For Text Mining assignment

TWO:

1) Extract reviews of any product from ecommerce website like amazon

2) Perform sentimental analysis

**Ans :**

**R Code :**

## Text Mining & Sentimental Analysis

########## Amazon Reviews #########

library(rvest)

library(XML)

library(magrittr)

aurl <- "https://www.amazon.in/Apple-MacBook-Air-13-3-inch-Integrated/product-reviews/B073Q5R6VR/ref=cm\_cr\_arp\_d\_paging\_btm\_3?showViewpoints=1&pageNumber"

amazon\_reviews <- NULL

for (i in 1:10){

murl <- read\_html(as.character(paste(aurl,i,sep="=")))

rev <- murl %>% html\_nodes(".review-text") %>% html\_text()

amazon\_reviews <- c(amazon\_reviews,rev)

}

length(amazon\_reviews)

write.table(amazon\_reviews,"apple.txt",row.names = F)

getwd()

str(amazon\_reviews)

fin\_txt <- amazon\_reviews

library(tm)

fin\_corpus<- Corpus(VectorSource(fin\_txt))

inspect(fin\_corpus[100])

fin\_clean<-tm\_map(fin\_corpus, removePunctuation)

fin\_clean<-tm\_map(fin\_clean, content\_transformer(tolower))

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

fin\_clean<-tm\_map(fin\_clean,removeNumbers)

fin\_clean<-tm\_map(fin\_clean, stripWhitespace)

fin\_clean<-tm\_map(fin\_clean, removeWords, c("gameofthrones")) ## clean some words

library(wordcloud)

wordcloud(fin\_clean, random.order = F, max.words = 5,colors=rainbow(50))

wordcloud(fin\_clean, rot.per=0.5, random.order=TRUE,colors=brewer.pal(8, "Dark2"))

# Term document matrix

# converting unstructured data to structured format using TDM

tdm <- TermDocumentMatrix(fin\_clean)

dtm <- t(tdm)

tdm <- as.matrix(tdm)

# Bar plot

w <- rowSums(tdm)

w

w\_sub <- subset(w, w >= 20)

w\_sub

windows()

barplot(w\_sub, las=3, col = rainbow(20))

fin\_clean <- tm\_map(fin\_clean, removeWords, c('apple','air','can','mcds','macbook','product','windows'))

fin\_clean <- tm\_map(fin\_clean, stripWhitespace)

tdm <- TermDocumentMatrix(fin\_clean)

tdm <- as.matrix(tdm)

# Word cloud

install.packages("wordcloud")

library(wordcloud)

windows()

wordcloud(words = names(w\_sub), freq = w\_sub) # wordcloud with only subset of words

w\_sub1 <- sort(rowSums(tdm), decreasing = TRUE)

wordcloud(words = names(w\_sub1), freq = w\_sub1) # all words are considered

windows()

wordcloud(words = names(w\_sub1), freq = w\_sub1, random.order = F, colors = rainbow(20), scale=c(3,1), rot.per = 0.3)

# lOADING +VE AND -VE dictonaries

pos.words = scan(file.choose(), what="character", comment.char=";") # read-in positive-words.txt

neg.words = scan(file.choose(), what="character", comment.char=";") # read-in negative-words.txt

pos.words = c(pos.words,"wow", "kudos", "hurray") # including our own positive words to the existing list

# Positive wordcloud

pos.matches = match(names(w\_sub1), c(pos.words))

pos.matches = !is.na(pos.matches)

freq\_pos <- w\_sub1[pos.matches]

p\_names <- names(freq\_pos)

windows()

wordcloud(p\_names,freq\_pos,scale=c(4,1),colors = rainbow(20))

# Negative wordcloud

neg.matches = match(names(w\_sub1), c(neg.words))

neg.matches = !is.na(neg.matches)

freq\_neg <- w\_sub1[neg.matches]

n\_names <- names(freq\_neg)

windows()

wordcloud(n\_names,freq\_neg,scale=c(5,1),colors = brewer.pal(8,"Dark2"))

#### emotion mining ####

install.packages("syuzhet")

library("syuzhet")

library(lubridate,ggplot2)

library(ggplot2)

library(scales)

library(dplyr)

library(reshape2)

x <- get\_nrc\_sentiment(fin\_txt)

head(x,n=5)

fin\_txt[4]

get\_nrc\_sentiment('happy')

get\_nrc\_sentiment('boring')

get\_sentiment('boring',method="afinn")

get\_sentiment('happy',method="afinn")

