M8L3

Xueyi Fan

July 10, 2016

# Assignment:

* Load the file ML.Tweets.csv and ML.Tweets.New.csv (it is online at '<http://nikbearbrown.com/YouTube/MachineLearning/M08/ML.Tweets.csv>' and '<http://nikbearbrown.com/YouTube/MachineLearning/M08/ML.Tweets.New.csv>' )
* Do the following with ML.Tweets.csv:
  + Extract and rank a list of the important hashtags (using td-idf or word entropy)
  + Cluster the tweets using these hashtags .
  + Optional - give the the clusters names based on their dominant hashtags.
  + Classify the tweets in ML.Tweets.New.csv using the cluster lables generated from ML.Tweets.csv.
  + Use the qdap polarity function to score the polarity of the tweets in ML.Tweets.csv.
  + Would creating a custom polarity.frame - A dataframe or environment containing a dataframe of positive/negative words and weights - based on the tags and words in these tweets improve the polarity score? Try it.

# Answer:

## Load the file ML.Tweets.csv and ML.Tweets.New.csv (it is online at '<http://nikbearbrown.com/YouTube/MachineLearning/M08/ML.Tweets.csv>' and '<http://nikbearbrown.com/YouTube/MachineLearning/M08/ML.Tweets.New.csv>' )

library("RTextTools")

## Loading required package: SparseM

##   
## Attaching package: 'SparseM'

## The following object is masked from 'package:base':  
##   
## backsolve

library("e1071")  
library("qdap")

## Warning: package 'qdap' was built under R version 3.2.5

## Loading required package: qdapDictionaries

## Loading required package: qdapRegex

## Loading required package: qdapTools

## Loading required package: RColorBrewer

##   
## Attaching package: 'qdap'

## The following object is masked from 'package:base':  
##   
## Filter

library("tm")

## Loading required package: NLP

##   
## Attaching package: 'NLP'

## The following object is masked from 'package:qdap':  
##   
## ngrams

##   
## Attaching package: 'tm'

## The following objects are masked from 'package:qdap':  
##   
## as.DocumentTermMatrix, as.TermDocumentMatrix

tweets.url <- "http://nikbearbrown.com/YouTube/MachineLearning/M08/ML.Tweets.csv"  
tweets.new.url <- "http://nikbearbrown.com/YouTube/MachineLearning/M08/ML.Tweets.New.csv"  
  
tweets <- read.csv(url(tweets.url), header=F)  
tweets.new <- read.csv(url(tweets.new.url), header=F)  
  
str(tweets)

## 'data.frame': 139318 obs. of 1 variable:  
## $ V1: Factor w/ 81878 levels " "," #BigData, Small #Credit https://t.co/SAhUzEwl9Z - @FastCompany writes on @OmidyarNetwork report re: #EmergingMarkets #Financia"| \_\_truncated\_\_,..: 1464 81298 79544 57502 81319 36057 54007 58073 2050 57999 ...

str(tweets.new)

## 'data.frame': 139753 obs. of 1 variable:  
## $ V1: Factor w/ 77835 levels " #BigData habla y el sistema financiero debe saber escucharlo via @Forbes\_Mexico @SunGard http://t.co/FAtAJVSuKU",..: 67103 77165 27896 62592 58810 67716 66322 75387 71673 48534 ...

Do the following with ML.Tweets.csv:

### Extract and rank a list of the important hashtags (using td-idf or word entropy)

data preprocessing methods are based on [Zangerle el. paper](http://ceur-ws.org/Vol-730/paper7.pdf)

# multiple precessing steps  
#remove all messgaes not containing hashtags at all.  
has\_tag <- function(x){  
 grepl("#[\_a-zA-Z1-9]+",x)  
}  
  
tag\_result <- apply(tweets, 1, has\_tag)  
tweets\_ht <- tweets[which(tag\_result),]  
  
head(tweets\_ht)

## [1] http://t.co/YbpPiAAinW No. 13 Davenport baseball downs Viterbo to complete three-game sweep http://t.co/J4HQQvDp88 #GrandRapids  
## [2] RT @celebrate1837: RT @UConnMBB: Boatright hits them both. Five-point game (53-48). | #UConn #BleedBlue   
## [3] RT @Noeblanch: Yeh, Press Start, let's play! Why it's not game over for #gamification. @scoopit http://t.co/FIYfp1fM3Z   
## [4] Beginning Big Data with Power BI and Excel 2013 http://t.co/89XrOzIDXZ #computerscience #microsoftexcel #bigdata #datamodeling   
## [5] RT @Marketineer: 7 Factors Limiting Benefits of #BigData #Healthcare http://t.co/I7XvjCBcBs http://t.co/HAhVwBWGVD   
## [6] RT http://t.co/NujgrEaML3 Applying machine learning to medicine. Good stuff: http://t.co/bIJNgnxcab #drsherriworth   
## 81878 Levels: ...

