MUSA500 Homework6: UK Parliament Debate Text Analysis

Hang Zhao, Ling Chen, Jiahang Li

2023-12-20

Contents

The aim of this Markdown is to demonstrate the use of various text analysis tools in R, including text clustering, word clouds and sentiment analysis.

Introduction

Climate change has emerged as one of the most pressing global challenges of our time, with far-reaching implications for economies, ecosystems, and societies worldwide. The United Kingdom has played a pivotal role in addressing this issue, participating in significant international climate conferences, including the 26th to 28th Conference of the Parties (COP26-28) and the United Nations Framework Convention on Climate Change (UNFCCC). These gatherings have brought together leaders, policymakers, and experts from around the world to deliberate on strategies and commitments to mitigate climate change's impacts. The UK Parliament, as the supreme legislative body of the United Kingdom, has been instrumental in shaping the country's climate policy and influencing global climate initiatives. To gain valuable insights into the discussions, deliberations, and decisions made during these critical COP conferences, this report presents a comprehensive text data analysis of the UK Parliament Hansard debates concerning COP26, COP27, and COP28. By delving into the transcripts of these debates through text mining and NLP algorithms, we aim to explore the primary issues, concerns, positions, and policy considerations raised by Members of Parliament (MPs) surrounding COP26, COP27, and COP28. This will provide valuable insights that can inform future climate policy decisions and contribute to the broader public discourse on climate change.

Method

First, we will collect edited transcripts of debates in the British Parliament, obtained from the UK Parliament website. These transcripts will then be transformed into a Corpus, a vast and unstructured collection of texts, typically stored and processed electronically, used for statistical analysis and hypothesis testing. Subsequently, we will refine the data by eliminating punctuation and common English stop words (e.g., "a," "to"), which are often frequent terms that do not provide substantial information. To facilitate further analysis, we will convert myCorpus into a Document-Term Matrix (DTM) format. This conversion enables us to create a histogram and tabulation, revealing the distribution of term frequencies across all terms.

Next, we will generate a word cloud to visually represent the most frequently occurring terms. In this visualization, larger font sizes correspond to higher term frequencies. This graphical representation allows us to identify the prevailing topics and primary concerns addressed by the British government, providing us with deeper insights into the matter.

Following this, we will employ K-means clustering to categorize the prominent keywords from the three COP conferences. Subsequently, we will explore the characteristics of each cluster by examining the frequency of occurrence of each word within those clusters. This comprehensive analysis will enhance our understanding of the COP topic.

Additionally, we will conduct sentiment analysis on the COP text materials using the 'syuzhet' R package, which offers a variety of sentiment lexicons, also known as sentiment dictionaries. These lexicons contain words and phrases, each annotated with sentiment polarity information, indicating whether they convey positive, negative, or neutral emotions. It's worth noting that these lexicons intentionally exclude many words that are considered neutral, such as "hair," "purple," or "walk."

The 'syuzhet' package includes several lexicons, each with its unique characteristics:

NRC Lexicon: This lexicon, integrated into the 'syuzhet' package, features an extensive word list with annotations for various sentiment categories, including positivity, negativity, anger, fear, joy, sadness, and more. It simplifies sentiment analysis by linking words to specific emotional dimensions.

AFINN Lexicon: Another 'syuzhet' lexicon, AFINN, relies on precomputed sentiment scores for English words. Each word in this lexicon is assigned a sentiment score, ranging from negative (-5) to positive (5), indicating the intensity of sentiment. Its simplicity makes it a popular choice for basic sentiment analysis tasks.

Bing Lexicon: The Bing lexicon, also known as the Bing Liu lexicon, is a sentiment resource provided within the 'syuzhet' package. It categorizes words as either positive or negative based on their sentiment. It is commonly used for sentiment analysis applications and aids in determining text polarity.

Syuzhet (Jockers) Lexicon: Specifically designed for the 'syuzhet' package, the Syuzhet lexicon adopts a distinctive approach to sentiment analysis. It focuses on capturing sentiment by analyzing shifts in emotional intensity over time within text. Unlike the NRC, AFINN, and Bing lexicons, the Syuzhet lexicon delves into temporal sentiment dynamics, offering a unique perspective on emotional trajectories within textual content. It may prove valuable in analyses focused on temporal sentiment changes.

In sentiment analysis, we utilize these lexicons to assess the sentiment expressed in an entire text body. This typically involves aggregating sentiment scores, often achieved through averaging or summing, derived from all the words present in the text. However, it is important to note that for lengthy texts, the presence of positive and negative terms can sometimes counterbalance each other. Consequently, shorter texts, such as a few sentences or paragraphs, tend to yield more accurate sentiment assessments.

