**Step 1: Below all packages are installed, which are needed for the text mining.**

#initializing all the packages needed for the text mining

install.packages("rvest", dependencies = TRUE)

install.packages("XML", dependencies = TRUE)

install.packages("dplyr", dependencies = TRUE)

install.packages("ggplot2", dependencies = TRUE)

install.packages("tm", dependencies = TRUE)

install.packages("SnowballC", dependencies = TRUE)

install.packages("magrittr", dependencies = TRUE)

install.packages("wordcloud", dependencies = TRUE)

install.packages("graphics", dependencies = TRUE)

install.packages("grDevices", dependencies = TRUE)

**Step 2: Path is set and all packages are loaded in R.**

setwd("C:/Users/Spéro/Documents/ITM/Data Analytics/Week 15/Exercise 15/Apple filings")

library(rvest)

library(XML)

library(dplyr)

library(ggplot2)

library(tm)

library(SnowballC)

library(magrittr)

library(wordcloud)

library(graphics)

library(grDevices)

**Step 3: Read lines in every filing downloaded (Executed below code for every year).**

**2008 year doesn’t have Q&A section; hence we have done from 2009 to 2016.**

apple\_proxy\_2009 <- readLines("C:/Users/Spéro/Documents/ITM/Data Analytics/Week 15/Exercise 15/Apple filings/apple\_2009.htm")

apple\_proxy\_2010 <- readLines("C:/Users/Spéro/Documents/ITM/Data Analytics/Week 15/Exercise 15/Apple filings/apple\_2010.htm")

apple\_proxy\_2011 <- readLines("C:/Users/Spéro/Documents/ITM/Data Analytics/Week 15/Exercise 15/Apple filings/apple\_2011.htm")

apple\_proxy\_2012 <- readLines("C:/Users/Spéro/Documents/ITM/Data Analytics/Week 15/Exercise 15/Apple filings/apple\_2012.htm")

apple\_proxy\_2013 <- readLines("C:/Users/Spéro/Documents/ITM/Data Analytics/Week 15/Exercise 15/Apple filings/apple\_2013.htm")

apple\_proxy\_2014 <- readLines("C:/Users/Spéro/Documents/ITM/Data Analytics/Week 15/Exercise 15/Apple filings/apple\_2014.htm")

apple\_proxy\_2015 <- readLines("C:/Users/Spéro/Documents/ITM/Data Analytics/Week 15/Exercise 15/Apple filings/apple\_2015.htm")

apple\_proxy\_2016 <- readLines("C:/Users/Spéro/Documents/ITM/Data Analytics/Week 15/Exercise 15/Apple filings/apple\_2016.htm")

cname <- file.path("C:/Users/Spéro/Documents/ITM/Data Analytics/Week 15/Exercise 15/Apple filings", "proxy\_sections")

cname

dir(cname)

**Step 3: Q&A session is defined (Executed below code for every year).**

**2009:**

qa\_sentence\_start\_2009 <- "Why am I receiving these materials?"

grep(qa\_sentence\_start\_2009, apple\_proxy\_2009, ignore.case = TRUE)

qa\_sentence\_end\_2009 <- "under the Exchange Act."

grep(qa\_sentence\_end\_2009, apple\_proxy\_2009, ignore.case = TRUE)

proxy\_2009\_section <- apple\_proxy\_2009[268:872]

**2010:**

qa\_sentence\_start\_2010 <- "Why am I receiving these materials?"

grep(qa\_sentence\_start\_2010, apple\_proxy\_2010, ignore.case = TRUE)

qa\_sentence\_end\_2010 <- "with notice on or prior to November 28, 2010."

grep(qa\_sentence\_end\_2010, apple\_proxy\_2010, ignore.case = TRUE)

proxy\_2010\_section <- apple\_proxy\_2010[242:643]

**2011:**

qa\_sentence\_start\_2011 <- "Why am I receiving these materials?"

grep(qa\_sentence\_start\_2011, apple\_proxy\_2011, ignore.case = TRUE)

qa\_sentence\_end\_2011<- " proposal that you intend to present at the 2012 annual meeting of shareholders"

grep(qa\_sentence\_end\_2011, apple\_proxy\_2011, ignore.case = TRUE)

proxy\_2011\_section <- apple\_proxy\_2011[208:590]