#remove all non-english message  
clean\_sub <- function(x){  
 x <- tolower((x))  
 #remove all puncuation except #  
 gsub("(?!#)[[:punct:]]", "",perl=T,  
 #remove all control characters  
 gsub("[[:cntrl:]]", "",  
 #remove all digit  
 gsub("\\d+","",  
 #remove the users  
 gsub("@([\_a-zA-Z1-9]+)","",  
 #remove all url  
 gsub("(https?://[^ ]+)","",x)))))  
}  
  
tweets\_clean <- apply(as.data.frame(tweets\_ht), 1, clean\_sub)  
head(tweets\_clean)

## [1] " no davenport baseball downs viterbo to complete threegame sweep #grandrapids"   
## [2] "rt rt boatright hits them both fivepoint game #uconn #bleedblue"   
## [3] "rt yeh press start lets play why its not game over for #gamification "   
## [4] "beginning big data with power bi and excel #computerscience #microsoftexcel #bigdata #datamodeling"  
## [5] "rt factors limiting benefits of #bigdata #healthcare "   
## [6] "rt applying machine learning to medicine good stuff #drsherriworth"

#Extract all hashtags  
hashtags\_re <- function(line){  
 m <- gregexpr("#([\_a-zA-Z1-9]+)", line)  
 return(regmatches(line,m))  
}  
tweets\_clean <- as.data.frame(tweets\_clean)  
result\_hashtag <- apply(tweets\_clean, 1, hashtags\_re)  
head(result\_hashtag)

## [[1]]  
## [[1]]$tweets\_clean  
## [1] "#grandrapids"  
##   
##   
## [[2]]  
## [[2]]$tweets\_clean  
## [1] "#uconn" "#bleedblue"  
##   
##   
## [[3]]  
## [[3]]$tweets\_clean  
## [1] "#gamification"  
##   
##   
## [[4]]  
## [[4]]$tweets\_clean  
## [1] "#computerscience" "#microsoftexcel" "#bigdata"   
## [4] "#datamodeling"   
##   
##   
## [[5]]  
## [[5]]$tweets\_clean  
## [1] "#bigdata" "#healthcare"  
##   
##   
## [[6]]  
## [[6]]$tweets\_clean  
## [1] "#drsherriworth"

#rank the hashtag according to their tf-idf  
hashtags.corpus <- Corpus(VectorSource(result\_hashtag))  
#inspect(hashtags.corpus)  
hashtags.tdm <- TermDocumentMatrix(hashtags.corpus, control=list(removePunctuation=T))  
  
dim(hashtags.tdm)

## [1] 14793 98974

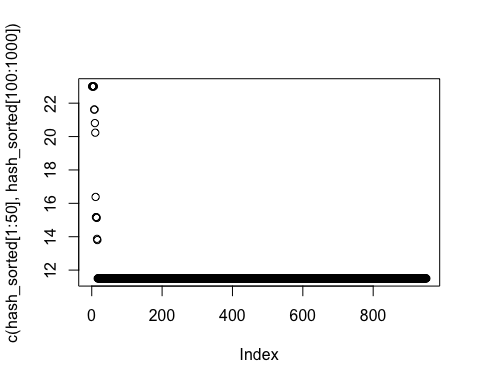
m<- as.matrix(hashtags.tdm)  
term\_nums <- rowSums(m)  
tweets\_sum <- rowSums(m!=0)  
  
  
#calculate tf of each hashtag  
tf <- term\_nums/tweets\_sum  
  
#calculate idf of each hashtag  
doc\_num <- nrow(tweets\_clean)  
idf <-log(doc\_num/tweets\_sum)  
  
#to get each hashtag tf-idf  
tf\_idf <- tf\*idf  
  
hash\_sorted <- sort(tf\_idf, decreasing = T)  
length(hash\_sorted)

## [1] 14793

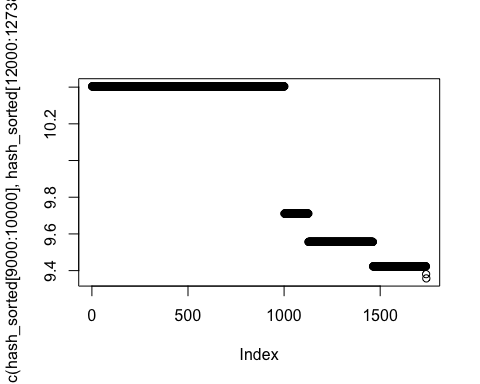
hash\_sorted[1:10]

## aeip blockfest cyberattack jews   
## 23.00522 23.00522 23.00522 23.00522   
## johnson oxforduniversity massivepotential mobileanalytics   
## 23.00522 23.00522 21.61893 21.61893   
## inac hds   
## 20.80800 20.23264

