**For Text Mining assignment**

**ONE:**

**1) Extract tweets for any user (try choosing a user who has more tweets)**

**2) Perform sentimental analysis on the tweets extracted from the above**

**Ans:**

User Choosed : Amitabh Bachhan(SrBachchan)

> library(twitteR)

> library(ROAuth)

> library(httpuv)

> library(base64enc)

**Authentication**

>user <- OAuthFactory$new(consumerKey="FXTquJNbgDG2dH81XYVqNZFAb", consumerSecret="3y0ALNFzJ8JKyxzFd0ba9FWSUpNSWhPisEIZOB6WCTtcGvP6SO", requestURL="https://api.twitter.com/oauth/request\_token", accessURL="https://api.twitter.com/oauth/access\_token",

authURL="https://api.twitter.com/oauth/authorize")

>save(user,file = "twitter authentication user.Rdata")

>getwd()

>setup\_twitter\_oauth("FXTquJNbgDG2dH81XYVqNZFAb",

"3y0ALNFzJ8JKyxzFd0ba9FWSUpNSWhPisEIZOB6WCTtcGvP6SO",

"529590041-qOXLd769cQEUTbXg3iRqCd33pC1K6xoORrGOMJDh",

"WlqZJwXFQzf64IuojkbKh1jdT5cnSY8U44pqmz6Sc1d4A")

**Twitter extraction**

> tweets\_ext <-userTimeline("SrBachchan",n=1000)

> tweets\_df <- twListToDF(tweets\_ext)

> write.csv(tweets\_df,"SrBachchan.csv")

> getwd()

[1] "C:/Users/Dell/Desktop/R directory"

> library(SnowballC)

> library(tm)

> library(wordcloud)

> library(topicmodels)

> library(RColorBrewer)

> text <- read.csv(file.choose())

> View(text)

> document <- Corpus(VectorSource(text$text))

> inspect(document[10])

<<SimpleCorpus>>

Metadata: corpus specific: 1, document level (indexed): 0

Content: documents: 1

[1] T 3444 -A Sajid - Farhad input : all good wishes !<U+0001F64F><U+0001F44F>\n\n https://t.co/phSp805iMw\nPiyush Shazia Back with \nDUS BAHANE… https://t.co/bDlKyxbaOv

**Function to clean the corpus**

> tospace <- function(x,pattern)gsub(pattern,"",x)

> document <- tm\_map(document,tospace,"/")

> document <- tm\_map(document,tospace,"@")

> document <- tm\_map(document,tospace,"\\|")

> document <- tm\_map(document,tospace,"#")

**Converting to lowerspace**

> document <- tm\_map(document,tolower)

**Removing numbers**

> document <- tm\_map(document,removeNumbers)

**Removing stopwords**

> document <- tm\_map(document,removeWords,stopwords("english"))

**Removing punctuations**

> document <- tm\_map(document,removePunctuation)

**Removing whitespace**

> document <- tm\_map(document,stripWhitespace)

> inspect(document[[10]])

<<PlainTextDocument>>

Metadata: 7

Content: chars: 105

t sajid farhad input good wishes uffuff httpstcophspimw piyush shazia back dus bahane… httpstcobdlkyxbaov

> inspect(document[10])

<<SimpleCorpus>>

Metadata: corpus specific: 1, document level (indexed): 0

Content: documents: 1

[1] t sajid farhad input good wishes uffuff httpstcophspimw piyush shazia back dus bahane… httpstcobdlkyxbaov

**Document term matrix**

> doctm <- TermDocumentMatrix(document)

> dim(doctm)

[1] 1949 582

> ctdm <- as.DocumentTermMatrix(doctm)

> rowtotals <- apply(ctdm,1,sum)

> ctdm.new <- ctdm[rowtotals>0,]

> lda <- LDA(ctdm.new,10)

> terms <- terms(lda,5)

> terms

Topic 1 Topic 2 Topic 3 Topic 4

[1,] "love" "uub" "walaalolylovely" "addzz"

[2,] "amitadaheur" "uuu" "ashokmistry" "proud"

[3,] "prashantkawadia" "uuuu" "janijasmine" "india"

[4,] "haha" "uuue" "juniorbachchan" "uuduuuuue"

[5,] "greetings" "ucuuu" "earthangel" "shpa"

