**Text Analysis in R**

Learn how to analyzing texts using the ‘**quanteda**’ package.

**Installing and loading the quanteda package**

The following line of code installs the quanteda package

install.packages(‘quanteda’)

We will be also using,

install.packages('quanteda.textstats')

install.packages('quanteda.textplots')

Once installed load all the above packages.

library(quanteda)

library(quanteda.textplots)

library(quanteda.textstats)

**Load the Data**

url = 'https://bit.ly/2QoqUQS'

data = read\_csv(url)

head(data)

A tibble: 6 x 5

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| paragraph  <dbl> | Date  <date> | President  <chr> | Party  <chr> | text  <chr> |
| 1 | 1790-01-08 | George Washin… | Other | I embrace with great satisfaction the opportunity which… |
| 2 | 1790-01-08 | George Washin… | Other | In resuming your consultations for the general good you… |
| 3 | 1790-01-08 | George Washin… | Other | Among the many interesting objects which will engage yo… |
| 4 | 1790-01-08 | George Washin… | Other | A free people ought not only to be armed, but disciplin… |
| 5 | 1790-01-08 | George Washin… | Other | The proper establishment of the troops which may be dee… |
| 6 | 1790-01-08 | George Washin… | Other | There was reason to hope that the pacific measures adop… |

OR get a sample of this data from quanteda which contains only 59 docs

corp ­<- data\_corpus\_inaugural

corp

**Corpus**

In text analysis, the term corpus is often used to refer to a collection of texts.

corp <- corpus(data, text\_field = 'text')

corp

Corpus consisting of **23,469 documents** and **4 docvars**.

text1 :

"I embrace with great satisfaction the opportunity which now ..."

text2 :

"In resuming your consultations for the general good you can ..."

text3 :

"Among the many interesting objects which will engage your at..."

text4 :

"A free people ought not only to be armed, but disciplined; t..."

text5 :

"The proper establishment of the troops which may be deemed i..."

text6 :

"There was reason to hope that the pacific measures adopted w..."

Docvars are the variables in the data. i.e. data, president, party, and text. We can view them with the function below

docvars(corp)

**The Document-term Matrix**

This is a matrix in which rows are documents, columns are terms, and cells indicate how often each term occurred in each document. This is like creating features for the computers with binary inputs.

This is also called a bag-of-words representation of texts, because documents have been reducted to only word frequencies (the matrix only shows how often a word occured in a text). Although, this format ignores a lot of important information regarding word order and syntax.

When we create a DTM there are some standard techniques for **preprocessing** the data. Specifically, we often want to filter out words that are not interesting, such as **stopwords** (e.g., the, it, is). Also, we need to normalize the terms in certain ways to help the computer understand that they mean the same thing. Most importantly, we often make all text **lowercase**, because a word like ‘Yes’ and ‘yes’ essentially means the same. We might also want to count different forms of a verb (e.g., go, going, gone) as one term. A simple way to achieve this is by **stemming** words. This technique cuts of certain parts of words to reduce them to their stem (e.g., go, go-ing, go-ne). We also have to remove the **punctuations** and other special characters.

dtm <- dfm(corp, stem=T, remove = stopwords('en'), remove\_punct=T, tolower = TRUE)

We now have a DTM with 23,469 documents and 20,201 features. The DTM is 99.82% sparse, which means that 99.82% of the cells in the DTM have the value zero. In general, DTM’s are very sparse, because individual documents (rows) contain only a small portion of all the words in the vocabulary (columns).

We cannot visualize the entire DTM. A subset is shown below

Document-feature matrix of: 23,469 documents, 20,201 features (99.82% sparse) and 4 docvars.

docs embrac great satisfact opportun now present congratul favor prospect public

text1 1 1 1 1 1 2 1 1 1 1

text2 0 0 0 0 0 1 0 0 0 0

text3 0 0 0 0 0 0 0 0 0 0

text4 0 0 0 0 0 0 0 0 0 0

text5 0 0 0 0 0 0 0 0 0 0

text6 0 0 0 0 0 0 0 0 0 0

[ reached max\_ndoc ... 23,463 more documents, reached max\_nfeat ... 20,191 more features ]

We can also trim the DTM to keep only words which have a minimum term frequency. Trimmingreturns a document by feature matrix reduced in size based on document and term frequency, usually in terms of a minimum frequency, but may also be in terms of maximum frequencies. Setting a combination of minimum and maximum frequencies will select features based on a range.

dtm <- dfm\_trim(dtm, min\_termfreq = 10)

dtm

Document-feature matrix of: 23,469 documents, **5,299 features** (99.36% sparse) and 4 docvars.