#plot the tf-idf score  
plot(c(hash\_sorted[1:50],hash\_sorted[100:1000]))



plot(c(hash\_sorted[9000:10000], hash\_sorted[12000:12738]))



quantile(hash\_sorted,probs=c(0.05,0.50,0.95))

## 5% 50% 95%   
## 8.206776 10.809465 11.502612

hashtag\_important\_list <- hash\_sorted[hash\_sorted>11.5]  
length(hashtag\_important\_list)

## [1] 5015

head(hashtag\_important\_list)

## aeip blockfest cyberattack jews   
## 23.00522 23.00522 23.00522 23.00522   
## johnson oxforduniversity   
## 23.00522 23.00522

Observing the quantile of the tf-idf score, only a small part of hashtags have score lower than ten, most of them have score larger then 12, Here I set 11.5 as the cut-off score to seperate the important hashtags and unimportant hashtags.

### Cluster the tweets using these hashtags .

Using the list of hashtags got from last question, Here, I choose about 500 hashtags and first 1000 tweets becasue of the computation problem which my computer can't calculate large numbers and it will take long time.

#use hierarchical clustering  
library("cluster")

## Warning: package 'cluster' was built under R version 3.2.5

hashtags.500.corpus <- Corpus(VectorSource(result\_hashtag))  
hashtags.500.tdm <- TermDocumentMatrix(hashtags.500.corpus, control=list(removePunctuation=T))  
tweets.matrix <- as.matrix(hashtags.500.tdm)  
  
#hclust clustering  
distMatrix <- dist(scale(t(tweets.matrix)[1:2000,1:500]))  
tweets.fit <- hclust(distMatrix,method = "ward.D")  
  
groups <- cutree(tweets.fit, k=5)  
table(groups)

## groups  
## 1 2 3 4 5   
## 1688 235 56 12 9

groups.1 <- groups[groups==1]  
groups.2 <- groups[groups==2]  
groups.3 <- groups[groups==3]  
groups.4 <- groups[groups==4]  
groups.5 <- groups[groups==5]  
  
  
head(groups.1)

## 1 2 3 4 5 6   
## 1 1 1 1 1 1

head(groups.2)

## 9 16 41 44 79 85   
## 2 2 2 2 2 2

head(groups.3)

## 28 54 161 163 179 218   
## 3 3 3 3 3 3

head(groups.4)

## 1945 1947 1950 1952 1955 1958   
## 4 4 4 4 4 4

head(groups.5)

## 1948 1949 1951 1953 1956 1959   
## 5 5 5 5 5 5

### Classify the tweets in ML.Tweets.New.csv using the cluster lables generated from ML.Tweets.csv.

#label the ML.Tweets.csv and use   
train\_tweets <- data.frame(tweets = tweets\_clean[1:2000,], cluster = groups)  
head(train\_tweets)

## tweets  
## 1 no davenport baseball downs viterbo to complete threegame sweep #grandrapids  
## 2 rt rt boatright hits them both fivepoint game #uconn #bleedblue  
## 3 rt yeh press start lets play why its not game over for #gamification   
## 4 beginning big data with power bi and excel #computerscience #microsoftexcel #bigdata #datamodeling  
## 5 rt factors limiting benefits of #bigdata #healthcare   
## 6 rt applying machine learning to medicine good stuff #drsherriworth  
## cluster  
## 1 1  
## 2 1  
## 3 1  
## 4 1  
## 5 1  
## 6 1

#preprocessing ML.Tweets.New.csv data  
  
#remove all messgaes not containing hashtags at all.  
new\_tag\_result <- apply(tweets.new, 1, has\_tag)  
tweets\_new\_ht <- tweets.new[which(new\_tag\_result),]  
  
#remove all non-english message  
tweets\_new\_clean <- apply(as.data.frame(tweets\_new\_ht), 1, clean\_sub)  
head(tweets\_new\_clean)

## [1] "witness the devastation of the #bombs great for #exploring fast #terraria #gamedev "   
## [2] "rt game day lets go spurs #gospursgo"   
## [3] "this game is nuckin futs wisconsin wants the seed and izzo wants to win the big #selectionsunday"   
## [4] "#packsb taking the field now were about to begin our second game of the day against fordham"   
## [5] "rt you train all summer practice all season prepare all year for the chance to play in an elimination game #day"  
## [6] "rt rt boatright hits them both fivepoint game #uconn #bleedblue"

#Classify the tweets\_new data  
matrix <- create\_matrix(train\_tweets[,1], language = "english", removeStopwords = T, stemWords = F,tm::weightTfIdf)  
  
mat <- as.matrix(matrix)  
classifier <- naiveBayes(mat, as.factor(train\_tweets[,2]))  
  
#to classify all the tweets in ML.Tweets.New.csv need a long time to run out the result, so I have to reduce the number of tweets for prediction.  
  
predicted <- predict(classifier,tweets\_new\_clean[1:50000])  
table(predicted)

## predicted  
## 1 2 3 4 5   
## 50000 0 0 0 0

Using naive Bayes method to classify the tweets in ML.Tweets.New.csv. Considering the computation ability of my own computer, I used 2000 tweets and their hashtags to do clustering, and use 50000 tweets to do prediction which lead to the bias in my final predict results which shows majority of tweets are classified in cluster 1.