Result

Data Description

We used UK Parliament Hansard text data (https://hansard.parliament.uk/) debates about United Nations Climate Change Conference in recent years. Specifically, we loaded COP28, COP27, and COP26 text data to further investigate their common features based on the cluster analysis in the second half of the assignment.

```
##                                                          Length
## Climate Change_ Aims for COP 28 2023-11-28.txt           2
## COP 26 2021-05-25.txt                                    2
## COP 26 2021-11-16.txt                                    2
## COP 26 2022-03-28.txt                                    2
## COP 27 2022-10-27.txt                                    2
## COP 27 2022-11-15.txt                                    2
## COP 27_ Commitments 2022-11-24.txt                       2
## COP 28 2023-05-17.txt                                    2
## COP26 2021-03-10.txt                                     2
## COP26 2021-11-15.txt                                     2
## COP26 and Air Pollution 2021-11-02.txt                   2
## COP26 Outcomes 2022-07-20.txt                            2
## COP26_ Limiting Global  Temperature Rises 2021-10-21.txt 2
## COP27 2022-09-06.txt                                     2
## COP27 2022-11-09.txt                                     2
## COP27 2022-11-21.txt                                     2
## COP28 2023-11-16.txt                                     2
## COP28 2023-11-29.txt                                     2
## COP28 2023-12-14.txt                                     2
## G20 and COP26 World Leaders Summit 2021-11-03.txt        2
## Outcome of COP26 2021-12-01.txt                          2
## Outcome of COP26 2022-03-02.txt                          2
##                                                          Class             Mode
## Climate Change_ Aims for COP 28 2023-11-28.txt           PlainTextDocument list
## COP 26 2021-05-25.txt                                    PlainTextDocument list
## COP 26 2021-11-16.txt                                    PlainTextDocument list
## COP 26 2022-03-28.txt                                    PlainTextDocument list
## COP 27 2022-10-27.txt                                    PlainTextDocument list
## COP 27 2022-11-15.txt                                    PlainTextDocument list
## COP 27_ Commitments 2022-11-24.txt                       PlainTextDocument list
## COP 28 2023-05-17.txt                                    PlainTextDocument list
## COP26 2021-03-10.txt                                     PlainTextDocument list
## COP26 2021-11-15.txt                                     PlainTextDocument list
## COP26 and Air Pollution 2021-11-02.txt                   PlainTextDocument list
## COP26 Outcomes 2022-07-20.txt                            PlainTextDocument list
## COP26_ Limiting Global  Temperature Rises 2021-10-21.txt PlainTextDocument list
## COP27 2022-09-06.txt                                     PlainTextDocument list
## COP27 2022-11-09.txt                                     PlainTextDocument list
## COP27 2022-11-21.txt                                     PlainTextDocument list
## COP28 2023-11-16.txt                                     PlainTextDocument list
## COP28 2023-11-29.txt                                     PlainTextDocument list
## COP28 2023-12-14.txt                                     PlainTextDocument list
## G20 and COP26 World Leaders Summit 2021-11-03.txt        PlainTextDocument list
## Outcome of COP26 2021-12-01.txt                          PlainTextDocument list
## Outcome of COP26 2022-03-02.txt                          PlainTextDocument list
```

Data Preprocessing

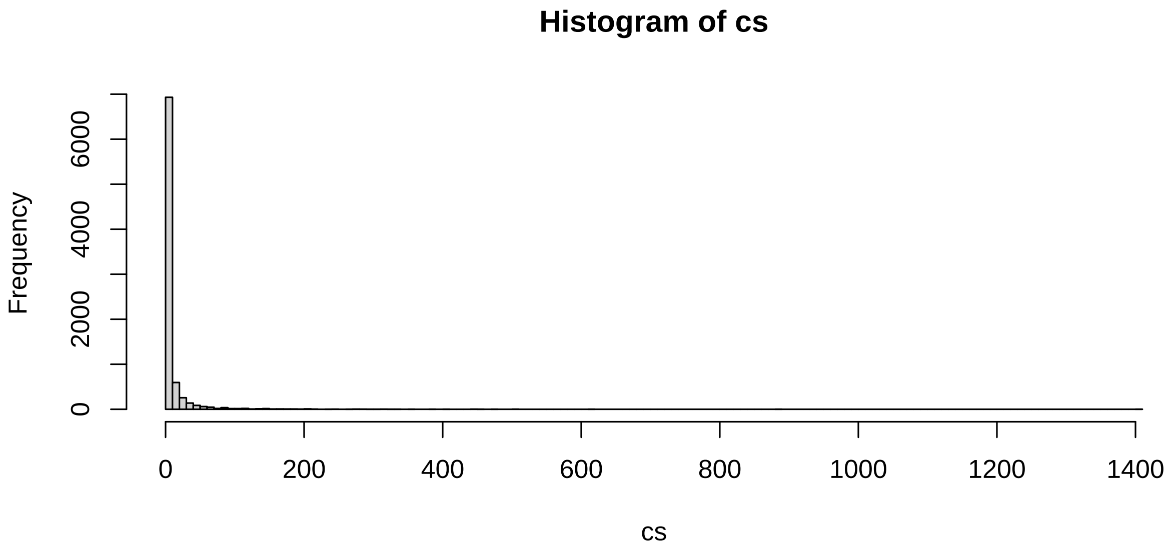
Firstly, we convert the text in all of these URLS into a Corpus. A text corpus (plural: *corpora*) “is a large and unstructured set of texts (nowadays usually electronically stored and processed) used to do statistical analysis and hypothesis testing, checking occurrences or validating linguistic rules within a specific language territory.”

Firstly, we convert a bunch of special characters(e.g., **@**,/,],**$**) to a space and by removing apostrophes. Then, numbers, punctuation and a list of English stop words are also removed. We also removed additional stop words like **parliament**, **minister**, **cop**, **gentleman**, **hon**, **lord**, **friend**.