docs embrac great satisfact opportun now present congratul favor prospect public

text1 1 1 1 1 1 2 1 1 1 1

text2 0 0 0 0 0 1 0 0 0 0

text3 0 0 0 0 0 0 0 0 0 0

text4 0 0 0 0 0 0 0 0 0 0

text5 0 0 0 0 0 0 0 0 0 0

text6 0 0 0 0 0 0 0 0 0 0

[ reached max\_ndoc ... 23,463 more documents, reached max\_nfeat ... 5,289 more features ]

**Word clouds**

We can create a word cloud with quanteda’s textplot\_wordcloud() function. We can set the minimum wordcount to 50 to ignore all words that occurred less than 50 times.

textplot\_wordcloud(dtm, max\_words=50)



We can also get a frequency of words in the entire document

textstat\_frequency(dtm, n=10)

feature frequency rank docfreq group

1 state 9234 1 5580 all

2 govern 8581 2 5553 all

3 year 7250 3 4934 all

4 nation 6733 4 4847 all

5 congress 5689 5 4494 all

6 unit 5223 6 3715 all

7 can 4731 7 3628 all

8 countri 4664 8 3612 all

9 peopl 4477 9 3388 all

10 upon 4168 10 3004 all

This tells us that the word state occurred 9234 times in the entire document and was used mostly by all the presidents in their speech combined.

One thing we can do is analyse only President George Washington’ speeches

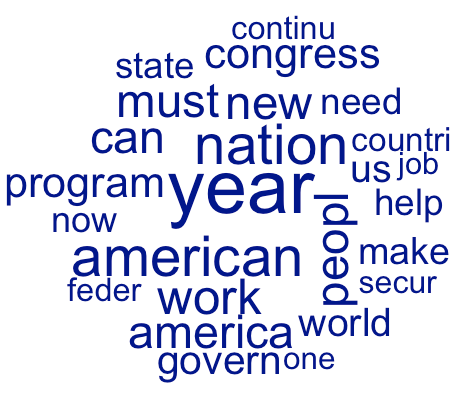
m\_GW dfm\_subset(dtm, President=="George Washington")

textplot\_wordcloud(m\_GW, max\_words = 25)



One more thing we can do is take a subset of the speeches, for example only speeches after 1945.

m\_postwar <- dfm\_subset(m, date > 1945)



**Keywords in context**

The wordcloud shows you the words that occur most frequently, but lack context. To get an idea of the context in which words are used, a **keyword-in-context** (kwic) is used. With quanteda, you can simply run the **kwic()** function on the corpus object and specifying the word you’re interested in.

For e.g.

I am doing the head to get only first 5 results

head( kwic(corp, 'state\*'), 5)

Keyword-in-context with 5 matches.

[text1, 31] recent accession of the important | state | of North Carolina to the

[text1, 41] the Constitution of the United | States | (of which official information

[text7, 6] The interests of the United | States | require that our intercourse with

[text9, 13] and measures of the United | States | is an object of great

[text16, 20] and interests of the United | States | are so obviously so deeply

**Corpus comparison**

If we want to see what words did George Washington more in comparison to other Presidents.

is\_GW <- docvars(dtm)$President == 'George Washington'

ts <- textstat\_keyness(dtm, is\_GW)

head(ts, 10)

feature chi2 p n\_target n\_reference

1 gentlemen 536.74498 0 20 64

2 militia 260.35125 0 20 142

3 whatsoev 229.33510 0 6 10

4 burthen 178.90431 0 7 21

5 agreeabl 163.84903 0 6 16

6 pennsylvania 140.67125 0 8 38

7 persuas 119.77537 0 5 15

8 particular 113.99607 0 29 610

9 creek 93.12124 0 6 31

10 requisit 87.67087 0 11 119

Here, we can see that George Washington used the word gentlemen 20 times, while it was used 64 times by any other presidents (combined). The chi2 score tells us the relative frequency of that word.

Similarly, you can do a tail()

tail(ts, 10)

feature chi2 p n\_target n\_reference

5290 help -11.32911 7.630144e-04 0 1415

5291 feder -13.90639 1.921441e-04 1 1979

5292 america -14.52987 1.379548e-04 0 1814

5293 work -15.37160 8.830540e-05 5 3042

5294 program -15.75795 7.198471e-05 0 1967

5295 peopl -17.13168 3.487575e-05 11 4466

5296 world -17.52944 2.828926e-05 2 2665

5297 american -19.89188 8.194745e-06 5 3625

5298 $ -30.79820 2.863038e-08 1 4083

5299 year -34.98432 3.323704e-09 13 7237

This tells us that G. Washington did not use the word year as much as other presidents. He used this word relatively less.