### Use the qdap polarity function to score the polarity of the tweets in ML.Tweets.csv.

First, I will clean the data, remove users, hashtags, url, number, punctuations, then calculate the scores

#pre-poccessing the data  
clean\_sub <- function(x){  
 #remove all punctuations  
 gsub("[[:punct:]]", "",  
 #remove all control characters  
 gsub("[[:cntrl:]]", "",  
 #remove all digit  
 gsub("\\d+","",  
 #remove the hashtag  
 gsub("#[\_a-zA-Z1-9]+", "",   
 #remove the users  
 gsub("@([\_a-zA-Z1-9]+)","",  
 #remove the url  
 gsub("(https?://[^ ]+)","",x))))))  
   
}  
  
tweets\_clean <- apply(tweets, 1, clean\_sub)  
tweets\_clean <-as.character(tweets\_clean)  
tweets\_clean <- stripWhitespace(tweets\_clean)  
tweets\_clean <- removeWords(tweets\_clean, stopwords())  
head(tweets\_clean)

## [1] " Im game either"   
## [2] "Wow runs scored scored one inningif game continued Oregon demolished Oregon state"  
## [3] "What hell UCONN For second time game foul shot clock "   
## [4] "RT le much cocaine football game "   
## [5] "Wow BG Championship game unbelievable"   
## [6] " No Davenport baseball downs Viterbo complete threegame sweep "

When I used funtion polarity(), I met a computation problem, so I have to choose first 3000 tweets to do sentiment analysis

#Computation problem, there are more then 130000 tweets, it will take long time to get the result, so here I use 10000 instead of 130000 tweets.  
ps <- polarity(tweets\_clean[1:3000])  
dim(ps$all)

## [1] 3000 6

ps$all[1:50,]

## all wc polarity pos.words neg.words  
## 1 all 3 0.0000000 - -  
## 2 all 12 0.2886751 wow -  
## 3 all 10 -0.6324555 - hell, foul  
## 4 all 6 0.0000000 - -  
## 5 all 5 0.0000000 wow unbelievable  
## 6 all 8 0.0000000 - -  
## 7 all 6 0.0000000 - -  
## 8 all 11 -0.6030227 - destroy, hard  
## 9 all 11 0.3015113 awesome -  
## 10 all 9 0.6666667 love, support -  
## 11 all 5 0.4472136 unreal -  
## 12 all 6 0.0000000 like fear  
## 13 all 6 0.4082483 like -  
## 14 all 7 0.3779645 great -  
## 15 all 12 0.2886751 fun -  
## 16 all 8 0.0000000 - -  
## 17 all 8 0.0000000 - -  
## 18 all 4 0.5000000 superb -  
## 19 all 3 0.5773503 right -  
## 20 all 8 0.0000000 strong fuck  
## 21 all 12 0.0000000 - -  
## 22 all 6 0.4082483 excel -  
## 23 all 4 0.5000000 benefits -  
## 24 all 7 0.3779645 good -  
## 25 all 8 0.0000000 - -  
## 26 all 9 0.0000000 - -  
## 27 all 11 0.0000000 - -  
## 28 all 9 0.0000000 - -  
## 29 all 7 0.3779645 wonderful -  
## 30 all 7 0.0000000 - -  
## 31 all 9 0.0000000 - -  
## 32 all 6 0.0000000 - -  
## 33 all 6 0.0000000 - -  
## 34 all 0 NaN - -  
## 35 all 5 0.0000000 - -  
## 36 all 3 0.0000000 - -  
## 37 all 3 0.0000000 - -  
## 38 all 1 0.0000000 - -  
## 39 all 11 0.3015113 cool -  
## 40 all 11 0.3015113 cool -  
## 41 all 8 0.0000000 - -  
## 42 all 4 1.0000000 exceptional, talent -  
## 43 all 4 0.0000000 - -  
## 44 all 12 0.2886751 top -  
## 45 all 5 0.0000000 - -  
## 46 all 2 0.0000000 - -  
## 47 all 5 0.0000000 - -  
## 48 all 5 0.0000000 - -  
## 49 all 0 NaN - -  
## 50 all 8 0.0000000 - -  
## text.var  
## 1 Im game either  
## 2 Wow runs scored scored one inningif game continued Oregon demolished Oregon state  
## 3 What hell UCONN For second time game foul shot clock   
## 4 RT le much cocaine football game   
## 5 Wow BG Championship game unbelievable  
## 6 No Davenport baseball downs Viterbo complete threegame sweep   
## 7 RT RT Boatright hits Fivepoint game   
## 8 RT guy game shirt said skate destroy im laugjkng hard help  
## 9 awesome dont mind subscribing new youtube channel im game reviews indies   
## 10 RT I love I special person coming game support   
## 11 This BG title game unreal  
## 12 Everyday like game call fear factor  
## 13 Got truck full bricks like contractor  
## 14 I want quote now What great game  
## 15 RT We got guys making plays one Been fun game watch sure  
## 16 RT Yeh Press Start lets play Why game   
## 17 see yall hooping Them s thing keeping game  
## 18 Simply Addictive Superb Gameplay   
## 19 game right tournament  
## 20 RT My push people away game strong fuck  
## 21 Sony PlayStation GB Video Game System Jet Black CUHA Full read eBay   
## 22 Beginning Big Data Power BI Excel   
## 23 RT Factors Limiting Benefits   
## 24 RT Applying machine learning medicine Good stuff   
## 25 Big Data experts discuss preparing data deep dive   
## 26 RT Big Data experts discuss preparing data deep dive   
## 27 CB Insights Predicting Startup Unicorns And Mastering For Private Companies via   
## 28 RT jopemoro Puede IBM Watson hundir Spotify y Pandora   
## 29 RT kdnuggets Watch JeremyPHoward wonderful terrifying implications   
## 30 The student rate keeps rising can help   
## 31 RT Big Data experts discuss preparing data deep dive   
## 32 RT Strategies big data Whats YOURS   
## 33 revenue growth Check s infographic learn   
## 34   
## 35 RT The difference lot data   
## 36 RT The Daily   
## 37 Special gt Friday   
## 38 teehee  
## 39 FusionOps Can make supply chain data cool business execs via amp   
## 40 FusionOps Can make supply chain data cool business execs via amp   
## 41 ScikitLearn video tutorial Kyle Kastner SciPy Part Video   
## 42 Spotting Exceptional Talent via   
## 43 From heady days calling   
## 44 RT Top rMachineLearning Posts Apr Andrew Ng AMA Autoencoders Deep Learning Textbooks   
## 45 RT RelayRides Data Scientist via   
## 46 teehee Gemini   
## 47 RT Build wearable Foundation IBM   
## 48 Big Data Engineer Atlassian Sydney   
## 49   
## 50 RT Panelists share tips rehearsal Whos coming today