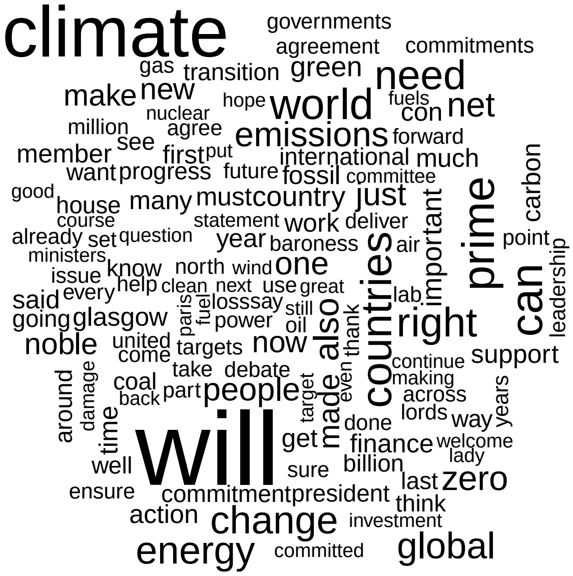
Document Term Matrix

Then, a term document matrix is created for further analysis.

Form the distribution of term frequency across all terms, we can see that there are a lot of terms that appear only a few times, while others appear over 200 times.



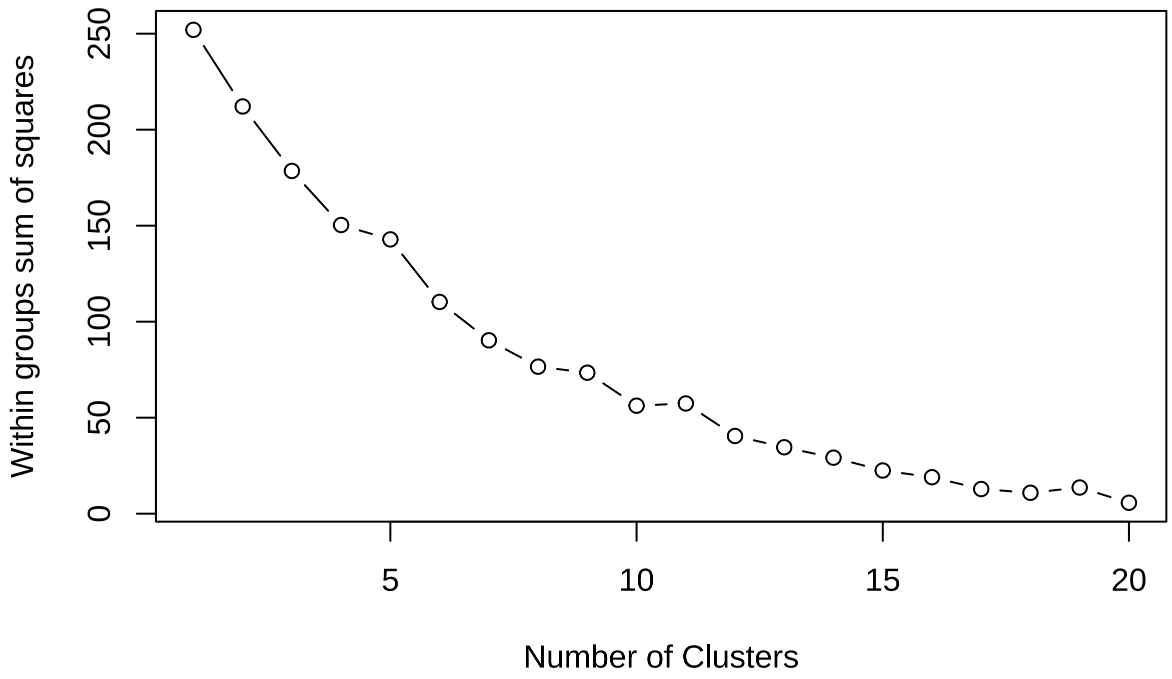
The world cloud further shows the terms that appear the most, for example, climate is sure to be the main topic across all the Climate Change Conferences. In general, the prominence of words ‘climate’,‘government’, ‘change’, ‘world’, ‘emissions’ underscores the global discourse on the action against global climate change. Also as indicated by the words ‘net’, ‘zero’,‘transition’, ‘target’, ‘clean’, ‘progress’,‘fossil’,‘nuclear’, we can see a tendency of achieving specific goals like cutting down greenhouse gases and aiming for carbon neutrality. Specifically, word like ‘glasgow’, ‘paris’,‘agreement’ further gives more contexts based on the references for The Glasgow Climate Pact and Paris Climate Agreement. It also reveals that there are ongoing dialogue happening based on the significant pledges made in previous Climate Change Conferences. Additionally, we can see an urgent and proactive attitudes towards the collaborative efforts in the future, with the words like ‘hope’, ‘future’, ‘commitment’,‘ensure’,‘action’.

