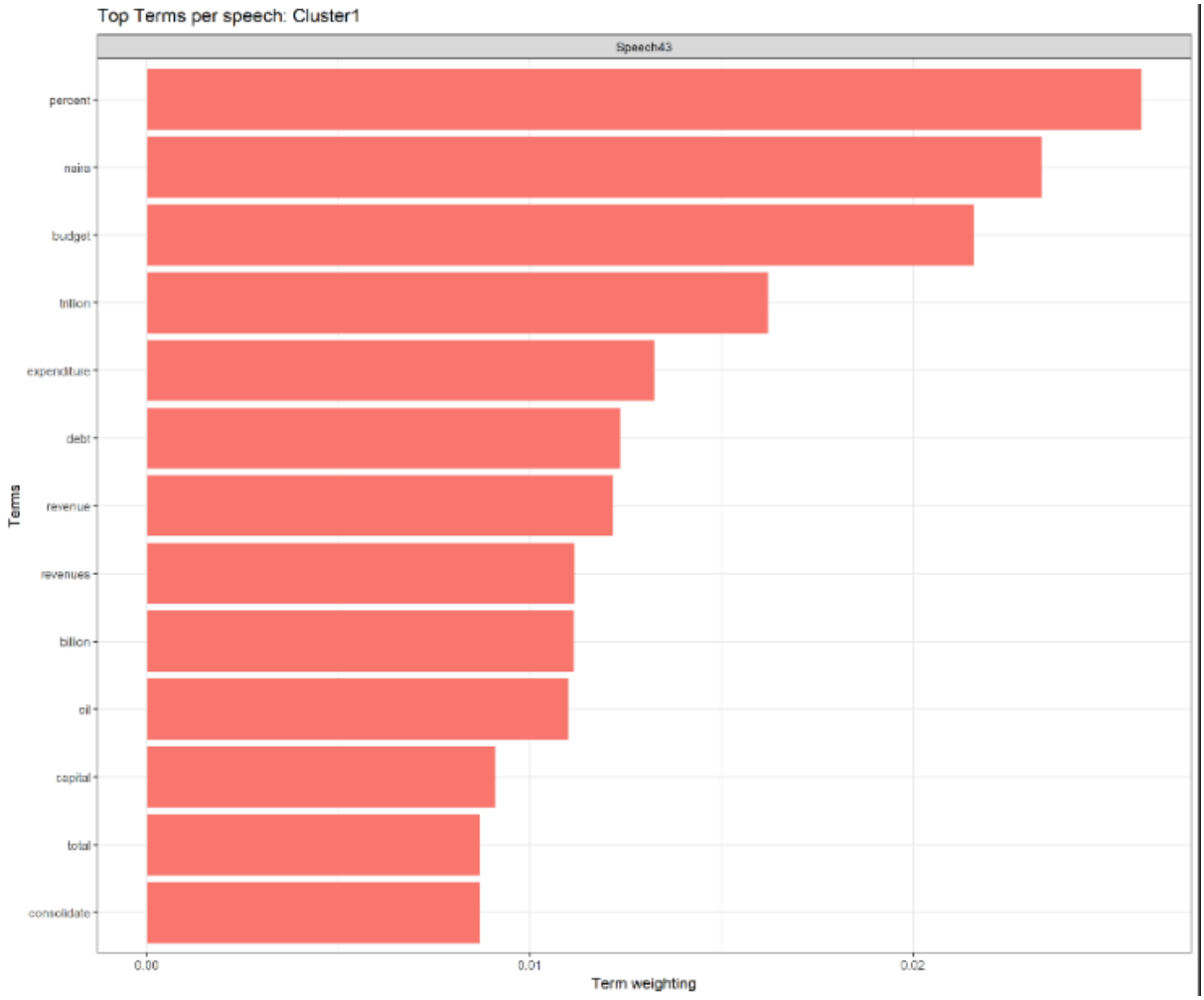
```
#tokenise the text using unnest function. This is the tidy version of the corpus.
tidyBuhariTxt <- buhariTxt%>%
select(text, id)%>%
group_by(id)%>%
unnest_tokens(word,text)%>%
ungroup()
```