#each sentences by eight

example<-get\_sentences(fin\_txt)

nrc\_data<-get\_nrc\_sentiment(example)

# Bar plot for emotion mining

windows()

barplot(colSums(nrc\_data), las = 1, col = rainbow(10), ylab = 'Count', main = 'Emotion scores')

sentiment\_vector<-get\_sentiment(example,method="bing")

sentiment\_afinn<-get\_sentiment(example,method="afinn")

sentiment\_nrc<-get\_sentiment(example,method="nrc")

sum(sentiment\_afinn)

mean(sentiment\_afinn)

summary(sentiment\_afinn)

windows()

plot(sentiment\_vector,type='l',maim='Plot trajectory',xlab='Narative time',ylab='Emotional valence')

abline(h=0,color='red')

plot(

sentiment\_vector,

type="h",

main="Example Plot Trajectory",

xlab = "Narrative Time",

ylab= "Emotional Valence"

)

##Shape smoothing and normalization using a Fourier based transformation and low pass filtering is achieved using the get\_transformed\_values function as shown below.

ft\_values <- get\_transformed\_values(

sentiment\_vector,

low\_pass\_size = 3,

x\_reverse\_len = 100,

padding\_factor = 2,

scale\_vals = TRUE,

scale\_range = FALSE

)

plot(

ft\_values,

type ="l",

main ="Apple Macbook Reviews using Transformed values",

xlab = "Narrative Time",

ylab = "Emotional Valence",

col = "red"

)

#Most Negative and Positive reviews

negative<-example[which.min(sentiment\_vector)]

positive<-example[which.max(sentiment\_vector)]

################################################

my\_example\_text <- "I begin this story with a neutral statement.

Basically this is a very silly test.

You are testing the Syuzhet package using short, inane sentences.

I am actually very happy today.

I have finally finished writing this package.

Tomorrow I will be very sad.

I won't have anything left to do.

I might get angry and decide to do something horrible.

I might destroy the entire package and start from scratch.

Then again, I might find it satisfying to have completed my first R package.

Honestly this use of the Fourier transformation is really quite elegant.

You might even say it's beautiful!"

#x-axis represents the passage of time from the beginning to the end of the text, and the y-axis measures the degrees of positive and negative sentiment

s\_v <- get\_sentences(my\_example\_text)

s\_v\_sentiment <- get\_sentiment(s\_v)

plot(

s\_v\_sentiment,

type="l",

main="Example Plot Trajectory",

xlab = "Narrative Time",

ylab= "Emotional Valence"

)

**Results :**

> length(amazon\_reviews)

[1] 100

> write.table(amazon\_reviews,"apple.txt",row.names = F)

> getwd()

[1] "D:/Data Science/Excelr/Assignments/Assignment/Text Mining"

> str(amazon\_reviews)

chr [1:100] "\n\n\n\n\n\n\n\n \n \n Okay. So you have been seeing people display their Macbooks and you have always fan"| \_\_truncated\_\_ ...

> fin\_corpus<- Corpus(VectorSource(fin\_txt))

> inspect(fin\_corpus[100])

<<SimpleCorpus>>

Metadata: corpus specific: 1, document level (indexed): 0

Content: documents: 1

[1] \n\n\n\n\n\n\n\n \n \n From the very beginning the battery life was horrible . hardly 2 hours. And after using it for just a year it just stopped working . and i cannot even charge it. Terrible product sold by Amazon\n \n

> tdm <- TermDocumentMatrix(fin\_clean)

> dtm <- t(tdm)

> tdm <- as.matrix(tdm)

> # Bar plot

> w <- rowSums(tdm)

