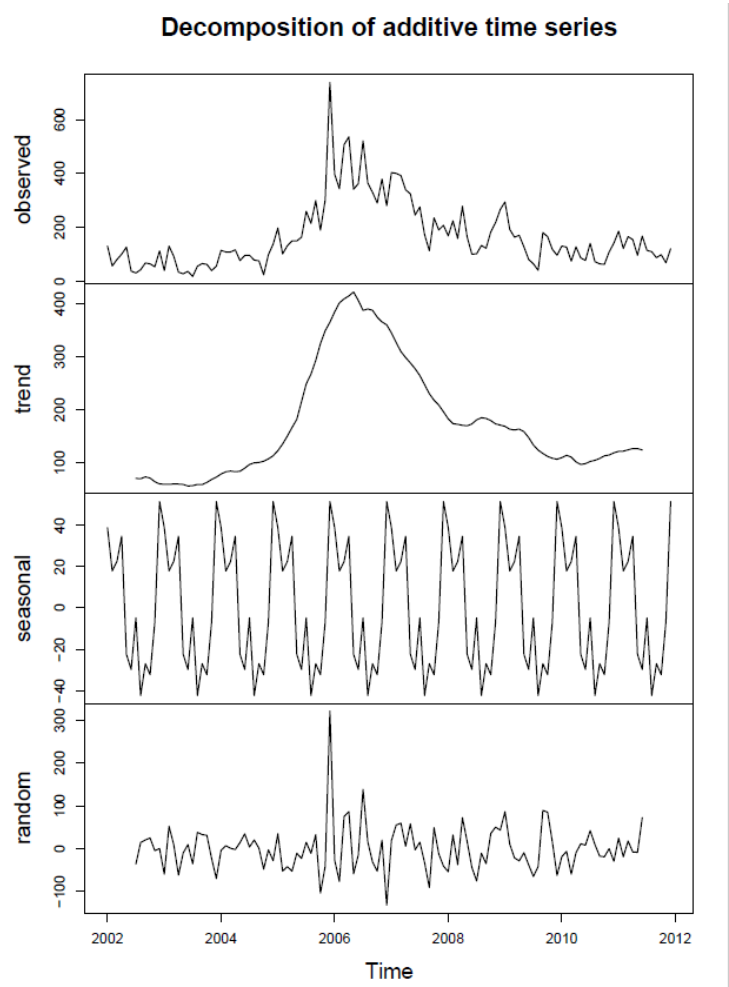
FIT3152 Data analytics – 2022: Assignment 1

Question(a)(1)



