[1] Prepare one WORD file and you should have cleaning text and graphical view of text mining with word search, word count, cluster view and star graphical view.

```
install.packages('officer')
install.packages('dplyr')
install.packages('tm')
install.packages('ggplot2')
install.packages('wordcloud')


require(officer)#Access word documents
require(dplyr)#Manipulate Data
require(tm) #Text Mining
require(ggplot2)#Data Visualization
require(wordcloud)#create word cloud



sample_data <- read_docx("MyDocument.docx")
content <- docx_summary(sample_data)

#read text from the content variable
paragraphs <- content %>% filter(content_type == "paragraph")
Doc_Data<-paragraphs$text # Access the actual text
Doc_Data
#A corpus is a collection of texts, written or spoken, usually stored in a
database.

# convert the vector Doc_Data to a corpus
new_corpus <- Corpus(VectorSource(Doc_Data))



word.tdm <- TermDocumentMatrix(new_corpus)
inspect(word.tdm[1:100,]) # Examine  100 words at a time
#Examine the frequently appearing words in the term document matrix
FrequentTerms <- findFreqTerms(word.tdm, lowfreq = 5, highfreq = Inf)



#Convert term document matrix to data frame
word.tdm <- TermDocumentMatrix(new_corpus)
```

```r
m <- as.matrix(word.tdm)
v <- sort(rowSums(m),decreasing=TRUE)
d <- data.frame(word = names(v),freq=v)

set.seed(10000)
wordcloud(words = d$word, freq = d$freq, min.freq = 1,
          max.words=500, random.order=FALSE, rot.per=0.35,
          colors=brewer.pal(8, "Set1"))

barplot(d[1:11,]$freq, las = 2, names.arg = d[1:11,]$word,
        col ="pink", main ="Most frequent words",
        ylab = "Word frequencies")
```

[2] Develop prediction model to predict that person will purchase or not based on the data given in purchasedata.xlxs. Add extra 10 records of your choice in that file. Achieve accuracy of the model at least 75%.

Step 1: Read the data set in R. Put the R file and data set in same folder.

```r
data1 <- read.csv(file.choose(), header=T)

# display the data

data1
```

To read the excel file install the following packages

```r
install.packages("readxl")
```

Try to load it using

```r
library("readxl")
```

Read both xls and xlsx files

```r
library("readxl")

# xls files

my_data <- read_excel("my_file.xls")

# xlsx files

my_data <- read_excel("my_file.xlsx")
```

Step 2: we need to compute a linear model for this data frame:

```
# Creates a linear model
my_linear_model <- lm(dist~Purchased, data = df)

# Prints the model results
my_linear_model
```

Step 3: Now that we have a model, we can apply predict().

```
#Creating a data frame
variable_ Purchased <- data.frame(1,1,1,1,0,0,0)

# Fiting the linear model
linear_model <- lm(dist~ Purchased, data = df)

# Predicts the future values
predict(linear_model, newdata = variable_ Purchased)
```

[4] Develop prediction model to predict that person will suffer heart problem or not based on the data given in heartproblem.csv. Add extra 20 records in the same file. Achieve accuracy of the model at least 75%.

[5] Develop prediction model to predict that student will be placed or not based on the data given in Placement Details.xlxs. Apply appropriate preprocessing before prediction. Achieve accuracy of the model at least 75%.

[6] Develop prediction model to predict that person will have diabetes or not based on the data given in diabetes.csv file. Add such 20 records more in the same file. Achieve accuracy of the model at least 75%.

NOTE: Apply the same code on all data sets.