> w

able abovea activities adapter adds

2 1 1 2 3

advantage ago air alone along

3 2 41 1 2

alongside alternatives always amazon amount

1 1 5 33 3

another anything apple apples around

5 2 99 2 21

asin asinbjwbkxf available back bare

1 1 8 13 1

bdvmgm best better beyond blue

1 18 18 1 2

bootcamp bootng bugs buy buying

1 1 1 42 14

can carrying cell certain chances

34 2 1 1 2

change cheap comes coming consider

2 1 4 2 4

course cozy crash cursor data

2 1 2 1 3

decide dell depend desktop device

2 5 1 2 12

devices differently difficult dinosaur discourage

5 1 2 2 1

display dual durability even evn

15 1 1 16 1

exfat expensive explain extend face

1 10 1 1 1

fact fancied feature feel feeling

2 1 5 3 2

find first forced fragile friends

6 7 1 5 2

functions generic get getting give

1 1 25 5 4

giving good grace great greatnow

7 52 2 9 1

handy hanged happy hardware hence

2 1 2 7 3

hey however improves increasing iphone

1 6 1 2 3

jail kept know laid laptop

2 2 10 1 61

last latest least left life

13 1 3 3 22

lifetime lightweight like line linux

1 1 37 2 6

linuxthe little load longer look

1 12 2 1 9

looks luxury mac macbook macbooks

3 1 57 65 2

machines make makes many matched

1 7 3 9 1

matter may money months mostly

1 6 11 15 4

mouse much multitasking myselfknow need

1 27 1 1 20

neighbouring never next nonstandard note

1 16 4 1 2

now obsolete offers official okay

30 4 4 1 1

one operating optimal option outside

30 7 1 9 2

overall owning page pain parallels

4 1 3 2 2

partitioning peddlesif people perfectly performance

2 1 11 5 11

period plan please plight plug

3 3 10 1 2

possible power premium premiumness prepared

3 15 11 1 1

pretend price pride profile protect

1 20 2 2 1

pulls puny purchased quite ram

2 1 10 4 2

rare rating reading real realise

1 2 3 3 2

reason rebooting recognise reduced resist

5 1 1 1 1

review risk run rushed saving

11 2 8 1 4

screens seamlessly second see seeing

1 1 1 3 1

sells shell shop showing sides

1 3 1 1 1

single sluggish smirk snobrace software

3 1 1 1 6

something soon spares spend spite

2 2 3 5 1

ssd stable stars stick still

7 2 1 3 15

stop stranded stuck successfully support

1 1 4 1 15

system systems take temptation tempted

8 2 3 1 1

thankfully thats think thousands till

1 3 12 3 5

today tomorrow toysif training transferring

2 1 1 2 1

truth try understand updates upgradable

1 7 2 5 1

upgrade upgradeeditadd usb use user

9 1 8 24 13

users using value virtual wil

6 21 1 1 1

will windows withdrawn without workhorse

51 37 1 8 2

works wrote year yes amazing

17 1 23 3 7

battery coders hrsbest students ahead

39 1 1 2 4

angry antivirus apps away bad

1 1 9 2 8

bag based become bird bother

2 2 2 1 2

cant carry cherished clarity compatibility

6 4 3 5 4

compatible con connect connected connectors

1 3 2 2 2

costs critical customer defunct different

6 2 11 1 5

disc disk distortion diwali document

1 6 2 2 1

doesnt drive economical either else

5 1 1 2 6

emotional environment existing external extra

2 2 1 7 3

figure finally format free google

3 3 1 7 1

got gradually handled happeningoverall hard

23 3 2 2 11

help huge inability india indian

6 3 1 5 2

issue issues keynote laptopgood laptops

20 14 1 1 10

laser less long machine manuals

1 11 8 21 2

market marketing master mentioned mfd

2 2 1 6 1

news newsluxury niche open otherwise

1 1 1 3 3

pages paid planned play ports

2 3 2 5 2

possession powerpoint presently pridebad printer

2 1 2 1 1

printers problems product projectors properly

1 9 49 2 5

reliable required ricoh sale section

1 4 1 14 3

security seem sense silent slides

2 2 3 2 2

space store strategy sweep switching

1 7 2 2 3

technically technology terrific theres thing

1 2 1 5 15

transactions transfer unless unreasonably used

1 1 4 1 7

verdictif word working writer years

2 4 23 3 18

accounting app appsnumbers busy convert

1 7 1 1 1

csv dont drain emulator every

1 23 1 1 8

excel excelpages file filesame fine

1 1 3 1 11

freebut games heat install just

1 2 4 3 37

keynoteso likeissue love lover\U0001f601 mecons

1 1 6 1 1

mobile monthafter movies numbers office