**2012:**

qa\_sentence\_start\_2012 <- "Why am I receiving these materials?"

grep(qa\_sentence\_start\_2012, apple\_proxy\_2012, ignore.case = TRUE)

qa\_sentence\_end\_2012<- " proposal that you intend to present at the 2013 annual meeting of shareholders."

grep(qa\_sentence\_end\_2012, apple\_proxy\_2012, ignore.case = TRUE)

proxy\_2012\_section <- apple\_proxy\_2012[206:600]

**2013:**

qa\_sentence\_start\_2013 <- "Why am I receiving these materials?"

grep(qa\_sentence\_start\_2013, apple\_proxy\_2013, ignore.case = TRUE)

qa\_sentence\_end\_2013 <- "the shareholder intends to present at the 2014 annual meeting of shareholders."

grep(qa\_sentence\_end\_2013, apple\_proxy\_2013, ignore.case = TRUE)

proxy\_2013\_section <- apple\_proxy\_2013[220:624]

**2014:**

qa\_sentence\_start\_2014 <- "Why am I receiving these materials?"

grep(qa\_sentence\_start\_2014, apple\_proxy\_2014, ignore.case = TRUE)

qa\_sentence\_end\_2014 <- "how to register your shares directly in your name as a shareholder of record."

grep(qa\_sentence\_end\_2014, apple\_proxy\_2014, ignore.case = TRUE)

proxy\_2014\_section <- apple\_proxy\_2014[240:739]

**2015:**

qa\_sentence\_start\_2015 <- "Why am I receiving these materials?"

grep(qa\_sentence\_start\_2015, apple\_proxy\_2015, ignore.case = TRUE)

qa\_sentence\_end\_2015 <- "a shareholder of record"

grep(qa\_sentence\_end\_2015, apple\_proxy\_2015, ignore.case = TRUE)

proxy\_2015\_section <- apple\_proxy\_2015[681:1132]

**2016:**

qa\_sentence\_start\_2016 <- "Why am I receiving these materials?"

grep(qa\_sentence\_start\_2016, apple\_proxy\_2016, ignore.case = TRUE)

qa\_sentence\_end\_2016 <- "Proxy Statement to ensure that your vote will be represented at the Annual Meeting."

grep(qa\_sentence\_end\_2016, apple\_proxy\_2016, ignore.case = TRUE)

proxy\_2016\_section <- apple\_proxy\_2016[1265:1704]

**Step 4: Texts files are created to aggregate in Corpus (Did for every year).**

#create .txt files to aggregate in Corpus

cat(proxy\_2009\_section,file="C:/Users/Spéro/Documents/ITM/Data Analytics/Week 15/Exercise 15/Apple filings/proxy\_sections/proxy\_2009\_section.htm", sep="n",append = TRUE)

cat(proxy\_2010\_section,file="C:/Users/Spéro/Documents/ITM/Data Analytics/Week 15/Exercise 15/Apple filings/proxy\_sections/proxy\_2010\_section.htm", sep="n",append = TRUE)

cat(proxy\_2011\_section,file="C:/Users/Spéro/Documents/ITM/Data Analytics/Week 15/Exercise 15/Apple filings/proxy\_sections/proxy\_2011\_section.htm", sep="n",append = TRUE)

cat(proxy\_2012\_section,file="C:/Users/Spéro/Documents/ITM/Data Analytics/Week 15/Exercise 15/Apple filings/proxy\_sections/proxy\_2012\_section.htm", sep="n",append = TRUE)

cat(proxy\_2013\_section,file="C:/Users/Spéro/Documents/ITM/Data Analytics/Week 15/Exercise 15/Apple filings/proxy\_sections/proxy\_2013\_section.htm", sep="n",append = TRUE)

cat(proxy\_2014\_section,file="C:/Users/Spéro/Documents/ITM/Data Analytics/Week 15/Exercise 15/Apple filings/proxy\_sections/proxy\_2014\_section.htm", sep="n",append = TRUE)

cat(proxy\_2015\_section,file="C:/Users/Spéro/Documents/ITM/Data Analytics/Week 15/Exercise 15/Apple filings/proxy\_sections/proxy\_2015\_section.htm", sep="n",append = TRUE)

cat(proxy\_2016\_section,file="C:/Users/Spéro/Documents/ITM/Data Analytics/Week 15/Exercise 15/Apple filings/proxy\_sections/proxy\_2016\_section.htm", sep="n",append = TRUE)

preprocess <- lapply(dir(cname), function(x) read\_html(x))

preprocess <- preprocess %>% lapply(function (x) html\_text(x))

docs <- as.VCorpus(preprocess)

inspect(docs[2])

for(j in seq(docs))

