Text_Mining

Chapter 1 :

#setwd("D:\Machine_Learning\02.R\All_abt_R\13.Text_Mining\file\corpus")

```
list.files()
```

```
## [1] "cr.txt"             "jv.txt"             "negative-words.txt"
## [4] "pl.txt"             "positive-words.txt" "project.Rproj"
## [7] "text_mining_files"  "Text_Mining.R"      "text_mining.Rmd"
```

STEP 1 : Basic of Reading file

```
readLines("pl.txt")
```

```
## [1] "FULL TIME : Crystal Palace 0-1 Tottenham Hotspur"
## [2] ""
## [3] "And that's that! christian Eriksen's stylish snapshot is enough to secure victory! It wasn't muc
## [4] "spurs' goalscorer christen Eriksen celebrates with goalkeeper Hugo Lloris after the final whistl
## [5] "spurs' goalscorer christen Eriksen celebrates with goalkeeper Hugo Lloris after the final Whistl
## [6] "Mauricio Pochettino soaks up the applause from the visiting fans after the final whistle."
## [7] "whilst spurs boss Mauricio Pochettino soaks up the applause from the visiting fans. Photograph:
```

```
str(readLines("pl.txt"))
```

```
##  chr [1:7] "FULL TIME : Crystal Palace 0-1 Tottenham Hotspur" "" ...
```

```
paste( str(readLines("pl.txt"),collapse = " "))
```

```
##  chr [1:7] "FULL TIME : Crystal Palace 0-1 Tottenham Hotspur" "" ...
## character(0)
```

STEP 2 : Read text

```
text <- paste(readLines("pl.txt"),collapse = " ")
```

STEP 3 : Clean Text

remove punctuation \W looks for spaces and punctuation.

```
text2 <- gsub(pattern = "\\W",replace=" ", text)
```

remove numbers

```
text2 <- gsub(pattern = "\\d",replace =" ",text)
```

remove capital

```
text2 <- tolower(text2)
```

Install package

```
library(tm)
```

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

```
library(stopwords)
```

```
##
## Attaching package: 'stopwords'
```

```
## The following object is masked from 'package:tm':
##
##     stopwords
```

remove stopwords

```
text2 <- removeWords(text2,stopwords())
```

remove single letters

```
text2 <- gsub(pattern = "\\b[A-z]\\b{1}",replace=" ",text2)
```

remove whitespaces

```
text2 <- stripWhitespace(text2)
```

HERE END THE CLEANING PROCESS

Chapter 2 : Sentiment Analysis

```
library(stringr)
library(RColorBrewer)
library(wordcloud)
library(NLP)
```

STEP 1 : Split the text

```
textbag <- str_split(text2,pattern = "\\s+")
class(textbag)
```

```
## [1] "list"
```

STEP 2 : Unlist textbag

```
textbag <- unlist(textbag)
class(textbag)
```

```
## [1] "character"
```

```
str(textbag)
```

```
##  chr [1:65] "full" "time" ":" "crystal" "palace" "-" "tottenham" ...
```

STEP 3 : Lexicon positive and negative words.

#setwd("D:\Machine_Learning\02.R\All_abt_R\13.Text_Mining\file")

```
getwd()
```

```
## [1] "/cloud/project"
```

```
#setwd("/cloud/project/lexicon")
poswords <- scan("positive-words.txt",what = 'character',comment.char = ';')
negwords <- scan("negative-words.txt",what = 'character',comment.char = ';')
```

Now need to check how many match with postive & negative words.

```
match(textbag,poswords)
```

```
##  [1]   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA 1711   NA  560
## [15] 1578   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
## [29]   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
## [43]   NA   NA   NA   NA   NA   NA   NA   NA  685   NA   NA   NA   NA   NA
## [57]   NA   NA   NA   NA   NA   NA   NA   NA   NA
```

```
match(textbag,negwords)
```

```
##  [1] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
## [24] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
## [47] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
```

Is na give us true words

```
is.na(match(textbag,poswords))
```

```
##  [1]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [12] FALSE  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [23]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [34]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [45]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE
## [56]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
```