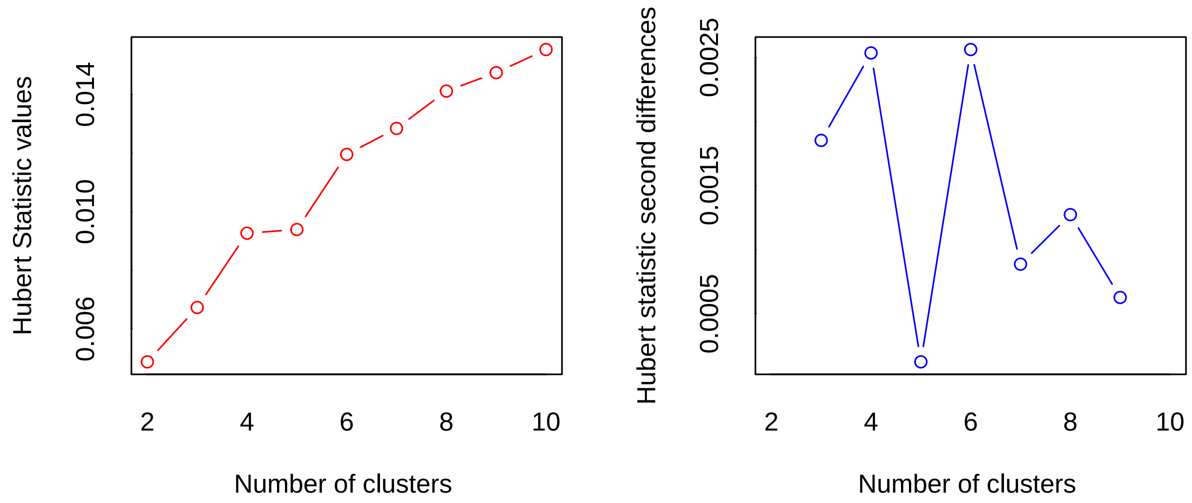


Furthermore, before moving to cluster analysis, we normalize all variables by removing all variables where the column sum is less than 350 and standardizing the columns and rows. Specifically, we calculate the the proportion of occurrences of each term's frequency within the same document, as well as a standardized version of frequency of occurrence of each term across different documents.

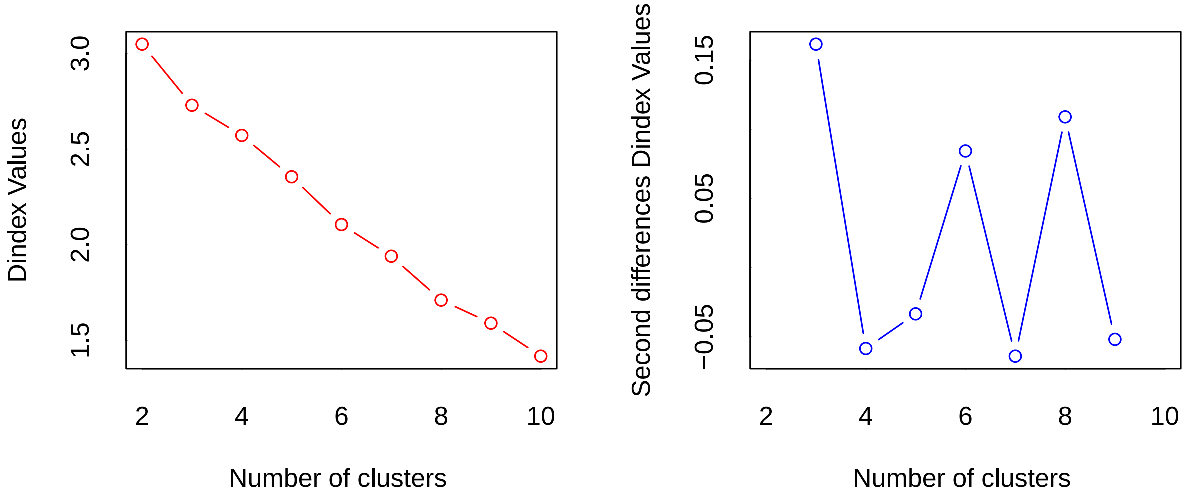
Text Clustering

By using the Scree Plot and the `NbClust` package in R, we further identify the optimal number of clusters to be 3. Specifically, the scree plot doesn't yield a clear recommendation. As such, we further look into NbClust approach which suggests the best number of clusters is 3, according to the majority rule.