1 1 7 1 10

officeits payyou pirated powerpointand pubg

1 1 1 1 2

shows sleek softwares storage substitute

1 2 4 6 2

talking tally term version want

1 1 2 5 13

wordkeynote work \U0001f60dnow avoids awesome

1 20 1 1 10

beautiful cost enough fast feels

1 31 2 7 2

highly kcons kind lag light

2 1 2 3 16

limited memory optimize prime process

1 5 1 1 1

pros response sound super time

3 3 9 6 22

userfriendlyprobably weight bought case diagnostic

1 8 30 1 1

expecting ive lasts pay recommended

1 2 1 10 2

repair saying across actually adaptor

14 2 3 2 2

additional adjustable adjusts advantagesi ambient

4 1 2 2 2

android angle answer audio automatically

3 2 2 3 4

aware backup basic basically batterywhom

3 3 4 2 1

benefit benefits big body bonus

1 4 3 2 1

brief brightness browse bucks buttons

2 5 1 2 1

calendar calls came camera careless

2 5 7 2 2

cashback chance charger check closed

1 2 4 4 2

color come comfortable comment compared

2 6 4 3 9

comparing confused construction consumption contains

2 3 2 1 2

conversion core cpu daily day

1 1 8 4 10

days deal deeper demerits despite

16 3 1 3 3

details difference displayfull documents dropped

1 4 2 1 2

editing encoding energy etc evolved

7 1 1 11 1

experience faster fifth finger force

12 2 1 2 2

formacbook front gaming generation ghz

1 2 11 2 2

hands hassle hdmi helps helpsalso

5 4 1 3 1

hope hours idle ignore impressions

2 15 2 1 2

intel intensive internet justifies keep

2 2 4 1 5

keyboard keys launch leave listing

18 5 4 4 2

lot loud loves low lower

7 2 1 10 2

mag mail mainly metal might

2 6 1 2 4

neat new newer non normal

2 16 3 6 3

notable notes old options optionsif

1 2 3 5 1

ordering osair pad pads part

1 2 4 2 5

photo photoshop pixels port premiumshould

1 1 2 5 2

pressure previous pro proair processor

5 1 36 2 4

processors programming put quality queries

1 5 3 32 1

quicktime range raw really regular

2 2 1 11 4

responsive retina safe save say

4 3 4 2 8

seems sensor side silver similar

3 2 2 1 4

skype slim smooth someone somewhere

2 3 5 1 2

speakers speed spreadsheets struggle studio

2 4 2 1 1

tasks testing tests thisi todos

4 1 1 1 2

touch travel twothanks type ultrabook

7 2 1 1 2

update usage uses video viewing

3 7 2 11 2

vqr watch weeks whether wider

1 2 5 3 2

world xcode asked board centre

3 2 3 18 13

checked collapsed december didn’t disappointed

9 1 2 1 5

don’t due expiry extension fixed

3 3 1 1 2

informed january luckily mother refused

3 1 3 4 6

replacement service start took upset

9 44 5 11 1

warranty april can’t charge dead

30 3 2 11 12

ever fix full recently request

6 2 2 1 2

responsibility says seller since solid

1 6 2 4 1

solution turn button defective flicker

1 6 6 10 1

motherboard piece replaced screen shutdown

10 4 16 18 1

sometime star started blanket bring

1 2 3 1 1

calm cold created cuddling date

1 1 1 1 3

deep electronic electrostatic enjoying excitement

2 2 1 1 1

expiration forever fortune icare included

1 2 1 2 1

jaipuri mood night nothing sleep

1 1 2 6 1

spacebar spending spike though thousand

1 2 1 7 2

week went youtube breeze build

3 7 1 2 4

design discount exactly excellen expect

3 1 1 1 7

incase keyboardcons laptopyou macos mark

1 1 1 6 1

media mojave notch nullifies resolution

2 2 1 1 3

ships sierra top upto awesomehowever

1 2 2 3 1

music production doubting products purchase

1 1 1 15 14

sold three white acting brands

6 3 1 1 3

experiences far mean monthsi priced

2 1 2 1 1

somebody’s station times visit heating

1 1 2 2 4

ignored issuesbut minebest programmingonly care

1 1 1 1 7

charging companies earn extended feedback

5 1 1 3 1

hear longlasting meant month prefer

1 1 1 6 1

selling shocked stopped suddenly told

1 4 13 5 13

carefully defected helpless invest logic

3 2 3 2 14

sent switch unable well worked

3 3 3 8 5

gud adding also appario bothered

1 1 11 1 1

corrosion disappointment electronics idea july

3 2 2 3 1

nov online refusal sometimes stating

1 7 1 2 2

turning wonder serviceno solutions worst

1 1 1 1 7

almost amazonapple conveniently customers dime

5 2 1 3 1

