## Loading libraries

library(tm) # Framework for text mining.

library(SnowballC) # Provides wordStem() for stemming.

library(qdap)# Quantitative discourse analysis of transcripts.

library(qdapDictionaries)

library(dplyr) # Data preparation and pipes %>%.

library(RColorBrewer) # Generate palette of colours for plots.library(ggplot2) # Plot word frequencies.

library(scales) # Include commas in numbers.

library(Rgraphviz) # Correlation plots.

library(RWeka)

##Loading corpus and files

cname <- file.path("E:/Dropbox/Courses/DATA SCIENCE TRACK/10. Capstone Project", "Dataset", "sampled")

length(dir(cname))

dir(cname)

docs <- Corpus(DirSource(cname))

docs

class(docs)

summary(docs)

## Transformations

toSpace <- function(x, pattern) gsub(pattern, " ", x) ##Replace char with space

docs <- tm\_map(docs, toSpace, "/")

docs <- tm\_map(docs, toSpace, "@")

docs <- tm\_map(docs, toSpace, "\\|")

docs <- tm\_map(docs, toSpace, ":")

docs <- tm\_map(docs, toSpace, ";")

docs <- tm\_map(docs, toSpace, "\\(")

docs <- tm\_map(docs, toSpace, "\\)")

docs <- tm\_map(docs, toSpace, "\\[")

docs <- tm\_map(docs, toSpace, "\\]")

docs <- tm\_map(docs, toSpace, "\\{")

docs <- tm\_map(docs, toSpace, "\\}")

docs <- tm\_map(docs, toSpace, "-")

toString <- function(x, from, to) gsub(from, to, x) ## specific transformations

docs <- tm\_map(docs, toString, ",", "")

docs <- tm\_map(docs, toString, "\\.", "")

docs <- tm\_map(docs, toString, "’", "'")

docs <- tm\_map(docs, toString, "â€™", "'")

docs <- tm\_map(docs, toString, "â€˜", "'")

docs <- tm\_map(docs, toString, "™", "")

docs <- tm\_map(docs, toString, " ™", "")

tonothing <- function(x, pattern) gsub(pattern, "", x) ##Replace char with nothing

docs <- tm\_map(docs, tonothing, "â€œ")

docs <- tm\_map(docs, tonothing, "â€")

docs <- tm\_map(docs, tonothing, "!")

docs <- tm\_map(docs, tonothing, "?")

docs <- tm\_map(docs, tonothing, " [0-9] ")

docs <- tm\_map(docs, tonothing, "[0-9][0-9] [0-9][0-9]")

docs <- tm\_map(docs, tonothing, "\\$")

docs <- tm\_map(docs, tonothing, "¦")

docs <- tm\_map(docs, tonothing, "’ ")

docs <- tm\_map(docs, tonothing, "’\n")

docs <- tm\_map(docs, tonothing, "“")

docs <- tm\_map(docs, tonothing, "”")

docs <- tm\_map(docs, tonothing, "‘")

docs <- tm\_map(docs, tolower) ## convert to lower case

##docs <- tm\_map(docs, removePunctuation) ## remove punctuation

docs <- tm\_map(docs, removeWords, stopwords("english")) ## remove common stop words

##docs <- tm\_map(docs, removeNumbers) ## remove numbers

setwd("E:/Dropbox/Courses/DATA SCIENCE TRACK/10. Capstone Project")

con = file("profanity.txt", "rb")

profanity=readLines(con)

close(con)

docs <- tm\_map(docs, removeWords, profanity) ## remove bad words

docs <- tm\_map(docs, stripWhitespace) ## remove whitespace

## writing the docs to txt file

setwd("E:/Dropbox/Courses/DATA SCIENCE TRACK/10. Capstone Project/Dataset/Processed file with stop words command")

lapply(docs[[1]], write, "blog.txt", append=TRUE)

lapply(docs[[2]], write, "news.txt", append=TRUE)

lapply(docs[[3]], write, "twitter.txt", append=TRUE)