con=file("speech.txt","r",blocking=FALSE)

text=readLines(con)

close(con)

df=data.frame(text)

textdata=df[df$text, ]

library(tm)

doc.vec=VectorSource(textdata)#Converted textdata into vector form

doc.corpus=Corpus(doc.vec)#corpus is the set of dataset specifically alphabets

summary(doc.corpus)

docs=tm\_map(doc.corpus,removePunctuation)

docs=tm\_map(docs,removeNumbers)

docs=tm\_map(docs,tolower)

docs

docs=tm\_map(docs,removeWords,stopwords("English"))

#particulasr errors can be removed using docs=tm\_map(docs,removeWords,c("department","email"))

library(SnowballC)

docs=tm\_map(docs,stemDocument)#stemDocument eg ing,es,s,ess

docs=tm\_map(docs,stripWhitespace)

dtm=DocumentTermMatrix(docs)#A document-term matrix or term-document matrix is a mathematical matrix that describes the frequency of terms that occur in a collection of documents

#In a document-term matrix, rows correspond to documents in the collection and columns correspond to terms

inspect(dtm)

tdm=TermDocumentMatrix(docs)

tdm

freq=colSums(as.matrix(dtm))

length(freq)

ord=order(freq)

m=as.matrix(dtm)

dim(m)

write.csv(m,file="mdtm.csv")

dtms=removeSparseTerms(dtm,0.9)

inspect(dtms)

freq[head(ord)]

freq[tail(ord)]

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

head(wf)

library(ggplot2)

p=ggplot(subset(wf,freq>15),aes(word,freq))

p=p+geom\_bar(stat="identity")#identity is standard name

p=p+theme(axis.text.x=element\_text(angle=45,hjust=1))

findAssocs(dtm,"health",corlimit=0.49)# corlimit is used to adjust approximation to the word

library(wordcloud)

set.seed(142)#for 142 times the orientation will not change

wordcloud(names(freq),freq,min.freq = 4)

wordcloud(names(freq),freq,min.freq = 10,colors=brewer.pal(6,"Dark2"))