Assignment No. 2.1

Aim: Perform text analysis using R/ Python.

Theory:

The objective of this practical is to perform text analysis on a dataset of customer reviews to uncover insights and patterns in textual data. This is achieved using various R libraries for text mining and visualization. The process begins with reading and preprocessing the data, where the text is cleaned by converting it to lowercase, removing numbers, punctuation, and common stopwords, and applying stemming to reduce words to their base forms. This ensures uniformity and highlights meaningful terms. A Term-Document Matrix (TDM) is then constructed to quantify the frequency of each term across the corpus, enabling a structured analysis of the text. Key insights are visualized through bar plots and word clouds, showcasing the most frequently occurring words. These visual tools provide a clear understanding of prevalent themes, helping to identify trends, common opinions, and potential areas of interest in the customer feedback.

CODE:

```
install.packages("tm") # for text mining

install.packages("SnowballC") # for text stemming

install.packages("wordcloud") # word-cloud generator

install.packages("RColorBrewer") # color palettes

install.packages("ggplot2") # for plotting graphs


library("tm")

library("wordcloud")

library("SnowballC")

library("RColorBrewer")

library("ggplot2")


data <-  read.csv('D:\MIT ADT\LY - Sem 1\BDA Lab\Amreen Mam\Assign
2\preprocessed_kindle_review.csv',stringsAsFactors = FALSE)
```

```
head(data)

text_column <- data$reviewText
TextDoc <- Corpus(VectorSource(text_column))

toSpace <- content_transformer(function(x, pattern)gsub(pattern, " ", x))
TextDoc <- tm_map(TextDoc, toSpace, "/")
TextDoc <- tm_map(TextDoc, toSpace, "@")
TextDoc <- tm_map(TextDoc, toSpace, "\\|")

TextDoc <- tm_map(TextDoc, content_transformer(tolower))
TextDoc <- tm_map(TextDoc, removeNumbers)

TextDoc <- tm_map(TextDoc, removeWords, stopwords("english"))



TextDoc <- tm_map(TextDoc, removeWords,c("kindle", "amazon", "product", "review",
"customer",
                                         "one", "two", "three", "four", "five"))

TextDoc <- tm_map(TextDoc, removePunctuation)
TextDoc <- tm_map(TextDoc, stripWhitespace)
TextDoc <- tm_map(TextDoc, stemDocument)
TextDoc_dtm <- TermDocumentMatrix(TextDoc)
dtm_m <- as.matrix(TextDoc_dtm)

dtm_v <- sort(rowSums(dtm_m), decreasing=TRUE)
dtm_d <- data.frame(word=names(dtm_v), freq=dtm_v)

head(dtm_d, 8)

#          word      freq
#book       book 15694
#stori     stori 11325
```