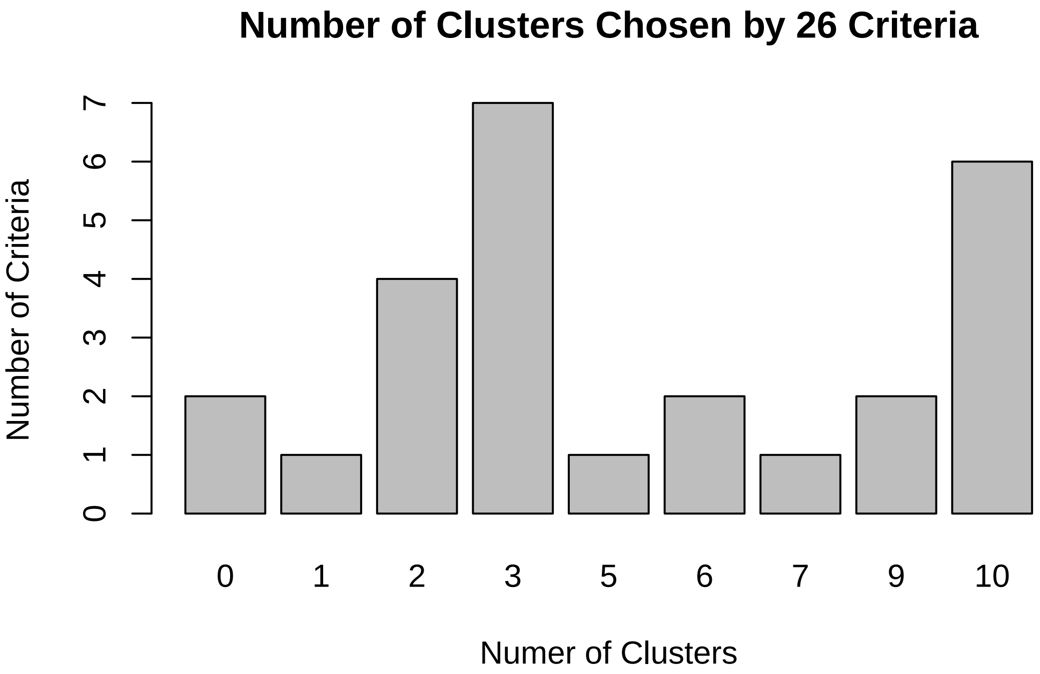




```
## *** : The Hubert index is a graphical method of determining the number of clusters.
##                 In the plot of Hubert index, we seek a significant knee that corresponds to a
##                 significant increase of the value of the measure i.e the significant peak in Hubert
##                 index second differences plot.
##
```



```
## *** : The D index is a graphical method of determining the number of clusters.
##                 In the plot of D index, we seek a significant knee (the significant peak in Dindex
##                 second differences plot) that corresponds to a significant increase of the value of
##                 the measure.
##
## *******************************************************************
## * Among all indices:
## * 4 proposed 2 as the best number of clusters
## * 7 proposed 3 as the best number of clusters
## * 1 proposed 5 as the best number of clusters
## * 2 proposed 6 as the best number of clusters
## * 1 proposed 7 as the best number of clusters
## * 2 proposed 9 as the best number of clusters
## * 6 proposed 10 as the best number of clusters
##
##                    ***** Conclusion *****
##
## * According to the majority rule, the best number of clusters is  3
##
##
## *******************************************************************
```