Topic 5 Topic 6 Topic 7 Topic 8 Topic 9

[1,] "love" "wishes" "ashabachchan" "ufu" "manish"

[2,] "yes" "uff" "mukherjiratna" "uuuu" "jhund"

[3,] "gratitude" "film" "allaukhinaa" "uuu" "vikasbansalef"

[4,] "thank" "best" "nileshbodawala" "uub" "care"

[5,] "well" "pride" "gangulysunetra" "uue" "will"

Topic 10

[1,] "uuuu"

[2,] "work"

[3,] "uu\u0085"

[4,] "uuf"

[5,] "uufuu"

> topic <- terms(lda)

> tab <- table(names(topic),unlist(topic))

> head(tab)

addzz ashabachchan love manish ufu uub uuuu walaalolylovely

Topic 1 0 0 1 0 0 0 0 0

Topic 10 0 0 0 0 0 0 1 0

Topic 2 0 0 0 0 0 1 0 0

Topic 3 0 0 0 0 0 0 0 1

Topic 4 1 0 0 0 0 0 0 0

Topic 5 0 0 1 0 0 0 0 0

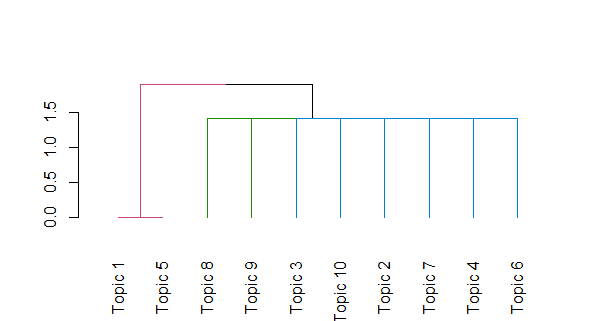
> library(cluster)

> library(dendextend)

> cluster <- hclust(dist(tab),method = "ward.D2")

> col\_bran <- color\_branches(cluster,k=3)

> plot(col\_bran)



**NLP**

> library(textcat)

> table(textcat(document))

afrikaans albanian basque

21 2 3

bosnian breton catalan

3 9 5

czech-iso8859\_2 danish dutch

3 11 10

english estonian finnish

128 9 3

french frisian hungarian

9 8 2

icelandic indonesian irish

4 13 2

italian latin latvian

4 47 2

malay manx middle\_frisian

8 6 13

nepali norwegian polish

1 1 3

portuguese romanian rumantsch

1 11 25

sanskrit scots scots\_gaelic

49 105 4

serbian-ascii slovak-ascii slovak-windows1250

3 1 4

slovenian-ascii slovenian-iso8859\_2 spanish

4 12 2

swahili swedish tagalog

16 7 4

> consider <- c(which(textcat(document)=="english"))

> documen2 <- document[consider]

> documen3 <- as.character(documen2)

> library(syuzhet)

> AmitabhBachhan\_tweets <- get\_sentences(documen3)

**Sentimental analysis**

> sentiments <- c("syuzhet","afinn","bing","nrc","stanford","custom")

> a <- NULL

> sent\_list <- NULL

>for(i in sentiments[1:4]){

sent\_list[[i]] <- get\_sentiment(documen3,method = i)

a[[i]] <- table(get\_sentiment(documen3,method = i))

}

> a

$syuzhet

0 54.15

2 1

$afinn

0 108

2 1

$bing

0 54

2 1

$nrc

0 61

2 1

> sent\_list

$syuzhet

[1] 54.15 0.00 0.00

$afinn

[1] 108 0 0

$bing

[1] 54 0 0

$nrc

[1] 61 0 0

**Plot bing**

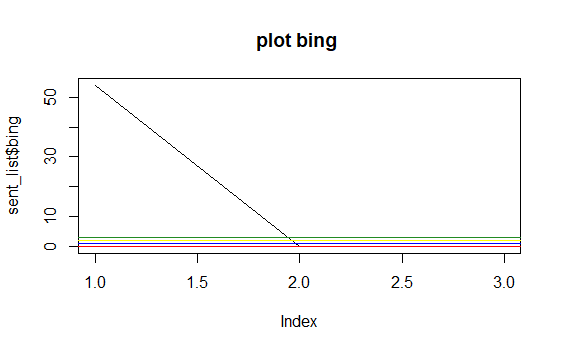
> plot(sent\_list$bing,type = "l",main="plot bing ")

> abline(h=0,col="red")