#read read 10920

#like like 6583

#charact charact 5763

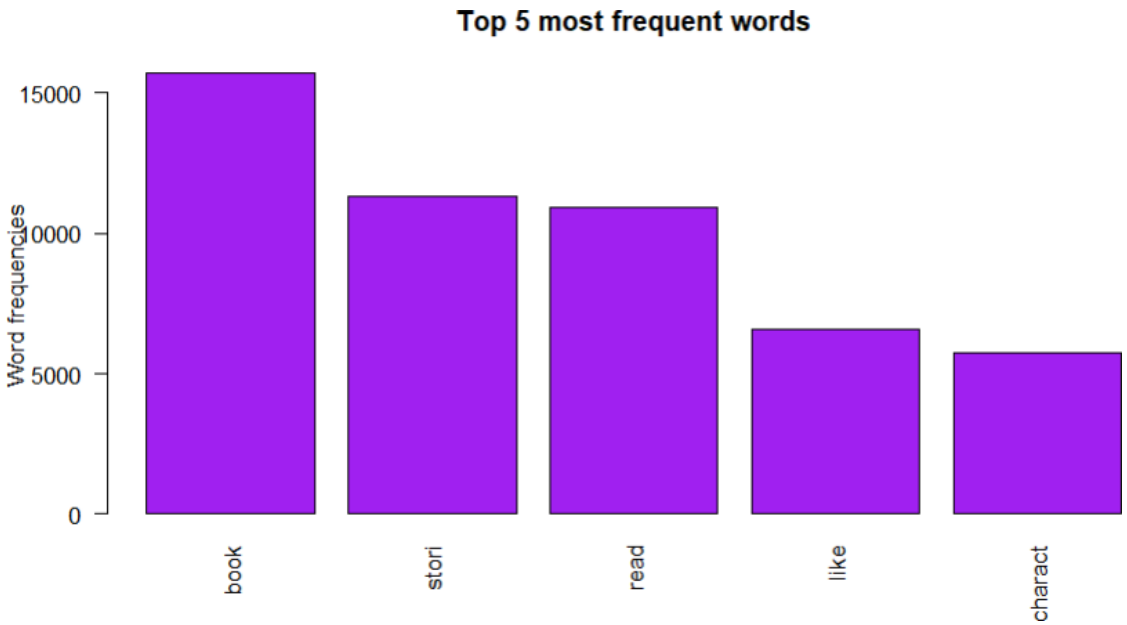
#just just 5492

#love love 5359

#good good 4411

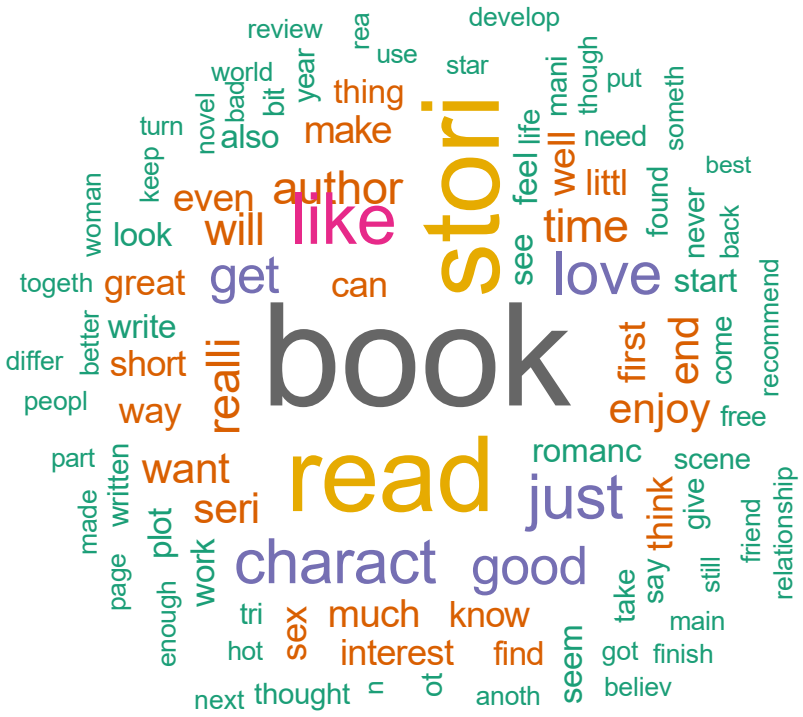
```
> head(dtm_d, 8)
           word  freq
book       book 15694
stori     stori 11325
read       read 10920
like       like  6583
charact charact  5763
just       just  5492
love       love  5359
good       good  4411
```

```
barplot(dtm_d[1:5,]$freq, las = 2, names.arg = dtm_d[1:5,]$word,
        col ="purple", main ="Top 5 most frequent words",
        ylab = "Word frequencies")
```




```
set.seed(1234)
wordcloud(words = dtm_d$word, freq = dtm_d$freq, min.freq = 5,
          max.words=100, random.order=FALSE, rot.per=0.40,
```

colors=brewer.pal(8, "Dark2"))



Conclusion:

Text analysis revealed key themes and sentiment in customer reviews by preprocessing the data and visualizing word frequencies using bar plots and word clouds.