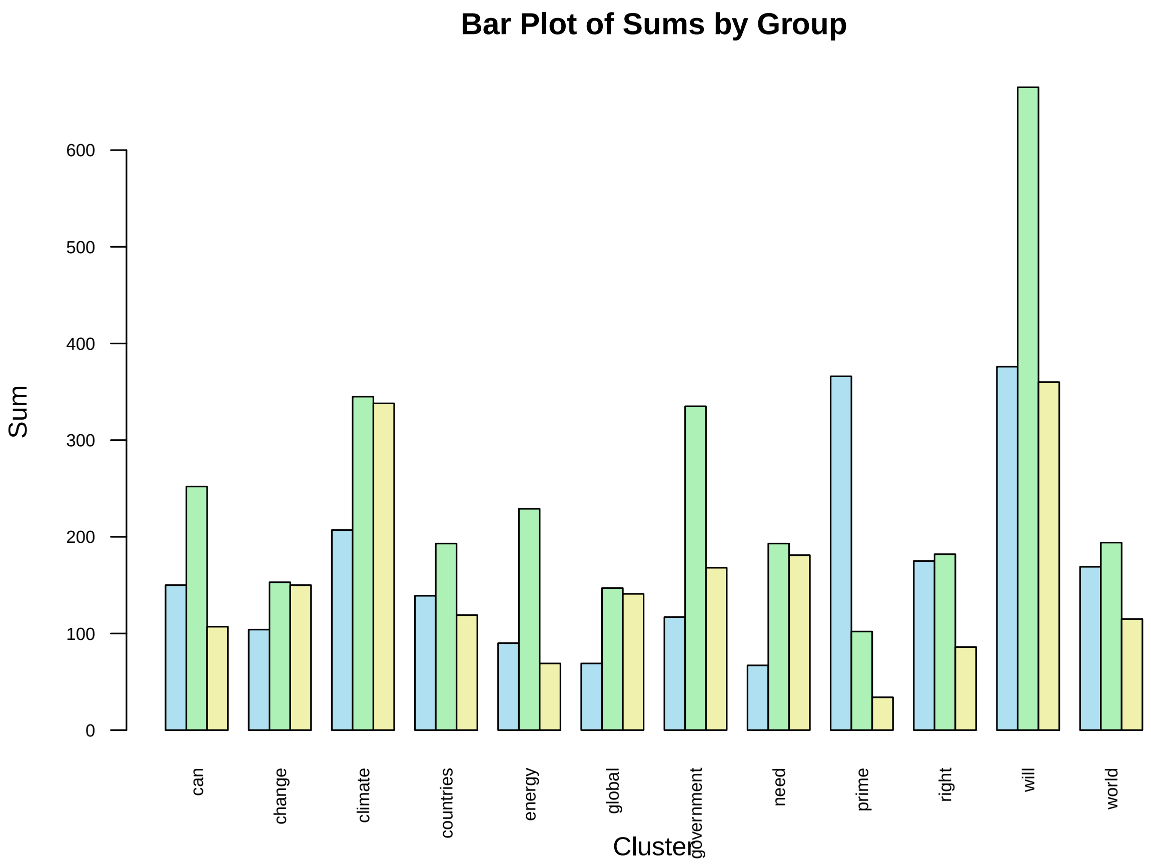


Furthermore, we can see which year's debate text falls in which cluster, as well as each cluster's size. We can see that half of them(11) are in cluster 2 and 5/6 respectively in cluster 1 and 3.

```
##            COP 26 March 10         COP 26 Lord May 25
##                          2                          3
##              COP 26 Oct 21 COP 26 Air Pollution Nov 2
##                          1                          2
##    COP 26 & G20 2021 Nov 3         COP 26 Lord Nov 16
##                          1                          2
##           COP 26 Nov 18/15               COP 26 Dec 1
##                          2                          2
##             COP 26 March 2         COP 27 Lord Oct 27
##                          2                          1
##             COP 26 July 20               COP 27 Sep 6
##                          2                          2
##         COP 27 Lord Oct 27               COP 27 Nov 9
##                          3                          3
##         COP 27 Lord Nov 15              COP 27 Nov 21
##                          1                          2
##  COP 27 Commit Lord Nov 24         COP 28 Lord May 17
##                          3                          2
##              COP 28 Nov 16          COP28 Lord Nov 28
##                          2                          1
##              COP 28 Nov 29              COP 28 Dec 14
##                          3                          3
```

```
## [1]  5 11  6
```

Finally, by looking at the number of times each of the terms appears in each cluster, we can see that different terms have different frequencies in the clusters. For example, ‘can’, ‘climate’,‘countries’,‘energy’,‘government’,‘will’ have higher frequencies in the cluster 2 generally.



Sentiment Analysis

Here, we will print the first 20 terms of each of the four lexicons mentioned in the method part above.

```
nrc <- syuzhet::get_sentiment_dictionary(dictionary="nrc")
head(nrc, n=20L)
```

```
##       lang           word sentiment value
## 1  english           abba  positive     1
## 2  english        ability  positive     1
## 3  english abovementioned  positive     1
## 4  english       absolute  positive     1
## 5  english     absolution  positive     1
## 6  english       absorbed  positive     1
## 7  english      abundance  positive     1
## 8  english       abundant  positive     1
## 9  english       academic  positive     1
## 10 english        academy  positive     1
## 11 english     acceptable  positive     1
## 12 english     acceptance  positive     1
## 13 english     accessible  positive     1
## 14 english       accolade  positive     1
## 15 english  accommodation  positive     1
## 16 english  accompaniment  positive     1
## 17 english     accomplish  positive     1
## 18 english   accomplished  positive     1
## 19 english accomplishment  positive     1
## 20 english         accord  positive     1
```

```r
afinn <- syuzhet::get_sentiment_dictionary(dictionary="afinn")
head(afinn, n=20L)
```

```
##          word value
## 1     abandon    -2
## 2   abandoned    -2
## 3    abandons    -2
## 4    abducted    -2
## 5   abduction    -2
## 6  abductions    -2
## 7       abhor    -3
## 8    abhorred    -3
## 9   abhorrent    -3
## 10     abhors    -3
## 11  abilities     2
## 12    ability     2
## 13     aboard     1
## 14    aborted    -1
## 15     aborts    -1
## 16   absentee    -1
## 17  absentees    -1
## 18    absolve     2
## 19   absolved     2
## 20   absolves     2
```

```r
bing <- syuzhet::get_sentiment_dictionary(dictionary="bing")
head(bing, n=20L)
```

```
##               word value
## 1               a+     1
## 2           abound     1
## 3          abounds     1
## 4        abundance     1
## 5         abundant     1
## 6       accessable     1
## 7       accessible     1
## 8          acclaim     1
## 9        acclaimed     1
## 10     acclamation     1
## 11        accolade     1
## 12       accolades     1
## 13   accommodative     1
## 14    accomodative     1
## 15      accomplish     1
## 16    accomplished     1
## 17  accomplishment     1
## 18 accomplishments     1
## 19        accurate     1
## 20      accurately     1
```

```
syuzhet <- syuzhet::get_sentiment_dictionary(dictionary="syuzhet")
head(syuzhet, n=20L)
```