### Would creating a custom polarity.frame - A dataframe or environment containing a dataframe of positive/negative words and weights - based on the tags and words in these tweets improve the polarity score? Try it.

I will use a word list [AFINN](http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010) which has 2477 words and phrases rated from -5 [very negative] to +5 [very positive].

afinn\_list <- read.delim(file="AFINN-111.txt", header=F)  
names(afinn\_list) <- c("word", "score")  
afinn\_list$word <- tolower(afinn\_list$word)  
positive\_afinn <- c(afinn\_list$word[afinn\_list$score>0])  
negative\_afinn <- c(afinn\_list$word[afinn\_list$score<0])  
  
#Base on the previous question, I use 5000 tweets to do sentiment analysis  
tweets.3000.corpus <- Corpus(VectorSource(tweets\_clean[1:3000]))  
tweets.3000.tdm <- TermDocumentMatrix(tweets.3000.corpus, control= list(stopwords("en"), removeWords=T, stripWhitespace=T))  
tweets.3000.matrix <- as.matrix(tweets.3000.tdm)  
tweets.3000.freq <- rowSums(tweets.3000.matrix)  
  
#get the word list of 5000 tweets  
words.3000 <- names(tweets.3000.freq)  
  
#to find positive and negative words according to AFINN word list  
positive\_tweets <- as.vector(NULL)  
negative\_tweets <- as.vector(NULL)  
for (word in words.3000){  
 if (word %in% positive\_afinn){  
 positive\_tweets <- c(positive\_tweets,word)  
 }  
 else if (word %in% negative\_afinn){  
 negative\_tweets <- c(negative\_tweets,word)  
 }  
}  
  