!is.na give us false words

```
!is.na(match(textbag,poswords))
```

```
##  [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [12]  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [23] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [34] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [45] FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE
## [56] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
sum(!is.na(match(textbag,poswords)))
```

```
## [1] 4
```

```
sum(!is.na(match(textbag,negwords)))
```

```
## [1] 0
```

```
score <- sum(!is.na(match(textbag,poswords))) - sum(!is.na(match(textbag,negwords)))
score
```

```
## [1] 4
```

visualize

```
#wordcloud(textbag)
#wordcloud(textbag,min.freq = 2)
#wordcloud(textbag,min.freq = 4,random.order = FALSE)
#wordcloud(textbag,min.freq = 4,random.order = FALSE,scale = c(3,0.5))
wordcloud(textbag,min.freq = 4,random.order = FALSE,scale = c(3,0.5),colors = rainbow(3))
```

```
## Warning in tm_map.SimpleCorpus(corpus, tm::removePunctuation):
## transformation drops documents
```

```
## Warning in tm_map.SimpleCorpus(corpus, function(x) tm::removeWords(x,
## tm::stopwords())): transformation drops documents
```


match
secure
pochettino
goalkeeper
celebrates
full
whistle
christian
crystal
mauricio
style
photograph
hugo
final
lloris
stylish
hotspur
time
eriksen
soaks
tom
spurs
christen
visiting
boss
fans
much
jenkins
victory
goalscorer
tottenham
whilst
matters
guardian
palace
enough
snapshot
applause

Chapter 3 : Working with Multiple files

STEP 1 : Import corpus

```
#file.choose()
#folder <- "D:\\Machine_Learning\\02.R\\All_abt_R\\13.Text_Mining\\file\\corpus"
#list.files(path = folder)
```

specify pattern

```
#filelist <- list.files(path = folder,pattern = "*.txt")
#filelist <- paste(folder,"\\",filelist,sep="")
#filelist <-
```

```
#filelist <- list.files(pattern = "*.txt")
```

STEP 2 : Read text document

```
#typeof(filelist)
#a <- lapply(filelist,FUN=readLines)
#corpus <- lapply(a,FUN=paste,collapse = " ")
```

STEP 3 : Cleaning process

1) Remove punctiontion

```
#corpus2 <- gsub(pattern = "\\W",replace=" ",corpus)
```

2) Remove digits

```
#corpus2 <- gsub(pattern = "\\d",replace=" ",corpus2)
```

3) change upper case to lowercase for simplicity

```
#corpus2 <- tolower(corpus2)
```

4) remove stopwords

```
#library(tm)
#library(NLP)
#corpus2 <- removeWords(corpus2,stopwords())
```

5) remove single words

```
#corpus2 <- gsub(pattern ="\\b[A-z]\\b{1}",replace=" ",corpus2)
```

6) remove whitespaces

```
#corpus2 <- stripWhitespace(corpus2)
```

STEP 4 : Visualization

```
#library(wordcloud)
#library(RColorBrewer)
#wordcloud(corpus2)
#wordcloud(corpus2,random.order = FALSE,colors = rainbow(3))
```

Chapter 4 : Comparison Wordcloud

```
#comparison.cloud(corpus2) # before corpus it will through error
```

```
#corpus3 <- Corpus(VectorSource(corpus2))
#corpus3
```

to open new graphic device use x11() function

```
#x11()
```

Term document matrix

term listed out in rows

Documents listed out in columns

```
#tdm <- TermDocumentMatrix(corpus3)
#tdm
```

```
#m <- as.matrix(tdm)
```

Change the default column names

```
#colnames(m) <- c("CR","JUVY","TOT")
#head(m)
```

```
#comparison.cloud(m)
```

Chapter 5 : Sentiment Analysis on Corpus

```
#corpus2
```

```
#setwd("/cloud/project/lexicon")
#opinion.lexicon.pos <- scan("positive-words.txt",what = 'character',comment.char = ';')
#opinion.lexicon.neg <- scan("negative-words.txt",what = 'character',comment.char = ';')
```

split document into bag of word

```
#library(tm)
#library(stringr)
#jj <- str_split(corpus2,pattern = "\\s+")
```

```
#lapply(jj, function(x){
 # sum(!is.na(match(x,opinion.lexicon.pos)))
#})
```

```
#jj <- str_split(corpus2,pattern = "\\s+")
```

```
#lapply(jj, function(x){
 # sum(!is.na(match(x,opinion.lexicon.neg)))
#})
```

Substract positive from neg

```
#lapply(jj, function(x){
  #sum(!is.na(match(x,opinion.lexicon.pos))) - sum(!is.na(match(x,opinion.lexicon.neg)))
#})
```

##Unlist

```
#SentimentScore <- unlist(lapply(jj, function(x){
 # sum(!is.na(match(x,opinion.lexicon.pos))) - sum(!is.na(match(x,opinion.lexicon.neg)))
#}))
```

```
#mean(SentimentScore)
```

```
#sd(SentimentScore)
```

```
#hist(SentimentScore)
```