disheartened expired fan happened hate

1 2 5 2 1

major read sell among anyone

1 3 1 1 5

appleappleapple applications appropriate atleast backgroundmacos

1 4 2 1 1

backlightthe badthe base bass bedroomvery

1 1 1 1 1

boot breif butter call camp

1 1 2 1 1

card ccant center changed checkedeveryday

1 1 18 2 1

checking cinebench cinebenchspeakers continuous cooling

2 1 1 1 2

criteria cut damn degrees detect

1 3 1 2 1

distributed dusted easily edit editingthe

1 1 4 1 1

egpages everything explained explorer exporting

1 7 1 1 2

extremely fanram final finethe flashed

2 1 4 1 1

fps gamerintel gbs geforce given

2 1 1 1 2

god going gpu graphics grthe

1 7 1 6 1

hardcore heated imac immediately inch

3 1 1 2 1

installedthe integrated iris issua itdisplay

1 1 1 1 1

kfps laptopgets laptopprosk lenovo lightweightmacos

2 1 1 1 1

living made management mind model

1 1 1 2 2

must network nits oct overpowered

8 1 1 2 1

partgood paste perfect plus priceno

1 2 2 2 1

radeon recommend room router running

1 4 1 1 3

saw school score scores seen

1 1 1 1 2

solved speaker spent ssdconsdisplay standardsno

1 1 2 1 1

structure thank thanks thermal thick

1 1 2 3 1

thin thisupdate thought throttling thunderbolt

1 1 1 1 1

todays total trackpad trackpadbezels vega

1 1 4 1 1

verdictthis videos view way whole

1 2 1 3 6

wifi wordmagsafe writing yay yearsmy

2 1 3 1 1

accessories aspect books doubt excellent

1 2 1 1 5

vouch amazons caps delivery exchange

1 2 1 2 4

fault fully gives hasslefree item

8 1 1 1 3

key kudos lesser lock modular

3 1 1 1 1

offer original packed received safety

2 4 1 4 1

[ reached getOption("max.print") -- omitted 789 entries ]

> w\_sub <- subset(w, w >= 20)

> w\_sub

air amazon apple around buy can get good laptop life like mac

41 33 99 21 42 34 25 52 61 22 37 57

macbook much need now one price use using will windows year battery

65 27 20 30 30 20 24 21 51 37 23 39

got issue machine product working dont just work cost time bought pro

23 20 21 49 23 23 37 20 31 22 30 36

quality service warranty

32 44 30

> x <- get\_nrc\_sentiment(fin\_txt)

> head(x,n=5)

anger anticipation disgust fear joy sadness surprise trust negative positive

1 6 12 2 14 6 11 6 14 21 22

2 1 0 0 0 0 0 0 0 1 0

3 4 4 4 4 4 5 3 7 10 19

4 2 8 1 2 5 1 2 4 4 14

5 2 3 1 1 3 2 1 3 4 5

> fin\_txt[4]

[1] "\n\n\n\n\n\n\n\n \n \n Best Laptop under 60k. If you don't want any accounting softwares like tally or busy then go for it, better than windows in every term. I just love it's battery life with sleek look \U0001f60dNow talking about MS officeIt's available on app store, free for 1 monthAfter that you will have to pay.You cannot install pirated version but if you install Windows 7 or any other version then u can use Ms office for freeBut in MacBook there are other apps which are good substitute for Ms office appsNumbers for ExcelPages for wordKeynote for PowerPointAnd you can convert your numbers file to Excel file or csv fileSame for the pages & keynoteSo no need of Ms office for me.Cons1. U cannot play games, your laptop will just heat up, battery will drain\* U cannot play pubg mobile on emulator if you are pubg lover\U0001f6012. Storage 128gb, it will work fine but if you want to store movies or TV shows u can't, u have to buy external hard disk\n \n"

> get\_nrc\_sentiment('happy')

anger anticipation disgust fear joy sadness surprise trust negative positive

1 0 1 0 0 1 0 0 1 0 1

> get\_nrc\_sentiment('boring')

anger anticipation disgust fear joy sadness surprise trust negative positive

1 0 0 0 0 0 0 0 0 1 0

> get\_sentiment('boring',method="afinn")

[1] -3

> get\_sentiment('happy',method="afinn")

[1] 3

> #each sentences by eight

> example<-get\_sentences(fin\_txt)

> nrc\_data<-get\_nrc\_sentiment(example)

> # Bar plot for emotion mining

> windows()

> barplot(colSums(nrc\_data), las = 1, col = rainbow(10), ylab = 'Count', main = 'Emotion scores')

> sentiment\_vector<-get\_sentiment(example,method="bing")

> sentiment\_afinn<-get\_sentiment(example,method="afinn")

> sentiment\_nrc<-get\_sentiment(example,method="nrc")

> sum(sentiment\_afinn)

[1] 419

> mean(sentiment\_afinn)