positive\_tweets

## [1] "ability" "active" "adopt"   
## [4] "advanced" "advantage" "agree"   
## [7] "allow" "amazed" "amazing"   
## [10] "asset" "assets" "award"   
## [13] "awards" "awesome" "beautiful"   
## [16] "benefit" "benefits" "best"   
## [19] "better" "big" "bless"   
## [22] "bold" "boost" "breakthrough"   
## [25] "calm" "capable" "care"   
## [28] "cheers" "clear" "clearly"   
## [31] "committing" "competitive" "comprehensive"   
## [34] "confident" "congrats" "congratulations"  
## [37] "cool" "dear" "dream"   
## [40] "easy" "effective" "effectively"   
## [43] "embrace" "encourage" "engage"   
## [46] "enjoy" "ethical" "excellence"   
## [49] "excellent" "excited" "exciting"   
## [52] "expands" "exploration" "extend"   
## [55] "fame" "fantastic" "fascinating"   
## [58] "favorite" "favorites" "fine"   
## [61] "focused" "free" "fresh"   
## [64] "fulfill" "fun" "funny"   
## [67] "gain" "gained" "genial"   
## [70] "god" "good" "great"   
## [73] "greatest" "growing" "growth"   
## [76] "happiness" "happy" "help"   
## [79] "helping" "helps" "hero"   
## [82] "highlight" "hilarious" "hope"   
## [85] "huge" "importance" "important"   
## [88] "impress" "impressive" "improve"   
## [91] "improvement" "improving" "increase"   
## [94] "increased" "innovate" "innovation"   
## [97] "innovative" "inspirational" "inspired"   
## [100] "intelligent" "interested" "interesting"   
## [103] "join" "joke" "laugh"   
## [106] "like" "love" "matters"   
## [109] "natural" "nice" "opportunity"   
## [112] "outstanding" "passionate" "perfect"   
## [115] "please" "powerful" "pretty"   
## [118] "promise" "promises" "promotes"   
## [121] "promoting" "proud" "reach"   
## [124] "reached" "recommend" "responsible"   
## [127] "safe" "safely" "safety"   
## [130] "save" "secure" "sexy"   
## [133] "share" "shared" "shares"   
## [136] "significance" "slick" "smart"   
## [139] "smarter" "solid" "solution"   
## [142] "solutions" "solve" "solved"   
## [145] "solving" "spark" "sparkling"   
## [148] "stimulate" "straight" "strength"   
## [151] "strengthen" "strong" "success"   
## [154] "successful" "super" "superb"   
## [157] "support" "supported" "supports"   
## [160] "thank" "thanks" "thoughtful"   
## [163] "top" "true" "trust"   
## [166] "united" "useful" "vision"   
## [169] "want" "welcome" "win"   
## [172] "winner" "winning" "wish"   
## [175] "wonderful" "wow" "yes"

negative\_tweets

## [1] "abuse" "accidents" "alert" "alone"   
## [5] "ass" "attacks" "avoid" "bad"   
## [9] "badass" "battle" "bias" "blind"   
## [13] "block" "blocking" "broken" "cancer"   
## [17] "challenge" "chaos" "cheat" "conflict"   
## [21] "crash" "crazy" "crime" "crisis"   
## [25] "cuts" "cutting" "dead" "deception"   
## [29] "demands" "destroy" "difficult" "disasters"   
## [33] "disruption" "drop" "dumb" "error"   
## [37] "fail" "failed" "fails" "fear"   
## [41] "fight" "forget" "fraud" "fuck"   
## [45] "grave" "greedy" "hard" "hell"   
## [49] "hide" "hiding" "ill" "lack"   
## [53] "limited" "limits" "losing" "lost"   
## [57] "lurks" "messed" "misread" "miss"   
## [61] "missed" "mistakes" "mistaking" "obsolete"   
## [65] "offline" "overload" "overlooked" "pain"   
## [69] "poor" "poverty" "prevent" "problem"   
## [73] "problems" "risk" "risks" "screaming"   
## [77] "shock" "shocks" "shortage" "skeptics"   
## [81] "stop" "stupid" "terrible" "terrified"   
## [85] "threat" "threaten" "unbelievable" "unclear"   
## [89] "undermining" "useless" "violate" "war"   
## [93] "warning" "worse" "wrong" "wtf"

#generate a sentiment loopup hash table   
pol\_frame <- sentiment\_frame(positives = positive\_tweets, negative=negative\_tweets, pos.weights = 1, neg.weights = -1)  
  
  
ps.new <- polarity(tweets\_clean[1:3000], polarity.frame = pol\_frame)  
dim(ps.new$all)

## [1] 3000 6

ps.new$all[1:50,]

