**Text Mining:**

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

setup\_twitter\_oauth(consumer\_key, consumer\_secret, access\_token, access\_secret)

|  |
| --- |
| > teste <- searchTwitter('sachin\_rt', n = 1000, since = '2020-01-01', retryOnRateLimit = 1e3)  > #registerTwitterOAuth(Cred)  > tweets\_ext <-userTimeline("sachin\_rt",n=1000)  > tweets\_df <- twListToDF(tweets\_ext)  > write.csv(tweets\_df,"sachin\_rt.csv")  > getwd()  [1] "/cloud/project"  > 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] Interesting start by Chris Morris. The first one left the batsman and the next one came back.\n\nWho says the ball do… https://t.co/R5suS2xmoR  > #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: 101  interesting start chris morris first one left batsman next one came back says ball … httpstcorsusxmor  > document <- tm\_map(document,stemDocument)  > inspect(document[10])  <<SimpleCorpus>>  Metadata: corpus specific: 1, document level (indexed): 0  Content: documents: 1  [1] interest start chris morri first one left batsman next one came back say ball … httpstcorsusxmor  > ################# document term matrix #####################  > doctm <- TermDocumentMatrix(document)  > dim(doctm)  [1] 3814 977  > 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 Topic 5 Topic 6  [1,] "thank" "…" "thank" "thank" "wish" "amp"  [2,] "alway" "can" "sport" "kind" "thank" "…"  [3,] "inning" "one" "amp" "word" "warm" "idbif"  [4,] "field" "play" "world" "cricket" "\U0001f642" "t…"  [5,] "togeth" "cricket" "laureussport" "game" "hope" "bat"  Topic 7 Topic 8 Topic 9 Topic 10  [1,] "के" "happi" "thank" "…"  [2,] "को" "birthday" "support" "back"  [3,] "से" "wish" "see" "hear"  [4,] "हम" "year" "soon" "memori"  [5,] "धन्यवाद" "good" "love" "time"  > topic <- terms(lda)  > tab <- table(names(topic),unlist(topic))  > head(tab)    … amp happi thank wish के  Topic 1 0 0 0 1 0 0  Topic 10 1 0 0 0 0 0  Topic 2 1 0 0 0 0 0  Topic 3 0 0 0 1 0 0  Topic 4 0 0 0 1 0 0  Topic 5 0 0 0 0 1 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 breton  3 4 1 6  catalan danish dutch english  7 2 6 274  esperanto french frisian hungarian  7 3 38 1  icelandic indonesian irish italian  1 17 1 6  latin latvian malay manx  18 1 6 8  middle\_frisian romanian rumantsch sanskrit  4 16 18 29  scots scots\_gaelic serbian-ascii slovak-ascii  475 3 1 1  slovenian-ascii spanish swahili swedish  1 3 2 11  tagalog turkish  2 1  > consider <- c(which(textcat(document)=="english"))  > documen2 <- document[consider]  > documen3 <- as.character(documen2)  > library(syuzhet)  > elon\_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))  + }  This warning is displayed once every 8 hours.  Call `lifecycle::last\_warnings()` to see where this warning was generated.  > a  $syuzhet  0 59.9  2 1  $afinn  0 113  2 1  $bing  0 48  2 1  $nrc  0 85  2 1  > sent\_list  $syuzhet  [1] 59.9 0.0 0.0  $afinn  [1] 113 0 0  $bing  [1] 48 0 0  $nrc  [1] 85 0 0  > #plot bing  > plot(sent\_list$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$nrc,type="l",main="plot nrc")  > abline(h=0,col="red")  > abline(h=1,col="blue")  > abline(h=2,col="yellow")  > abline(h=3,col="forestgreen")    >  > nrc\_data <- get\_nrc\_sentiment(documen3)  This warning is displayed once every 8 hours.  Call `lifecycle::last\_warnings()` to see where this warning was generated.  This warning is displayed once every 8 hours.  Call `lifecycle::last\_warnings()` to see where this warning was generated.  > #emotions plot  > barplot(sort(colSums(prop.table(nrc\_data))),cex.names = 0.8,main="emotion plot",col =1:8)    >  >  > #wordcloud  > #wordcloud    > freq <- rowSums(as.matrix(doctm))  > length(freq)  [1] 3814  > ord <- order(freq,decreasing = TRUE)  > freq[head(ord)]  thank wish happi amp … birthday  332 225 134 112 94 88  > freq[tail(ord)]  rahulakerkar httpstcoflbvgdpd individu extra  1 1 1 1  httpstcoayffroda mean…  1 1  > df <- data.frame(word =names(freq),freq = freq)  > wordcloud(words = df$word,freq = df$freq,min.freq = 3,max.words = 100,random.order = F,col = brewer.pal(20,"Dark2"))  > findFreqTerms(doctm,lowfreq = 8)  [1] "ball" "play"  [3] "shot" "alway"  [5] "amaz" "bat"  [7] "fun" "ipl"  [9] "watch" "even"  [11] "one" "\U0001f60b"  [13] "away" "day"  [15] "dear" "friend"  [17] "pass" "sinc"  [19] "word" "best"  [21] "birthday" "catch"  [23] "futur" "happi"  [25] "hope" "soon"  [27] "wish" "ahead"  [29] "success" "year"  [31] "forward" "look"  [33] "see" "start"  [35] "back" "good"  [37] "think" "wonder"  [39] "…" "last"  [41] "run" "came"  [43] "first" "famili"  [45] "hear" "sadden"  [47] "come" "inspir"  [49] "will" "amp"  [51] "covid" "need"  [53] "toward" "bless"  [55] "healthi" "stay"  [57] "season" "total"  [59] "congratul" "final"  [61] "win" "like"  [63] "special" "sport"  [65] "time" "’ve"  [67] "home" "experi"  [69] "never" "share"  [71] "thank" "world"  [73] "health" "game"  [75] "realli" "life"  [77] "new" "today"  [79] "children" "heart"  [81] "everyon" "forc"  [83] "indian" "warm"  [85] "way" "bowl"  [87] "mipaltan" "perform"  [89] "well" "wicket"  [91] "can" "take"  [93] "continu" "match"  [95] "batsmen" "get"  [97] "may" "’s"  [99] "import" "just"  [101] "show" "big"  [103] "bring" "imvkoh"  [105] "enjoy" "hard"  [107] "keep" "work"  [109] "th…" "much"  [111] "make" "contribut"  [113] "great" "partnership"  [115] "inning" "laureussport"  [117] "yuvstrong" "got"  [119] "\U0001f642" "cricket"  [121] "team" "moment"  [123] "शुभेच्छा" "हार्दिक"  [125] "love" "nice"  [127] "bowler" "support"  [129] "feel" "help"  [131] "lot" "rememb"  [133] "safe" "field"  [135] "kind" "also"  [137] "की" "से"  [139] "sir" "\U0001f1ee\U0001f1f3"  [141] "mani" "togeth"  [143] "t…" "beauti"  [145] "masterblastr" "\U0001f609"  [147] "india" "passion"  [149] "shri" "vvslaxman"  [151] "celebr" "fit"  [153] "test" "\U0001f3cf"  [155] "care" "vinodkambl"  [157] "\U0001f61c" "peopl"  [159] "cup" "msdhoni"  [161] "fond" "memori"  [163] "seri" "part"  [165] "pray" "proud"  [167] "selfless" "\U0001f64f\U0001f3fc"  [169] "meet" "appreci"  [171] "icc" "live"  [173] "\U0001f600" "dream"  [175] "mumbai" "across"  [177] "के" "को"  [179] "recent" "धन्यवाद"  [181] "\U0001f64f" "idbif"  [183] "mean" "हम"  [185] "nation" "joy"  [187] "और" "में"  [189] "teamindia" "cricketworldcup"  > findAssocs(doctm,terms ="uuuu",corlimit =0.3)  $uuuu  numeric(0)  > head(df,10)  word freq  backli… backli… 1  ball ball 15  clean clean 3  httpstcoinqysooc httpstcoinqysooc 1  nichola nichola 2  pack pack 1  play play 55  power power 7  shot shot 8  stanc stanc 2  > barplot(df[1:10,]$freq,names.arg = df[1:10,]$word,col="forestgreen",main= "most used terms",ylab ="word frequency")    > # Bar plot of the frequency for the top10  > barplot(df[1:10,]$freq, las = 2,  + names.arg = df[1:10,]$word,  + col ="lightblue", main ="Most frequent words",  + ylab = "Word frequencies") |
|  |
| |  | | --- | | > | |