[1] 0.592645

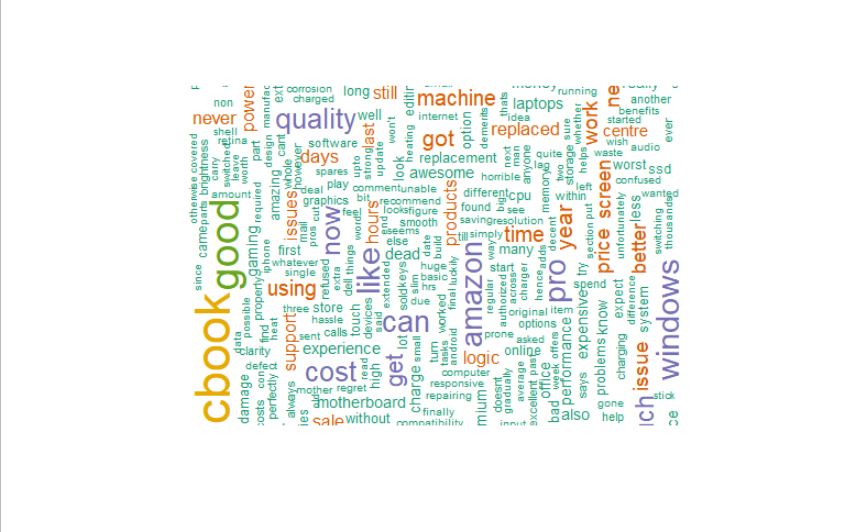
> summary(sentiment\_afinn)

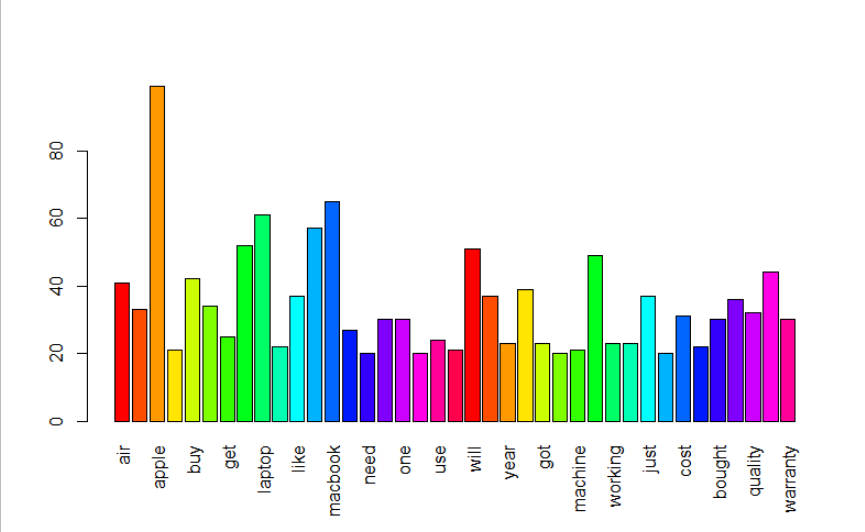
Min. 1st Qu. Median Mean 3rd Qu. Max.

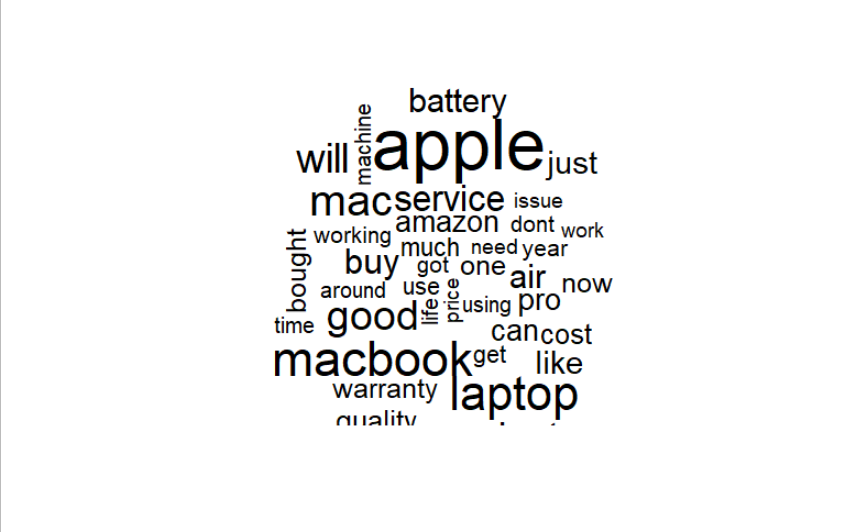
-7.0000 0.0000 0.0000 0.5926 2.0000 15.0000

