Assignment7_A

33140 (Sahil Naphade)

22/04/2020

Roll no. 33140 Batch: L9 Assignment 7: Generate wordcloud (1)

1. Install and load libraries

```
setwd("G:/College/SL6/Assignment7/")
# Install
#install.packages("tm")  # for text mining
#install.packages("SnowballC") # for text stemming
#install.packages("wordcloud") # word-cloud generator
#install.packages("RColorBrewer") # color palettes
#install.packages("wordcloud2") # word-cloud generator
#install.packages('readtext')
# Load
library("tm")
```

```
## Loading required package: NLP
```

```
library("SnowballC")
library("wordcloud")
```

```
## Loading required package: RColorBrewer
```

```
library("RColorBrewer")
library("wordcloud2")
library("readtext")
```

2. Read the text file, load as Corpus and inspect the file

```
#load the text

text <- readtext("../../Sl-VI DataSets/TextMining/NarendraModi.txt")

#Load the data as a corpus
docs <- Corpus(VectorSource(text))

#Inspect part of the content of the document
inspect(docs)
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 1
##
##
## Narendra Damodardas Modi (Gujarati: ['n??e?nd?r? d?a?mo?d???'d?a?s 'mo?d?i?] (About this sound liste
```

3. Preparation of data

a. Remove White spaces from data

```
  # remove white spaces
text_data <- tm_map(docs,stripWhitespace)
  inspect(text_data)
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 1
##
##
## Narendra Damodardas Modi (Gujarati: ['n??e?nd?r? d?a?mo?d???'d?a?s 'mo?d?i?] (About this sound listen
```

b. Convert all the words to lower alphabets

```
  # convert to lower
  text_data <- tm_map(text_data,tolower)
  inspect(text_data)
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 1
##
##
## narendra damodardas modi (gujarati: ['n??e?nd?r? d?a?mo?d???'d?a?s 'mo?d?i?] (about this sound listen
```

c. Remove the numbers

```
# Remove numbers
  text_data <- tm_map(text_data,removeNumbers)
  inspect(text_data)
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 1
##
##
## narendra damodardas modi (gujarati: ['n??e?nd?r? d?a?mo?d???'d?a?s 'mo?d?i?] (about this sound listen
```

d. Remove punctuations in the text

```
# Remove punctuations
  text_data <- tm_map(text_data,removePunctuation)
  inspect(text_data)
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 1
##
##
## narendra damodardas modi gujarati nendr damoddas modi about this sound listen born  september  is an
```

e. Remove stop-words

```
# Remove stop-words
  text_data <- tm_map(text_data,removeWords,stopwords('english'))
  inspect(text_data)
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 1
##
##
## narendra damodardas modi gujarati nendr damoddas modi   sound listen born  september    indian politi
```

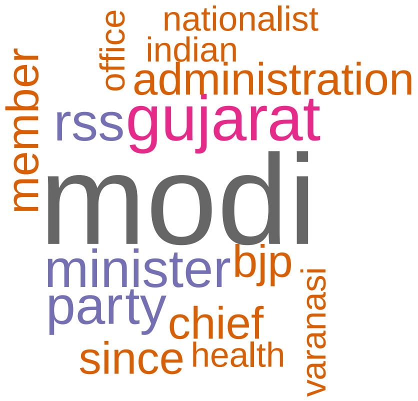
4. Load the data into Term Document Matrix, convert in a matrix, sort the data as increasing number of occurances, load as a dataframe

```r
  # Create a TDM
dtm <- TermDocumentMatrix(text_data)
matrix <- as.matrix(dtm)
words <- sort(rowSums(matrix),decreasing=TRUE)
df <- data.frame(word = names(words),freq=words)
str(df)
```

```
## 'data.frame':    195 obs. of  2 variables:
##  $ word: Factor w/ 195 levels "aayog","abolished",..: 114 75 113 128 161 6 24 30 112 171 ...
##  $ freq: num  12 5 4 4 4 3 3 3 3 3 ...
```

5. Generate Word Cloud

```r
set.seed(1234) # for reproducibility
wordcloud(words = df$word, freq = df$freq, min.freq = 1,max.words=15, random.order=FALSE,rot.per=0.35,c
```


nationalist
office
indian
administration
member
rss
gujarat
modi
minister
bjp
party
chief
varanasi
since
health