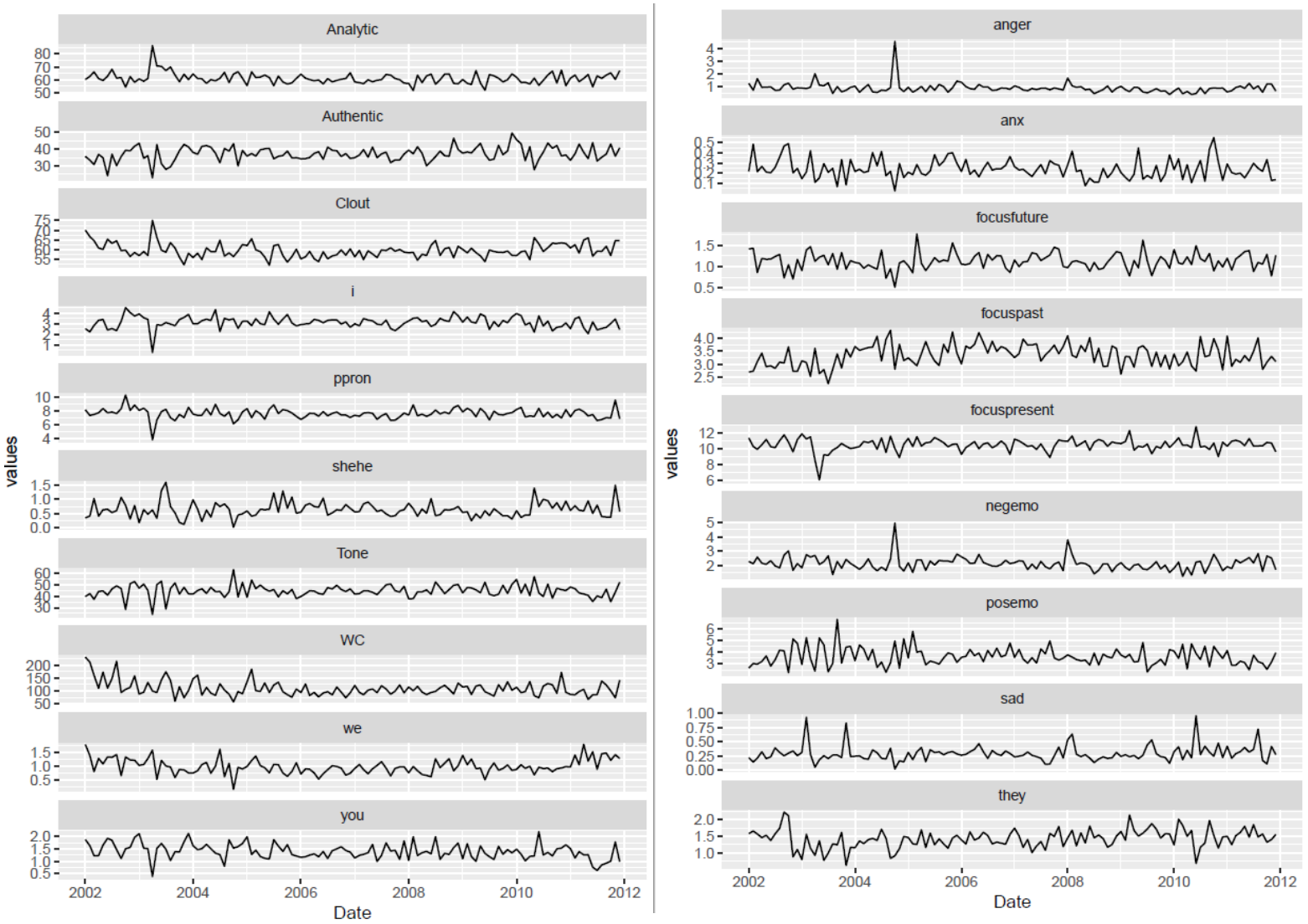
To see how active are participants over the longer term, I would sum up the number of posts posted each and every month from year 2002 to 2012.It is because I think that number of posts posted can determine how active are participants on the forum, if the number of posts are large, it indicates that the participants are active posting on the forum.

I am using a fine time division of year-month for time-series analysis so that I won't miss out the details and I can identify specifically the participants are most active and least active on which month of the year.

To observe the trend and activity of participants on the forum over time, I plotted a time-series decomposition. We can clearly see that that's a change in the direction of the trend in the end of year 2006.It is observed that there's a trend of increasing activity of the participants from year 2002 to 2006, then the trend changed in direction and we can see that there's also unexpected drop in the observed values. The peak of the observed values is on December 2005. Moreover, we have a downtrend of activity of participants from year 2007 to year 2010.

The activity of the forum is the least in around the end of year 2003.From my analysis, the amount of posts posted on that month (July 2003) was only 17.

Question (a)(2)



In order to observe whether the linguistic variables change over the duration of the forum, I plotted two graphs which showing the monthly average values of each linguistic variable over time.

Firstly, for Analytic, there's no obvious trend in the monthly average values of Analytical thinking of the posts, the values are around between 50-70 over time, except on one month in 2003, the Analytic was above 80.

For Authentic, there's no obvious trend in monthly average values of Authentic, the monthly average values are between 30-50, except for 2 months in 2003, there's an obvious drop of Authentic values to below 30.

For Clout, there's also no obvious trend in monthly average values, the values are between 50-70, except for one month in 2003, it went up to above 70, the overall pattern is quite similar to the pattern of Analytic, so I performed a t-test to check if their true difference in means is 0. The hypothesis testing is performed below.

Hypothesis testing:

Null hypothesis: true difference in means of Analytic and Clout is equal to 0

Alternative hypothesis: true difference in means of Analytic and Clout is not equal to 0

P-value is 0.001346, which is smaller than the significance p-value of 0.05, so the null hypothesis is rejected, so there's no evidence that the true difference in means of Analytic and Clout is equal to 0.