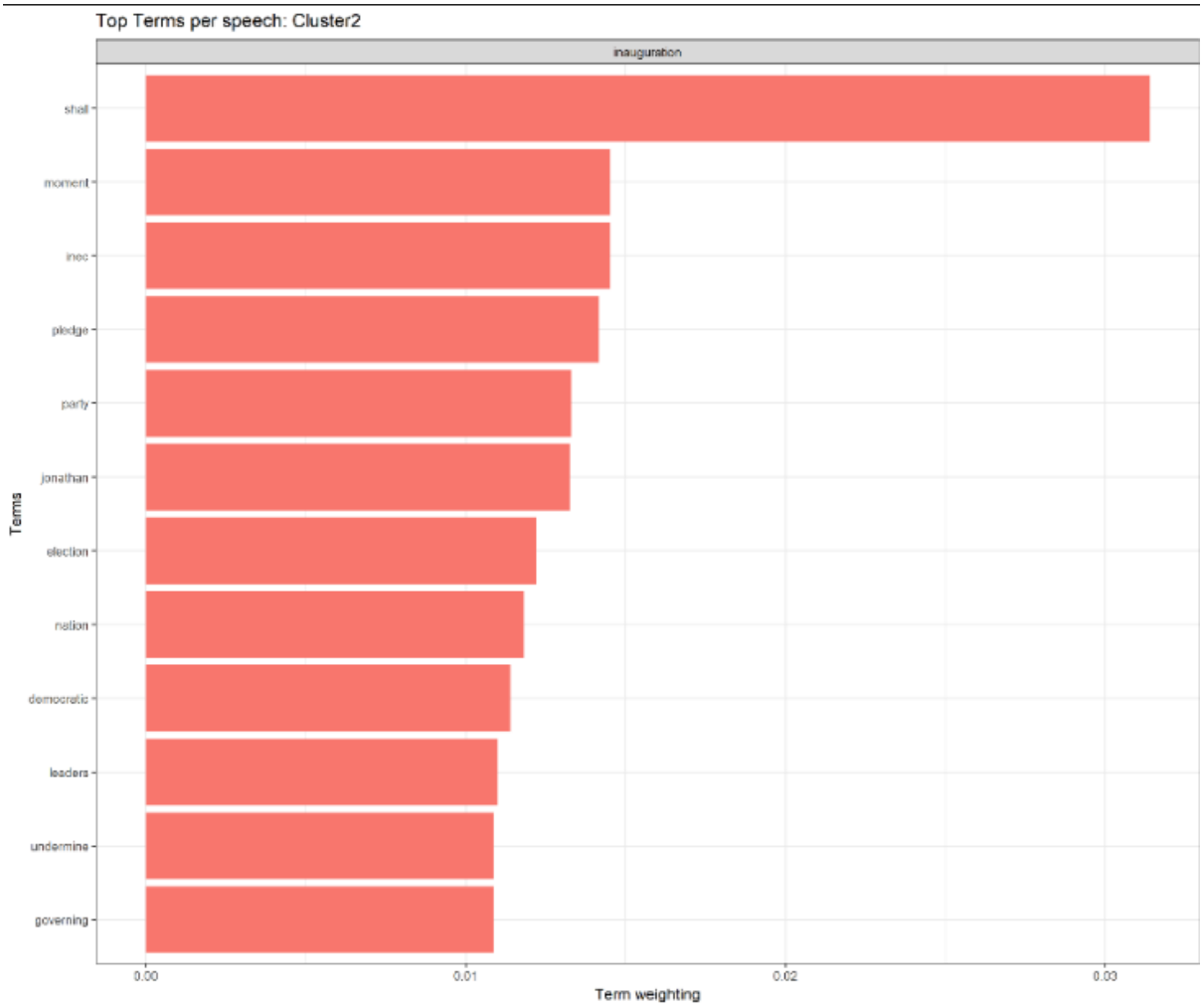
```
buhariTFIDF <- tidyBuhariTxt %>%
count(id, word, sort=TRUE) %>%
bind_tf_idf(word, id, n) %>%
arrange(desc(tf_idf))
```

