Hp\_proxy2016 <- readLines("C:/Users/gururaj/Desktop/Hewlett Packard/Hp\_proxy2016.html")

Hp\_proxy2015 <- readLines("C:/Users/gururaj/Desktop/Hewlett Packard/Hp\_proxy2015.html")

Hp\_proxy2014 <- readLines("C:/Users/gururaj/Desktop/Hewlett Packard/Hp\_proxy2014.html")

Hp\_proxy2013 <- readLines("C:/Users/gururaj/Desktop/Hewlett Packard/Hp\_proxy2013.html")

Hp\_proxy2012 <- readLines("C:/Users/gururaj/Desktop/Hewlett Packard/Hp\_proxy2012.html")

Hp\_proxy2011 <- readLines("C:/Users/gururaj/Desktop/Hewlett Packard/Hp\_proxy2011.html")

Hp\_proxy2010 <- readLines("C:/Users/gururaj/Desktop/Hewlett Packard/Hp\_proxy2010.html")

Hp\_proxy2009 <- readLines("C:/Users/gururaj/Desktop/Hewlett Packard/Hp\_proxy2009.html")

Hp\_proxy2008 <- readLines("C:/Users/gururaj/Desktop/Hewlett Packard/Hp\_proxy2008.html")

Hp\_proxy2007 <- readLines("C:/Users/gururaj/Desktop/Hewlett Packard/Hp\_proxy2007.html")

Hp\_proxy2006 <- readLines("C:/Users/gururaj/Desktop/Hewlett Packard/Hp\_proxy2006.html")

qa\_sentence\_start <- "QUESTIONS AND ANSWERS"

grep(qa\_sentence\_start, Hp\_proxy2016, ignore.case = TRUE)

qa\_sentence\_end <- "you should contact our proxy solicitor"

grep(qa\_sentence\_end, Hp\_proxy2016, ignore.case = TRUE)

qa\_section2016 <- Hp\_proxy2016[13208:13939]

cat(qa\_section2016,

file="C:/Users/gururaj/Desktop/Hewlett Packard/qa\_section/Hp\_proxyqa\_section2016.txt", sep="n",

append = TRUE)

qa\_sentence\_start <- "QUESTIONS AND ANSWERS"

grep(qa\_sentence\_start, Hp\_proxy2015, ignore.case = TRUE)

qa\_sentence\_end <- "you should contact our proxy solicitor"

grep(qa\_sentence\_end, Hp\_proxy2015, ignore.case = TRUE)

qa\_section2016 <- Hp\_proxy2015[1153:1780]

cat(qa\_section2015,

file="C:/Users/gururaj/Desktop/Hewlett Packard/qa\_section/Hp\_proxyqa\_section2015.txt", sep="n",

append = TRUE)

qa\_sentence\_start <- "QUESTIONS AND ANSWERS"

grep(qa\_sentence\_start, Hp\_proxy2014, ignore.case = TRUE)

qa\_sentence\_end <- "you should contact our proxy solicitor"

grep(qa\_sentence\_end, Hp\_proxy2014, ignore.case = TRUE)

qa\_section2014 <- Hp\_proxy2014[1763:2385]

cat(qa\_section2014,

file="C:/Users/gururaj/Desktop/Hewlett Packard/qa\_section/Hp\_proxyqa\_section2014.txt", sep="n",

append = TRUE)

qa\_sentence\_start <- "QUESTIONS AND ANSWERS"

grep(qa\_sentence\_start, Hp\_proxy2013, ignore.case = TRUE)

qa\_sentence\_end <- "you should contact HP's proxy solicitor"

grep(qa\_sentence\_end, Hp\_proxy2013, ignore.case = TRUE)

qa\_section2013 <- Hp\_proxy2013[1731:2307]

cat(qa\_section2013,

file="C:/Users/gururaj/Desktop/Hewlett Packard/qa\_section/Hp\_proxyqa\_section2013.txt", sep="n",

append = TRUE)

qa\_sentence\_start <- "QUESTIONS AND ANSWERS"

grep(qa\_sentence\_start, Hp\_proxy2012, ignore.case = TRUE)

qa\_sentence\_end <- "you should contact HP's proxy solicitor"

grep(qa\_sentence\_end, Hp\_proxy2012, ignore.case = TRUE)

qa\_section2013 <- Hp\_proxy2012[1102:1724]

cat(qa\_section2012,

file="C:/Users/gururaj/Desktop/Hewlett Packard/qa\_section/Hp\_proxyqa\_section2012.txt", sep="n",