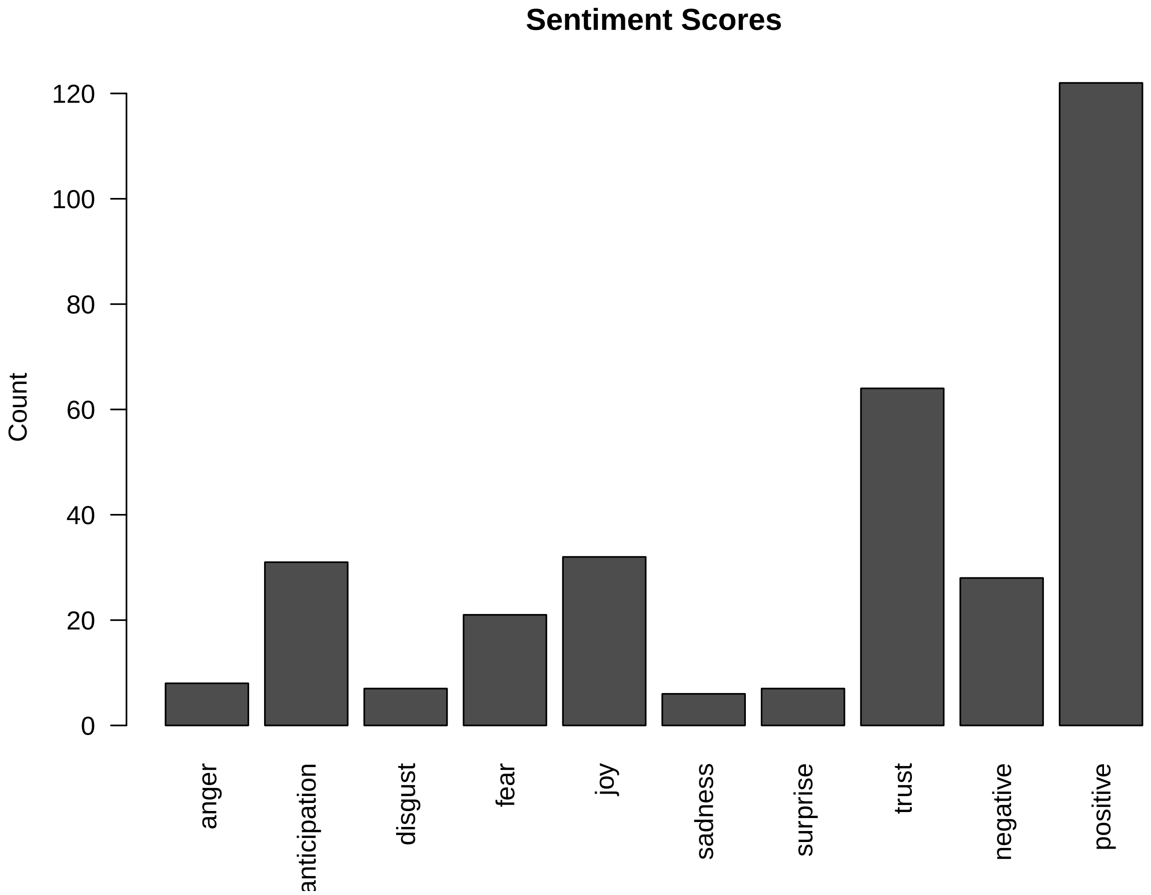
```
##            word value
## 1      abandon -0.75
## 2    abandoned -0.50
## 3    abandoner -0.25
## 4  abandonment -0.25
## 5     abandons -1.00
## 6     abducted -1.00
## 7    abduction -0.50
## 8   abductions -1.00
## 9     aberrant -0.60
## 10  aberration -0.80
## 11       abhor -0.50
## 12    abhorred -1.00
## 13   abhorrent -0.50
## 14      abhors -1.00
## 15   abilities  0.60
## 16     ability  0.50
## 17      abject -1.00
## 18      ablaze -0.25
## 19    abnormal -0.50
## 20      aboard  0.25
```

```
# `nrc` lexicon
get_nrc_sentiment("gorgeous")
```

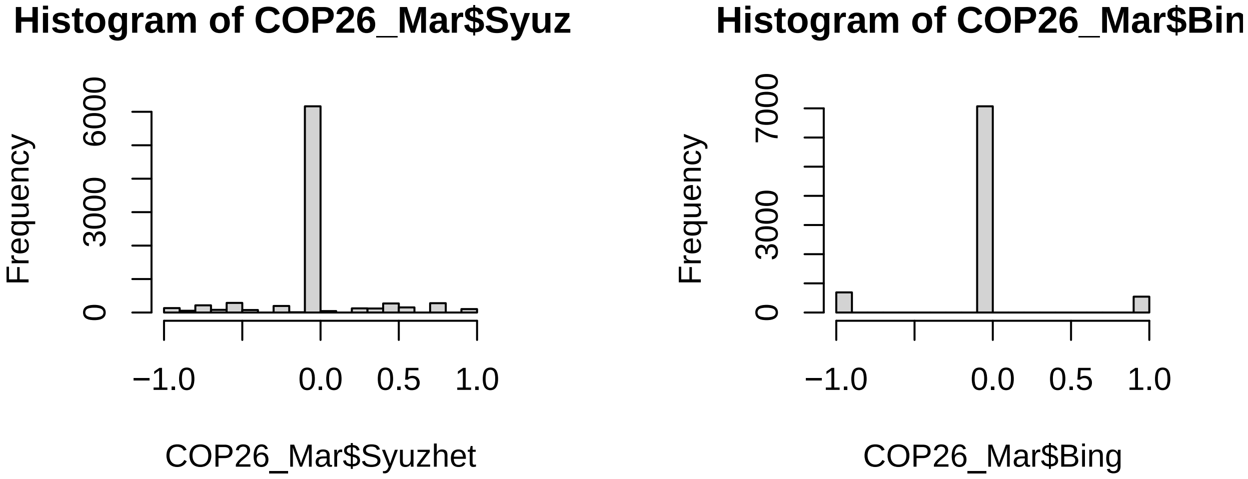
```
##   anger anticipation disgust fear joy sadness surprise trust negative positive
## 1     0            0       0    0   1       0        0     0        0        1
```