****

In short, although the pattern of their values over time looks similar, we cannot conclude that their true difference in means is equal to 0.

Furthermore, we are not so interested at analysing the values of "ppron","i","we","you",and "they",as they are just personal pronouns. Not to mention,we can see from the plotted graph that there is not obvious trend in their monthly average values over time.

Besides,we also do not see obvious trend in the monthly average values of Tone over time,the values are mostly between 30-60.Next,WC,which is word count. We can see that there's an obvious downtrend of monthly average word count from 2002 to 2005.From the beginning of year 2002,the monthly average wc is at peak,which is over 200,then it went down to 50 in the end of year 2004.After year 2004,theres no obvious trend in the monthly average word count, the values are mostly between 75-150.

Next,"anger",there's no obvious trend over time,the monthly average values are mostly between 0.75-1.50,except for 1-2 months in year 2004,it went above 4.0,it is likely that there were some global issues which made people angry happened at that period of time.

Moreover,for "anx",theres no obvious trend over time, the monthly average values are mostly between 0.1 and 0.4. Then,for "focusfuture", theres no obvious trend over time, the monthly average values are mostly between 0.75-1.50,except for in the year 2005,the monthly average values are relatively higher than the others.

For "focuspast",the monthly average values from year 2002-2004 are relatively low,and mostly below 3.5.Then,for the following years there's no obvious trend observed and the values are mostly between 3.0-4.0.

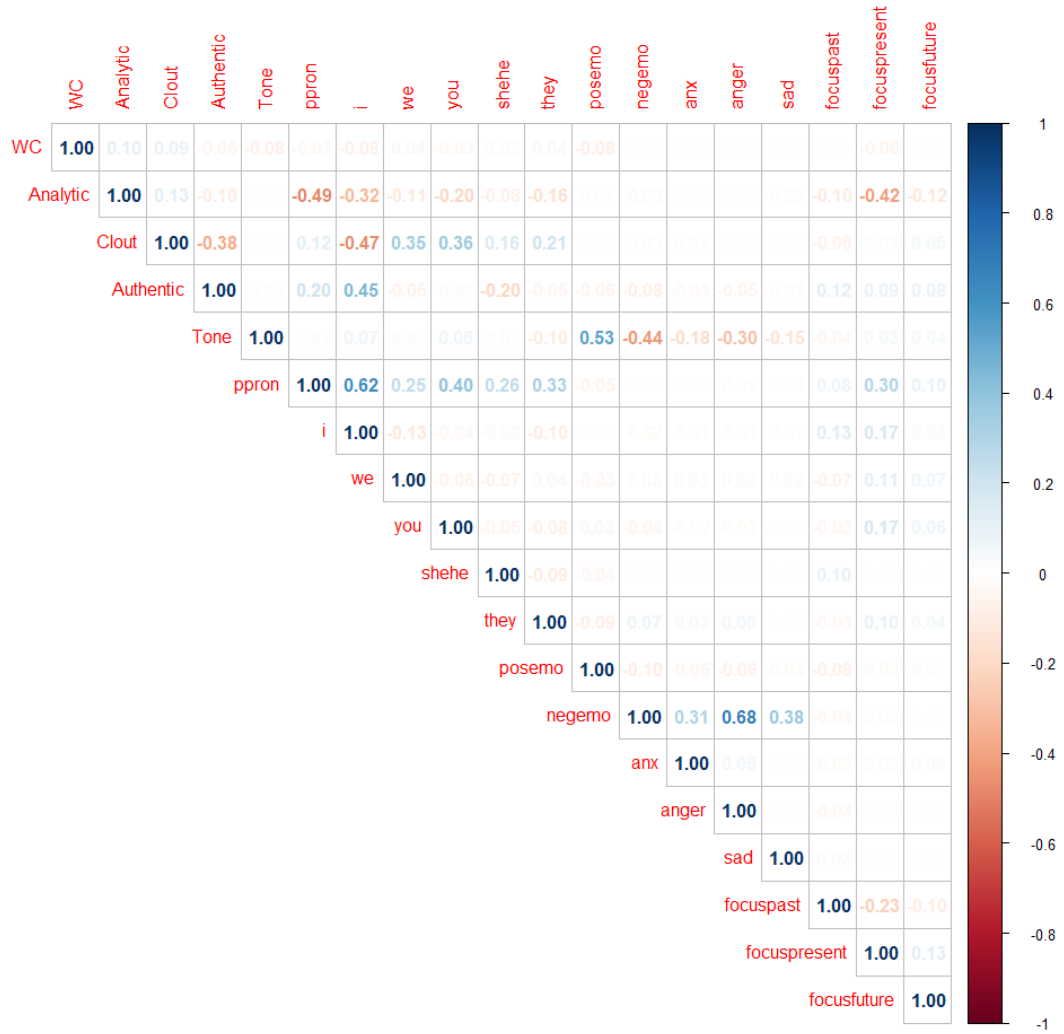
For "focuspresent", theres no obvious trend over time, the monthly average values are mostly between 10-12,except for one month in 2003,it went down to 6.0.