append = TRUE)

qa\_sentence\_start <- "QUESTIONS AND ANSWERS"

grep(qa\_sentence\_start, Hp\_proxy2011, ignore.case = TRUE)

qa\_sentence\_end <- "you should contact HP's proxy solicitor"

grep(qa\_sentence\_end, Hp\_proxy2011, ignore.case = TRUE)

qa\_section2011 <- Hp\_proxy2011[1046:1631]

cat(qa\_section2011,

file="C:/Users/gururaj/Desktop/Hewlett Packard/qa\_section/Hp\_proxyqa\_section2011.txt", sep="n",

append = TRUE)

qa\_sentence\_start <- "QUESTIONS AND ANSWERS"

grep(qa\_sentence\_start, Hp\_proxy2010, ignore.case = TRUE)

qa\_sentence\_end <- "you should contact HP's proxy solicitor"

grep(qa\_sentence\_end, Hp\_proxy2010, ignore.case = TRUE)

qa\_section2010 <- Hp\_proxy2010[897:1575]

cat(qa\_section2010,

file="C:/Users/gururaj/Desktop/Hewlett Packard/qa\_section/Hp\_proxyqa\_section2010.txt", sep="n",

append = TRUE)

qa\_sentence\_start <- "QUESTIONS AND ANSWERS"

grep(qa\_sentence\_start, Hp\_proxy2009, ignore.case = TRUE)

qa\_sentence\_end <- "you should contact HP's proxy solicitor"

grep(qa\_sentence\_end, Hp\_proxy2009, ignore.case = TRUE)

qa\_section2009 <- Hp\_proxy2009[677:1224]

cat(qa\_section2009,

file="C:/Users/gururaj/Desktop/Hewlett Packard/qa\_section/Hp\_proxyqa\_section2009.txt", sep="n",

append = TRUE)

qa\_sentence\_start <- "QUESTIONS AND ANSWERS"

grep(qa\_sentence\_start, Hp\_proxy2008, ignore.case = TRUE)

qa\_sentence\_end <- "you should contact HP's proxy solicitor"

grep(qa\_sentence\_end, Hp\_proxy2008, ignore.case = TRUE)

qa\_section2008 <- Hp\_proxy2008[665:1169]

cat(qa\_section2008,

file="C:/Users/gururaj/Desktop/Hewlett Packard/qa\_section/Hp\_proxyqa\_section2008.txt", sep="n",

append = TRUE)

qa\_sentence\_start <- "QUESTIONS AND ANSWERS"

grep(qa\_sentence\_start, Hp\_proxy2007, ignore.case = TRUE)

qa\_sentence\_end <- "you should contact HP's proxy solicitor"

grep(qa\_sentence\_end, Hp\_proxy2007, ignore.case = TRUE)

qa\_section2007 <- Hp\_proxy2007[704:1174]

cat(qa\_section2007,

file="C:/Users/gururaj/Desktop/Hewlett Packard/qa\_section/Hp\_proxyqa\_section2007.txt", sep="n",

append = TRUE)

qa\_sentence\_start <- "QUESTIONS AND ANSWERS"

grep(qa\_sentence\_start, Hp\_proxy2006, ignore.case = TRUE)

qa\_sentence\_end <- "you should contact HP's proxy solicitor"

grep(qa\_sentence\_end, Hp\_proxy2006, ignore.case = TRUE)

qa\_section2006 <- Hp\_proxy2006[833:1050]

cat(qa\_section2006,

file="C:/Users/gururaj/Desktop/Hewlett Packard/qa\_section/Hp\_proxyqa\_section2006.txt", sep="n",

append = TRUE)

cname <- "C:/Users/gururaj/Desktop/Hewlett Packard/qa\_section/"

qa\_sections <- Corpus(DirSource(cname))

dir(cname)

**#To lower case:**

qa\_sections <- tm\_map(qa\_sections, content\_transformer(tolower))

**#Remove Numbers:**

qa\_sections <- tm\_map(qa\_sections, removeNumbers)

**#Remove Stop Words:**

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

viewDocs(qa\_sections, 1)

**# Replacing certin expressions with spaces:**

toSpace <- content\_transformer(function(x, pattern) gsub(pattern, " ",x))

qa\_sections <- tm\_map(qa\_sections, toSpace, "/|<|>|"|=|@|\\|:|;|-|\"")

**#to remove punctuations**

qa\_sections <- tm\_map(qa\_sections, removePunctuation)