> abline(h=1,col="blue")

> abline(h=2,col="yellow")

> abline(h=3,col="forestgreen")



Plot nrc

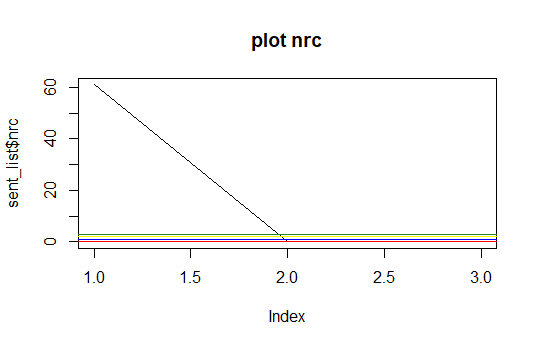
> plot(sent\_list$nrc,type="l",main="plot nrc")

> abline(h=0,col="red")

> abline(h=1,col="blue")

> abline(h=2,col="yellow")

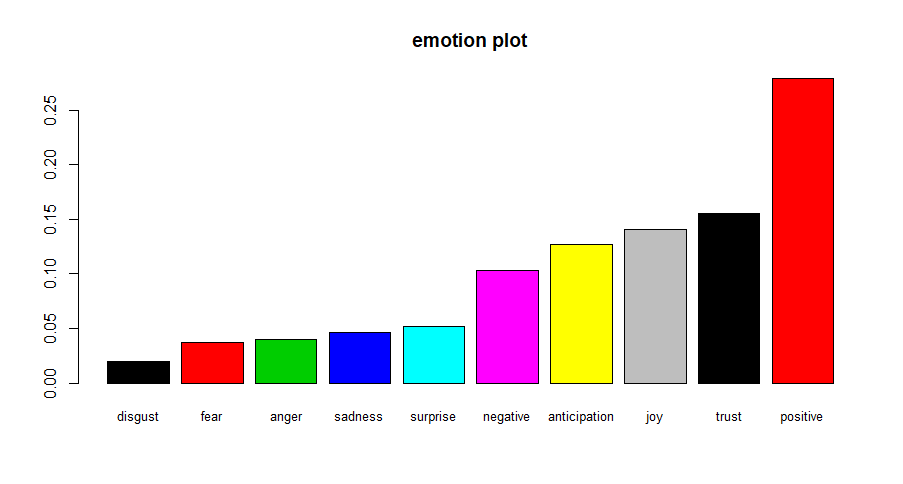
> abline(h=3,col="forestgreen")



> nrc\_data <- get\_nrc\_sentiment(documen3)

Emotions plot

> barplot(sort(colSums(prop.table(nrc\_data))),cex.names = 0.8,main="emotion plot",col =1:8)



**Wordcloud**

> freq <- rowSums(as.matrix(doctm))

> length(freq)

[1] 1949

> ord <- order(freq,decreasing = TRUE)

> freq[head(ord)]

uuuu uuu uub love

76 62 62 43

ufu walaalolylovely

38 36

> freq[tail(ord)]

httpstcoegynxkvv httpstcohxdmhnazom httpstcopawhmmtas

1 1 1

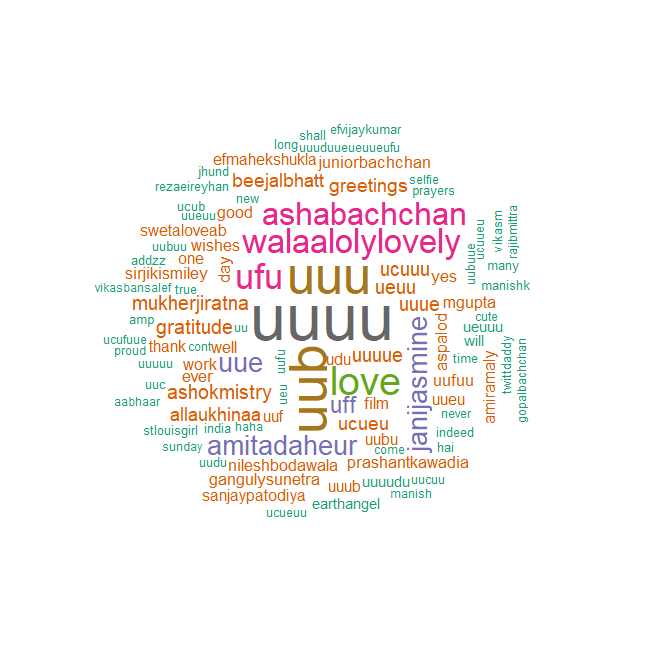
nitesh tiwari ubucufuudue

1 1 1

> df <- data.frame(word =names(freq),freq = freq)