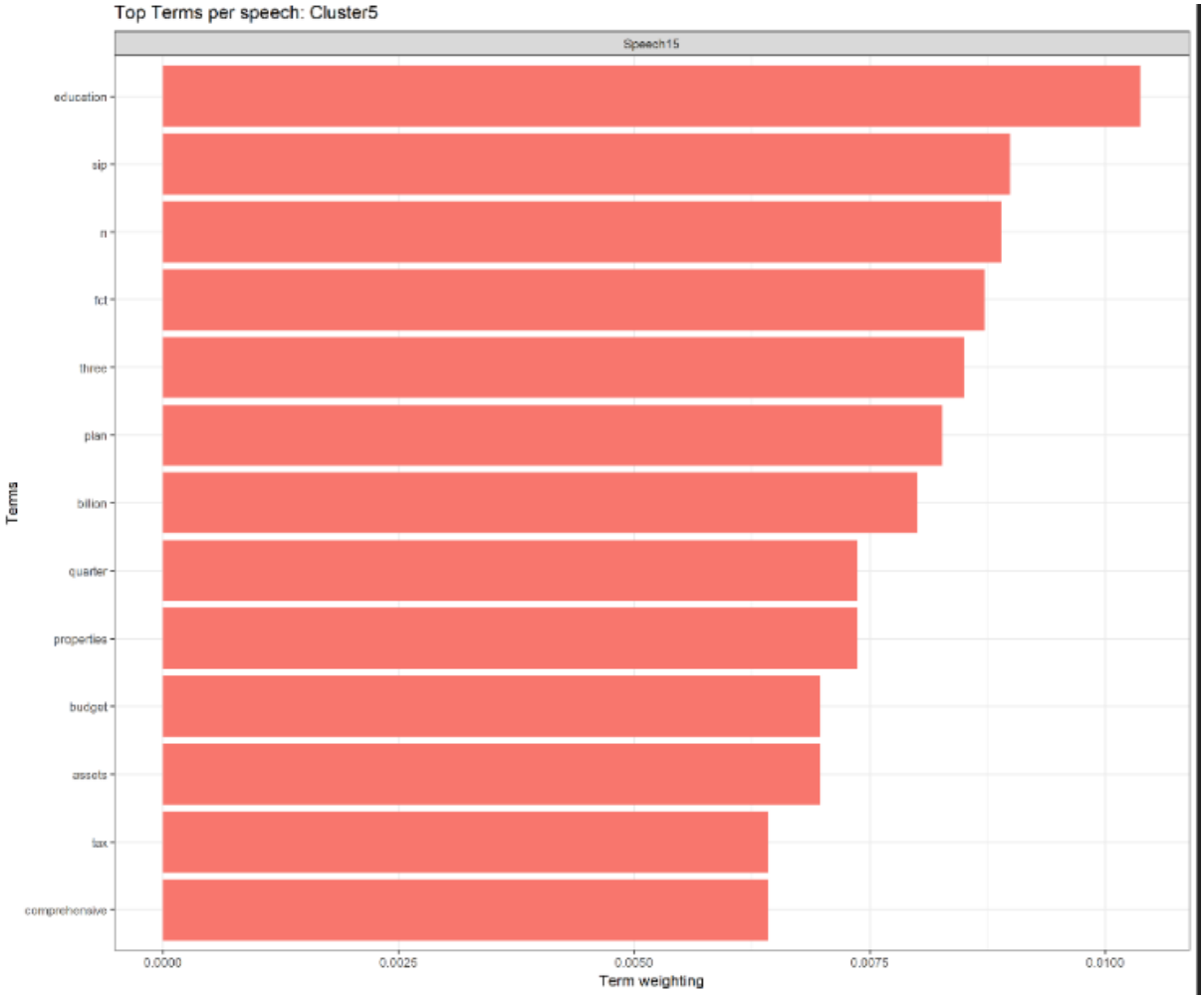
```
#Term weighting bar chary
buhariTFIDF %>%
  group_by(id) %>%
  filter(id == c('Speech36'))%>%
  top_n(12, tf_idf) %>%
  ungroup() %>%
  mutate(word = reorder(word, tf_idf)) %>%
  ggplot(aes(word, tf_idf, fill = id)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ id, scales = "free") +
  labs(x = "Terms", y = "Term weighting", title="Top Terms per speech: Cluster6") +
  theme_bw()+
  coord_flip()
ggsave('Cluster6.png', width=12, height=10)
```