**#Strip white spaces:**

qa\_sections <- tm\_map(qa\_sections, stripWhitespace)

#Remove known often words:

qa\_sections <- tm\_map(qa\_sections, removeWords, c("qnbsp","zeq","chk","efw","","ul","dt","dd","nn","b","br","pt", "q", "a", "i", "e", "font",

"style","n","trim","size","font", "can", "also", "e", "mail", "via", "td", "align","border",

"familytimes", "roman", "p", "tr", "nbsp", "with", "table", "cellspacing", "valign", "cellpadding",

"width", "top", "left", "sizepx", "telephone", "if", "may", "help", "us", "will", "please", "unless",

"visit","decembernbsp","januarynbsp","marchnbsp","formnbsp" ,"thnbsp","toppx","bottompx", "nnn", "address", "nonbsp", "new", "bottom", "em"))

**#Specific Transformation**:

toString <- content\_transformer(function(x, from, to) gsub(from, to, x))

qa\_sections <- tm\_map(qa\_sections, toString, "decembernbsp", "december")

qa\_sections <- tm\_map(qa\_sections, toString, "januarynbsp", "january")

qa\_sections <- tm\_map(qa\_sections, toString, "marchnbsp", "march")

qa\_sections <- tm\_map(qa\_sections, toString, "formnbsp", "form")

qa\_sections <- tm\_map(qa\_sections, toString, "informnbsp", "inform")

**#stemming**

qa\_sections <-tm\_map(qa\_sections, stemDocument)

**#plain text doc**

qa\_sections <-tm\_map(qa\_sections, PlainTextDocument)

qa\_dtm <- DocumentTermMatrix(qa\_sections)

**#inspect the element**

inspect(qa\_dtm[1:5, 100:105])

freq <- colSums((as.matrix(qa\_dtm)))

**#matrix**

m<-as.matrix(qa\_dtm)

**#save in CSV format**

write.csv(m,file= "C:/Users/gururaj/Desktop/Hewlett Packard/qa\_section/qa\_dtm.csv")

**#remove sparese items**

dtms <- removeSparseTerms(qa\_dtm, 0.1)

**#finding the frequency**

freq <-colSums(as.matrix(dtms))

freq

**#Patterns and Analytics**

**#find the words with minimum number of time occured**

findFreqTerms(dtms,lowfreq = 500)

**#finding co-relations for word "vote"**

jh<-findAssocs(dtms, "vote", 0.95)

**#sort out**

freq<-sort(colSums(as.matrix(dtms)),decreasing = TRUE)

freq

**#dataframe**

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

#Top 5 word counts and their plots

top<-head(wf,5)

subset(top)%>%ggplot(aes(word,freq))+geom\_bar(stat = "identity")

#Least 5 word counts and their plots

least<-tail(wf,5)

subset(least)%>%ggplot(aes(word,freq))+geom\_bar(stat = "identity")

**#for finding each and every word count for particular count from the corpus**

library(slam)

library(reshape2)

dtms.dense <- as.matrix(dtms)

dtms.dense = melt(dtms.dense, value.name = "count" )

head(dtms.dense)

#subjmeans <- cast(dtms.dense, id~count(), mean)

#ggplot subsetting and plotting purpose

**#plot for more than 400 counts**

subset(wf,freq>400)%>%ggplot(aes(word,freq))+geom\_bar(stat = "identity")

**#plot for words with counts more than average word counts**

average<-floor(mean(wf[["freq"]]))

subset(wf,freq<floor(mean(wf[["freq"]])))%>%ggplot(aes(word,freq))+geom\_bar(stat = "identity")

**#installing wordcloud**

set.seed(123)

wordcloud(names(freq),freq,min.freq = 150,scale = c(4,0.5),colors=brewer.pal(8, "Dark2"))

**#finding minimum terms per document**

minterms<-apply(dtms, 1, which.min)

dtms$dimnames$Terms[minterms]

**#finding maximum terms per document**

maxterms<-apply(dtms, 1, which.max)

dtms$dimnames$Terms[maxterms]

df.agg <- aggregate(count ~ Terms, dtms.dense, max)

# then simply merge with the original

df.max <- merge(df.agg, df.orig)

**#word counts of particular document in corpus**

word\_count<-rowSums(as.matrix(dtms))

names(word\_count) <- c("2005", "2", "3", "4", "5", "6", "7", "8", "9", "10", "11", "12")

barplot(word\_count,xlab = "Document\_numbers", ylab = "Word-Freq")