> windows()

> wordcloud(words = df$word,freq = df$freq,min.freq = 3,max.words = 100,random.order = F,col = brewer.pal(20,"Dark2"))



> findFreqTerms(doctm,lowfreq = 8)

[1] "uufu" "uuuu" "uuuduueueuueufu"

[4] "ucueu" "udu" "ufu"

[7] "uuu" "uub" "film"

[10] "uff" "uubuue" "uue"

[13] "uueuu" "uuuue" "good"

[16] "wishes" "gratitude" "love"

[19] "well" "will" "ucuuu"

[22] "ueuu" "day" "uuue"

[25] "uuf" "uuub" "allaukhinaa"

[28] "gangulysunetra" "rezaeireyhan" "earthangel"

[31] "juniorbachchan" "mgupta" "sirjikismiley"

[34] "one" "uubu" "amitadaheur"

[37] "indeed" "aspalod" "efmahekshukla"

[40] "sanjaypatodiya" "ashokmistry" "swetaloveab"

[43] "walaalolylovely" "ashabachchan" "work"

[46] "ueuuu" "uueu" "uuuuu"

[49] "janijasmine" "mukherjiratna" "ever"

[52] "manishk" "yes" "beejalbhatt"

> findAssocs(doctm,terms ="uuuu",corlimit =0.3)

$uuuu

ucufuuu httpstcogitftltufc ubueufuue

0.47 0.42 0.42

uucufuu uueuuduu uuucuuu

0.42 0.42 0.42

uuf uuu uub

0.38 0.37 0.37

uue uubuu uubu

0.35 0.35 0.35

uudufubuuuf uuuduueueuueufu uueueuu

0.35 0.33 0.31

uuuub

0.31

> head(df,10)

word freq

httpstcooorozea httpstcooorozea 1

ucuf ucuf 2

ueuue ueuue 2

uueub uueub 3

uueud uueud 4

uueuuduuf uueuuduuf 2

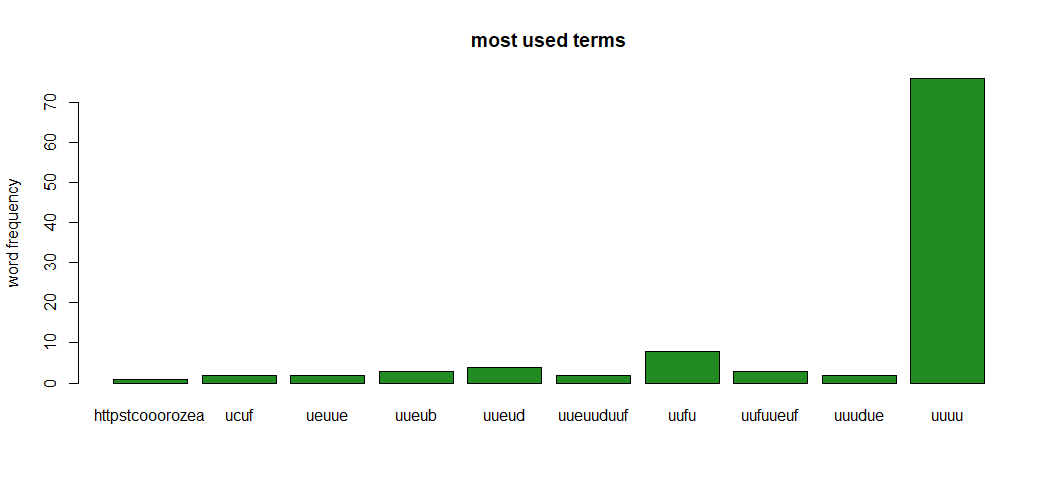
uufu uufu 8

uufuueuf uufuueuf 3

uuudue uuudue 2

uuuu uuuu 76

> barplot(df[1:10,]$freq,names.arg = df[1:10,]$word,col="forestgreen",main= "most used terms",ylab ="word frequency")



**uuuu is the most used word by Amitabh Bachhan (an Indian film actor)**