**Plots :**

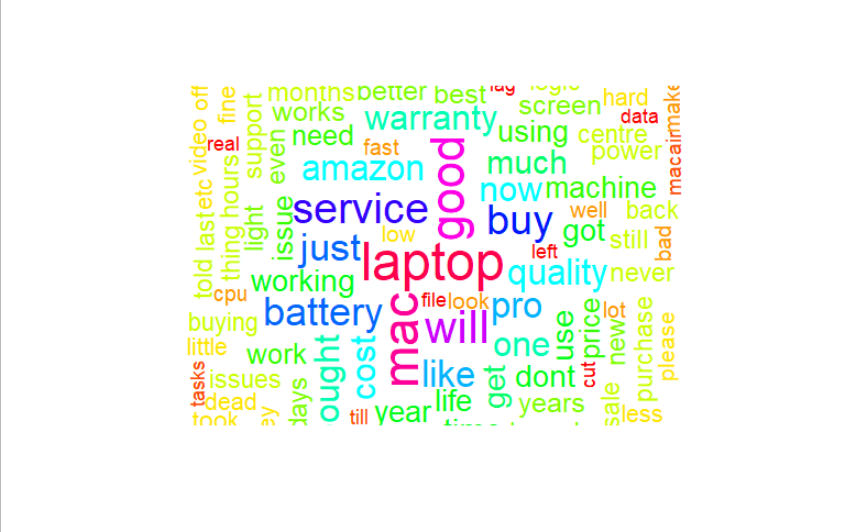




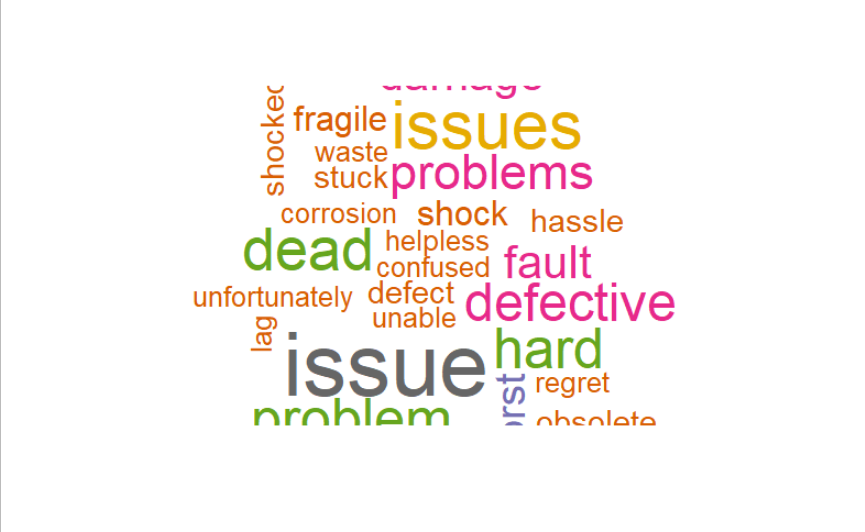


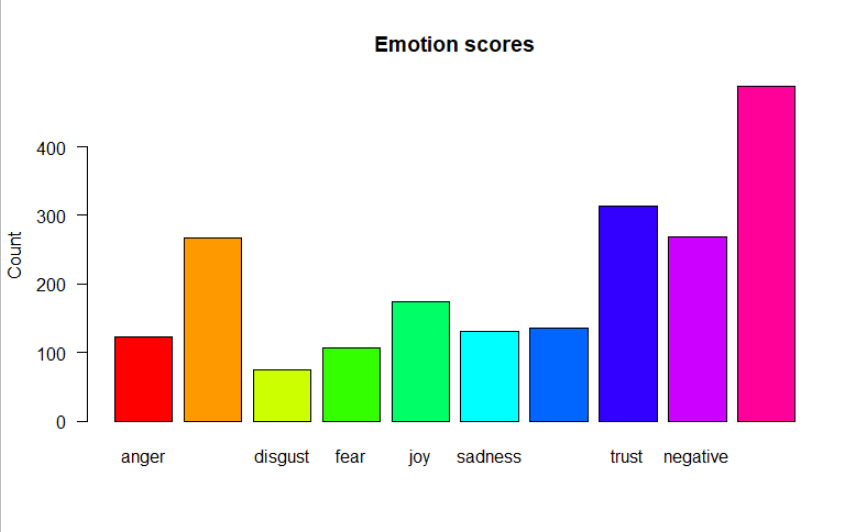


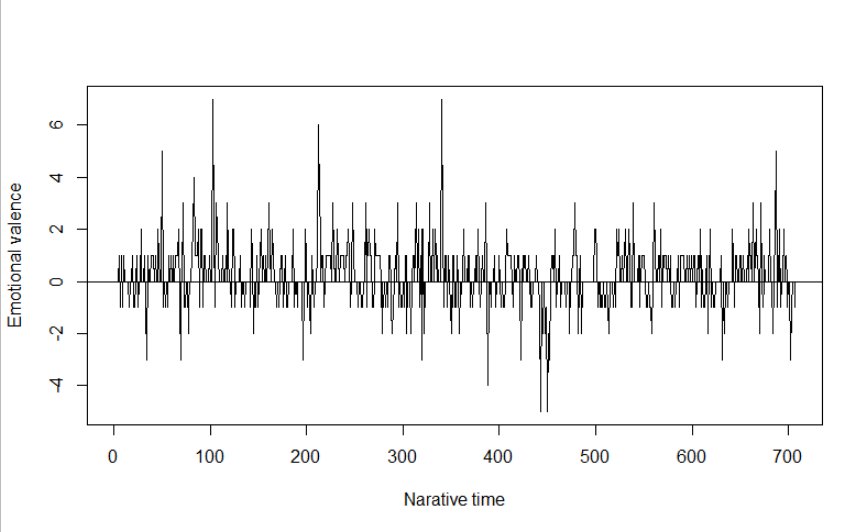