{

docs[[j]] <- gsub("\n", " ", docs[[j]])

docs[[j]] <- gsub("Â", " ", docs[[j]])

docs[[j]] <- gsub("nÂ", " ", docs[[j]])

}

# Check to see if it worked.

inspect(docs[3])

**Step 5: Preprocessing is done on texts, once it is loaded.**

docs <- tm\_map(docs, removePunctuation)

docs <- tm\_map(docs, removeNumbers)

docs <- tm\_map(docs, tolower)

# For a list of the stopwords, see:

# stopwords("english")

# length(stopwords("english"))

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

docs <- tm\_map(docs, stripWhitespace)

docs <- tm\_map(docs, stemDocument)

docs <-tm\_map(docs, PlainTextDocument)

docs <- tm\_map(docs, removeWords, c("trim","size","font", "can", "also", "mail", "familytimes", "with", "telephone", "if", "may", "help", "us", "will", "please", "unless", "visit", "nnn", "address", "new", "em"))

**Step 5: Document Term Matrix is created.**

dtm\_apple <- DocumentTermMatrix(docs)

dtm\_apple

**Step 5: Transpose of Matrix is done.**

tdm\_apple <- TermDocumentMatrix(docs)

tdm\_apple

**Step 5: Spare terms are removed and data visualization displayed.**

# Start by removing sparse terms:

dtms\_apple <- removeSparseTerms(dtm\_apple, 0.1) # This makes a matrix that is 10% empty space, maximum.

inspect(dtms\_apple)

freq\_dtm <- colSums(as.matrix(dtm\_apple))

length(freq\_dtm)

ord <- order(freq\_dtm)

freq\_dtm[head(ord)]

freq\_dtm[tail(ord)]

head(table(freq\_dtm), 20)

tail(table(freq\_dtm), 20)

freq\_dtm <- colSums(as.matrix(dtms\_apple))

freq\_dtm

freq\_dtm <- sort(colSums(as.matrix(dtm\_apple)), decreasing=TRUE)

head(freq\_dtm, 14)

findFreqTerms(dtm\_apple, lowfreq=50)

#find correlations

findAssocs(dtm\_apple, "apple", corlimit = .9)

#Plotting wordcloud

set.seed(123)

wordcloud(names(freq\_dtm), freq\_dtm, min.freq\_dtm=50)

#if you want to draw a heatmap

x <- as.matrix(dtm\_apple)

rc <- rainbow(nrow(x), start = 0, end = .3)

cc <- rainbow(ncol(x), start = 0, end = .3)

hv <- heatmap(x, col = cm.colors(256), scale = "column",

RowSideColors = rc, ColSideColors = cc, margins = c(5,10),

xlab = "Word frequency", ylab = "Word",

main = "heatmap(<dtm\_apple data>, ..., scale = \"column\")")

**Step 6: CSV file of Document Term Matrix is created.**

matrix\_apple<-as.matrix(dtm\_apple)

dim(matrix\_apple)

write.csv(matrix\_apple,file= "C:/Users/Spéro/Documents/ITM/Data Analytics/Week 15/Exercise 15/Apple filings/apple\_proxy\_text\_mining.csv")

**Step 6: CSV file of Transpose of Matrix of is created.**

matrix\_transposed\_apple<-as.matrix(tdm\_apple)

dim(matrix\_transposed\_apple)

write.csv(matrix\_transposed\_apple,file= "C:/Users/Spéro/Documents/ITM/Data Analytics/Week 15/Exercise 15/Apple filings/apple\_proxy\_text\_mining\_transposed.csv")