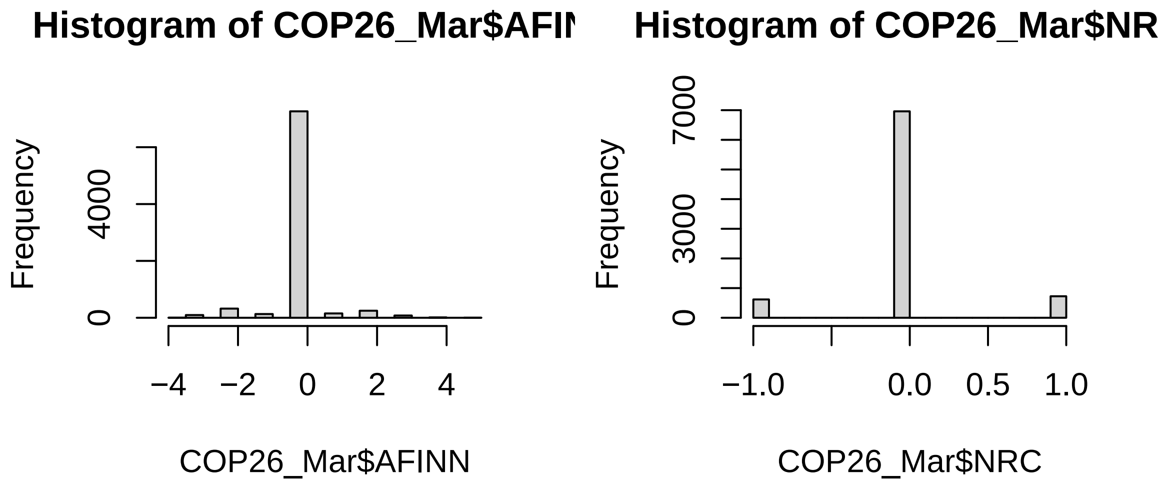
For this part, we take the COP.26.March.10 text as an example. the `get_nrc_sentiment` command obtains the sentiment score for each word in COP.26.March.10 that wasn't removed in the cleanning process above.

From the bar plot, we can see that most of the sentiment scores are either positive, trust, joy or anticipation. However, there is also a certain amount of negative emotion or fear.



The histogram further shows that most of the terms in COP.26.March.10 have relatively neutral sentiment scores.





Furthermore, we compare the sentiment scores from the different lexicons. We can see that in general, most of the terms are neutral, with about 6-12% of terms being negative and 6-13% of the terms being positive, depending on which lexicon we use. For Syuzhet, there're a higher frequency in either positive or negative terms relatively. Comparatively, AFINN yields a lower frequency in them.

```
##    Syuzhet Bing AFINN  NRC
## -1    1047  690   551  618
## 0     6165 7071  7263 6962
## 1     1093  544   491  725
```

```
##       Syuzhet       Bing      AFINN        NRC
## -1 0.1260686 0.08308248 0.06634557 0.07441300
## 0  0.7423239 0.85141481 0.87453341 0.83829019
## 1  0.1316075 0.06550271 0.05912101 0.08729681
```

Discussion

The word cloud analysis reveals a strong commitment by the British Government to distinct environmental objectives, particularly centered around reducing greenhouse gas emissions and achieving carbon neutrality. Furthermore, the clustering results highlight the prominence of these key themes, as approximately half of the dataset is categorized under cluster 2, which also has the highest cumulative frequency among all the keywords. Additionally, the sentiment analysis indicates a predominantly neutral standpoint within the debates, which is in line with the subjective nature of the climate change discourse.

Looking ahead, to enhance cluster clarity and gain a more comprehensive understanding of environmental topics, future analyses should involve expanding the dataset to encompass a wider range of environmental texts beyond COP conferences. Additionally, given the time constraints, we conducted sentiment analysis only for COP26 March. To obtain a more comprehensive result, future research should consider including additional text data for sentiment analysis. This approach will provide a more holistic perspective on sentiment across various environmental discussions and events.

References

1. Hill, Chelsey. 2023. “Sentiment Analysis (Lexicons)”. Rstudio-Pubs-Static.S3.Amazonaws.Com. https://rstudio-pubs-static.s3.amazonaws.com/676279_2fa8c2a7a3da4e7089e24442758e9d1b.html.

2. “Sentiment Analysis In R | R-Bloggers”. 2021. R-Bloggers. https://www.r-bloggers.com/2021/05/sentiment-analysis-in-r-3/.

3. Robinson, Julia. 2023. “2 Sentiment Analysis With Tidy Data | Text Mining With R”. Tidytextmining.Com. https://www.tidytextmining.com/sentiment.html.

4. “Text Mining: Sentiment Analysis · AFIT Data Science Lab R Programming Guide”. 2023. Afit-R.Github.Io. https://afit-r.github.io/sentiment_analysis.

5. “TDM (Term Document Matrix) And DTM (Document Term Matrix)”. 2023. Medium. https://medium.com/analytics-vidhya/tdm-term-document-matrix-and-dtm-document-term-matrix-8b07c58957e2.

6. “Text Clustering With R: An Introduction For Data Scientists”. 2018. Medium. https://medium.com/@SAPCAI/text-clustering-with-r-an-introduction-for-data-scientists-c406e7454e76.

7. “Introductory Tutorial To Text Clustering With R”. 2023. Rstudio-Pubs-Static.S3.Amazonaws.Com. https://rstudio-pubs-static.s3.amazonaws.com/445820_c6663e5a79874afdae826669a9499413.html.

8. “Library Guides: Text Mining & Text Analysis: Language Corpora”. 2023. Guides.Library.Uq.Edu.Au. https://guides.library.uq.edu.au/research-techniques/text-mining-analysis/language-corpora.