## all wc polarity pos.words neg.words  
## 1 all 3 0.0000000 - -  
## 2 all 12 0.2886751 wow -  
## 3 all 10 -0.3162278 - hell  
## 4 all 6 0.0000000 - -  
## 5 all 5 0.0000000 wow unbelievable  
## 6 all 8 0.0000000 - -  
## 7 all 6 0.0000000 - -  
## 8 all 11 -0.3015113 help destroy, hard  
## 9 all 11 0.3015113 awesome -  
## 10 all 9 0.6666667 love, support -  
## 11 all 5 0.0000000 - -  
## 12 all 6 0.0000000 like fear  
## 13 all 6 0.4082483 like -  
## 14 all 7 0.7559289 want, great -  
## 15 all 12 0.2886751 fun -  
## 16 all 8 0.0000000 - -  
## 17 all 8 0.0000000 - -  
## 18 all 4 0.5000000 superb -  
## 19 all 3 0.0000000 - -  
## 20 all 8 0.0000000 strong fuck  
## 21 all 12 0.0000000 - -  
## 22 all 6 0.4082483 big -  
## 23 all 4 0.5000000 benefits -  
## 24 all 7 0.3779645 good -  
## 25 all 8 0.3535534 big -  
## 26 all 9 0.3333333 big -  
## 27 all 11 0.0000000 - -  
## 28 all 9 0.0000000 - -  
## 29 all 7 0.3779645 wonderful -  
## 30 all 7 0.3779645 help -  
## 31 all 9 0.3333333 big -  
## 32 all 6 0.4082483 big -  
## 33 all 6 0.4082483 growth -  
## 34 all 0 NaN - -  
## 35 all 5 0.0000000 - -  
## 36 all 3 0.0000000 - -  
## 37 all 3 0.0000000 - -  
## 38 all 1 0.0000000 - -  
## 39 all 11 0.3015113 cool -  
## 40 all 11 0.3015113 cool -  
## 41 all 8 0.0000000 - -  
## 42 all 4 0.0000000 - -  
## 43 all 4 0.0000000 - -  
## 44 all 12 0.2886751 top -  
## 45 all 5 0.0000000 - -  
## 46 all 2 0.0000000 - -  
## 47 all 5 0.0000000 - -  
## 48 all 5 0.4472136 big -  
## 49 all 0 NaN - -  
## 50 all 8 0.3535534 share -  
## text.var  
## 1 Im game either  
## 2 Wow runs scored scored one inningif game continued Oregon demolished Oregon state  
## 3 What hell UCONN For second time game foul shot clock   
## 4 RT le much cocaine football game   
## 5 Wow BG Championship game unbelievable  
## 6 No Davenport baseball downs Viterbo complete threegame sweep   
## 7 RT RT Boatright hits Fivepoint game   
## 8 RT guy game shirt said skate destroy im laugjkng hard help  
## 9 awesome dont mind subscribing new youtube channel im game reviews indies   
## 10 RT I love I special person coming game support   
## 11 This BG title game unreal  
## 12 Everyday like game call fear factor  
## 13 Got truck full bricks like contractor  
## 14 I want quote now What great game  
## 15 RT We got guys making plays one Been fun game watch sure  
## 16 RT Yeh Press Start lets play Why game   
## 17 see yall hooping Them s thing keeping game  
## 18 Simply Addictive Superb Gameplay   
## 19 game right tournament  
## 20 RT My push people away game strong fuck  
## 21 Sony PlayStation GB Video Game System Jet Black CUHA Full read eBay   
## 22 Beginning Big Data Power BI Excel   
## 23 RT Factors Limiting Benefits   
## 24 RT Applying machine learning medicine Good stuff   
## 25 Big Data experts discuss preparing data deep dive   
## 26 RT Big Data experts discuss preparing data deep dive   
## 27 CB Insights Predicting Startup Unicorns And Mastering For Private Companies via   
## 28 RT jopemoro Puede IBM Watson hundir Spotify y Pandora   
## 29 RT kdnuggets Watch JeremyPHoward wonderful terrifying implications   
## 30 The student rate keeps rising can help   
## 31 RT Big Data experts discuss preparing data deep dive   
## 32 RT Strategies big data Whats YOURS   
## 33 revenue growth Check s infographic learn   
## 34   
## 35 RT The difference lot data   
## 36 RT The Daily   
## 37 Special gt Friday   
## 38 teehee  
## 39 FusionOps Can make supply chain data cool business execs via amp   
## 40 FusionOps Can make supply chain data cool business execs via amp   
## 41 ScikitLearn video tutorial Kyle Kastner SciPy Part Video   
## 42 Spotting Exceptional Talent via   
## 43 From heady days calling   
## 44 RT Top rMachineLearning Posts Apr Andrew Ng AMA Autoencoders Deep Learning Textbooks   
## 45 RT RelayRides Data Scientist via   
## 46 teehee Gemini   
## 47 RT Build wearable Foundation IBM   
## 48 Big Data Engineer Atlassian Sydney   
## 49   
## 50 RT Panelists share tips rehearsal Whos coming today

#using default polarity frame  
ps$all[1:50,]