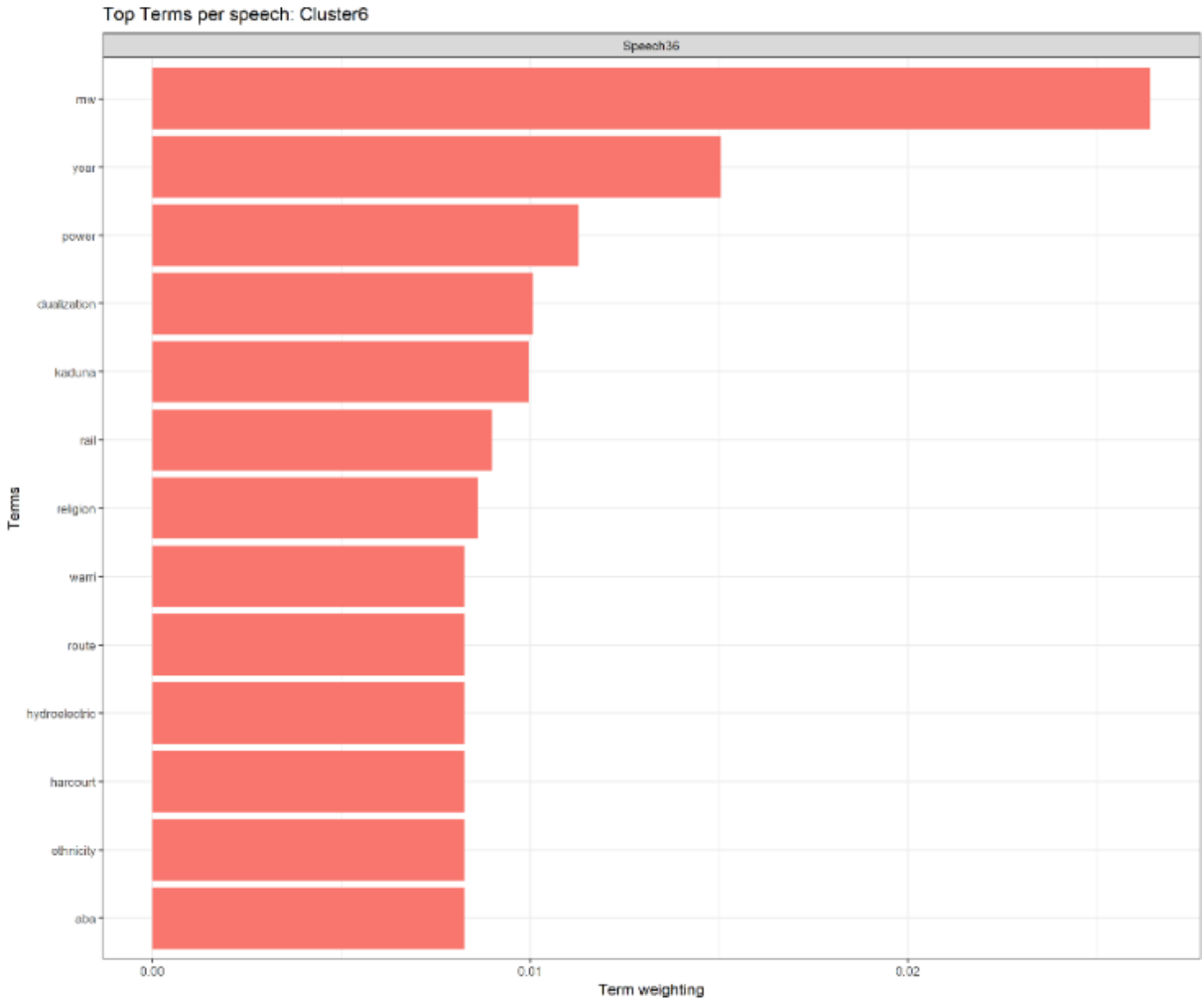
10. Cluster 1,2,5,6 terms:

Here a term weighting which ensures rare words are not penalised




Top Terms per speech: Cluster2
inauguration
shall
moment
inec
pledge
party
jonathan
election
nation
democratic
leaders
undermine
governing
Terms
0.00
0.01
0.02
0.03
Term weighting


Top Terms per speech: Cluster5
Speech15
education
sip
n
fct
three
plan
billion
quarter
properties
budget
assets
tax
comprehensive
Terms
0.0000
0.0025
0.0050
0.0075
0.0100
Term weighting


Top Terms per speech: Cluster6
Speech36
mw
year
power
dualization
kaduna
rail
religion
warri
route
hydroelectric
harcourt
ethnicity
aba
Terms
0.00
0.01
0.02
Term weighting