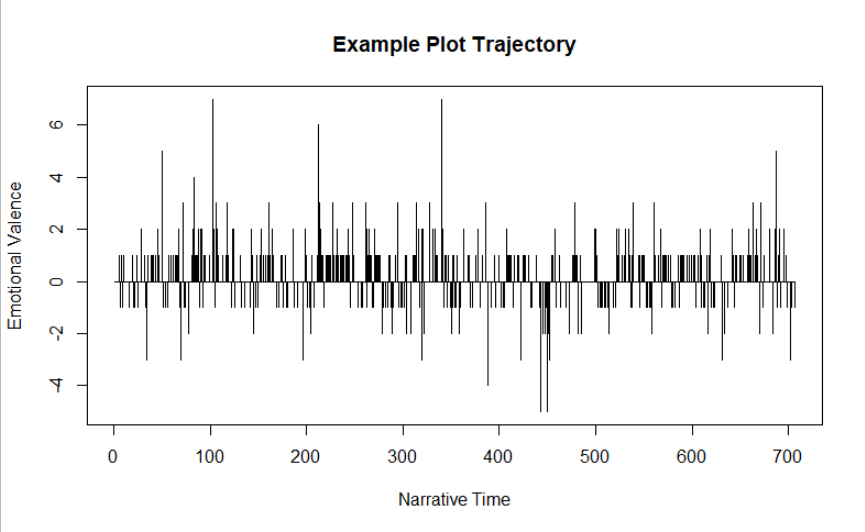


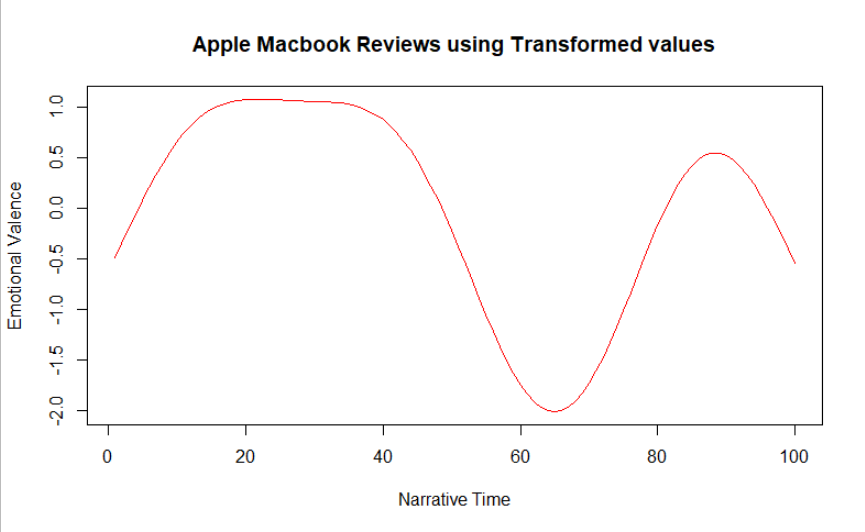














**Inference :**

The mean sentiment value is of 0.59. Hence, we can infer that the reviews are mostly on positive side.