## all wc polarity pos.words neg.words  
## 1 all 3 0.0000000 - -  
## 2 all 12 0.2886751 wow -  
## 3 all 10 -0.6324555 - hell, foul  
## 4 all 6 0.0000000 - -  
## 5 all 5 0.0000000 wow unbelievable  
## 6 all 8 0.0000000 - -  
## 7 all 6 0.0000000 - -  
## 8 all 11 -0.6030227 - destroy, hard  
## 9 all 11 0.3015113 awesome -  
## 10 all 9 0.6666667 love, support -  
## 11 all 5 0.4472136 unreal -  
## 12 all 6 0.0000000 like fear  
## 13 all 6 0.4082483 like -  
## 14 all 7 0.3779645 great -  
## 15 all 12 0.2886751 fun -  
## 16 all 8 0.0000000 - -  
## 17 all 8 0.0000000 - -  
## 18 all 4 0.5000000 superb -  
## 19 all 3 0.5773503 right -  
## 20 all 8 0.0000000 strong fuck  
## 21 all 12 0.0000000 - -  
## 22 all 6 0.4082483 excel -  
## 23 all 4 0.5000000 benefits -  
## 24 all 7 0.3779645 good -  
## 25 all 8 0.0000000 - -  
## 26 all 9 0.0000000 - -  
## 27 all 11 0.0000000 - -  
## 28 all 9 0.0000000 - -  
## 29 all 7 0.3779645 wonderful -  
## 30 all 7 0.0000000 - -  
## 31 all 9 0.0000000 - -  
## 32 all 6 0.0000000 - -  
## 33 all 6 0.0000000 - -  
## 34 all 0 NaN - -  
## 35 all 5 0.0000000 - -  
## 36 all 3 0.0000000 - -  
## 37 all 3 0.0000000 - -  
## 38 all 1 0.0000000 - -  
## 39 all 11 0.3015113 cool -  
## 40 all 11 0.3015113 cool -  
## 41 all 8 0.0000000 - -  
## 42 all 4 1.0000000 exceptional, talent -  
## 43 all 4 0.0000000 - -  
## 44 all 12 0.2886751 top -  
## 45 all 5 0.0000000 - -  
## 46 all 2 0.0000000 - -  
## 47 all 5 0.0000000 - -  
## 48 all 5 0.0000000 - -  
## 49 all 0 NaN - -  
## 50 all 8 0.0000000 - -  
## text.var  
## 1 Im game either  
## 2 Wow runs scored scored one inningif game continued Oregon demolished Oregon state  
## 3 What hell UCONN For second time game foul shot clock   
## 4 RT le much cocaine football game   
## 5 Wow BG Championship game unbelievable  
## 6 No Davenport baseball downs Viterbo complete threegame sweep   
## 7 RT RT Boatright hits Fivepoint game   
## 8 RT guy game shirt said skate destroy im laugjkng hard help  
## 9 awesome dont mind subscribing new youtube channel im game reviews indies   
## 10 RT I love I special person coming game support   
## 11 This BG title game unreal  
## 12 Everyday like game call fear factor  
## 13 Got truck full bricks like contractor  
## 14 I want quote now What great game  
## 15 RT We got guys making plays one Been fun game watch sure  
## 16 RT Yeh Press Start lets play Why game   
## 17 see yall hooping Them s thing keeping game  
## 18 Simply Addictive Superb Gameplay   
## 19 game right tournament  
## 20 RT My push people away game strong fuck  
## 21 Sony PlayStation GB Video Game System Jet Black CUHA Full read eBay   
## 22 Beginning Big Data Power BI Excel   
## 23 RT Factors Limiting Benefits   
## 24 RT Applying machine learning medicine Good stuff   
## 25 Big Data experts discuss preparing data deep dive   
## 26 RT Big Data experts discuss preparing data deep dive   
## 27 CB Insights Predicting Startup Unicorns And Mastering For Private Companies via   
## 28 RT jopemoro Puede IBM Watson hundir Spotify y Pandora   
## 29 RT kdnuggets Watch JeremyPHoward wonderful terrifying implications   
## 30 The student rate keeps rising can help   
## 31 RT Big Data experts discuss preparing data deep dive   
## 32 RT Strategies big data Whats YOURS   
## 33 revenue growth Check s infographic learn   
## 34   
## 35 RT The difference lot data   
## 36 RT The Daily   
## 37 Special gt Friday   
## 38 teehee  
## 39 FusionOps Can make supply chain data cool business execs via amp   
## 40 FusionOps Can make supply chain data cool business execs via amp   
## 41 ScikitLearn video tutorial Kyle Kastner SciPy Part Video   
## 42 Spotting Exceptional Talent via   
## 43 From heady days calling   
## 44 RT Top rMachineLearning Posts Apr Andrew Ng AMA Autoencoders Deep Learning Textbooks   
## 45 RT RelayRides Data Scientist via   
## 46 teehee Gemini   
## 47 RT Build wearable Foundation IBM   
## 48 Big Data Engineer Atlassian Sydney   
## 49   
## 50 RT Panelists share tips rehearsal Whos coming today

In fact, using the custom polarity frame improve the polarity score. For example, the fourteenth tweets, the old score is 0.37796, and the new score is 0.7559289. Using the new polarity frame, it is more sensitive for positive tweets.