For "negemo", theres no obvious trend over time, the monthly average values are mostly between 1.5-3.0,except for 1 month in 2004 and 1 month in 2008,the values went above 3.0.We can tell that the participants of the forum were expressing their negative emotions during that time.

For "posemo", theres no obvious trend over time, the monthly average values are mostly between 3.0 to 5.0,except for 1 month in 2003 which went above 6.0.

For "sad", theres no obvious trend over time, the monthly average values are mostly between 0.125 and 0.50,except for 2 months in 2003,the values went above 0.75 and for 1 month in 2010,it went up to 1.0.We can tell that participants of the forum were expressing their sad feeling, maybe some bad incidents happened.

Question(a)(2) part 2



In order to know if there is a relationship between linguistic variables over the longer term, I plotted a correlogram with the correlation and corrplot package, the darkness of the blue colour increases with the increasing strength of the positive correlation, darkness of the red colour increases with the increasing strength of the negative correlation. Furthermore, the correlations with p-value of > 0.05 is considered as insignificant so the correlation coefficient values are left empty as we do not have enough evidence to reject the null hypothesis that the correlation coefficient values are equal to 0.

Strong relationship : 0.6 < correlation coefficient values < 0.8

Moderate relationship : 0.4 < correlation coefficient values < 0.6

Weak relationship : 0.2 < correlation coefficient values < 0.4

From the correlogram,we can see that negemo and anger are strongly positive correlated.It is logical because anger is a negative feeling,so when participants of the forum are expressing their anger,they are expressing their negative emotions.The other strong positive correlation is between “ppron and “i”, ppron is the indication of appearance of “I, we, you” words while ”i” is the indication of “I, me, mine” words. The “I” word overlaps for both linguistic variables, so it is logical that ppron and i are strongly positive correlated.

Besides,we can see that Tone and posemo are moderately positive correlated whereas Tone and negemo are moderately negative correlated. It shows that when the forum authors were expressing positive emotions,they tend to have strong emotional tone.However,when the forum authors were expressing negative emotions,they tend to have weak emotional tone.

In addition,we can see that Analytic and ppron are moderately negative correlated,it indicates that when the posts are about analytic thinking,the appearance of “I, me, mine” words will be lesser. we can see that Analytic and focuspresent are moderately negative correlated,it indicates that when the posts are about analytic thinking, the expressing a focus on the present from the authors will be lesser.

Next,”Clout” and “i” are moderately negative correlated,it indicates that when the posts are about power,force and impact,the number of appearance of “I, me, mine” words will be lesser.

Some other weakly positive correlated linguistic variables are “Clout” and “we”, ”Clout” and “they”, “Authentic” and “ppron”,“ppron” and “they”,”ppron” and “focuspresent”,”ppron” and “shehe”,”ppron” and “we”,”negemo” and “anx”,”negemo” and “sad”.

On the contrary,some other weakly negative correlated linguisetic variables are “Analytic” and “I”,”Tone” and “anger”,”focuspast” and “focuspresent”.

There are two strong positive correlations, in order to prove that they are correlated,I performed two hypothesis testing.

****

Hypothesis testing for negemo and anger:

Null hypothesis: The true correlation coefficient value for negemo and anger is 0.

Alternative hypothesis: The true correlation coefficient value for negemo and anger is greater than 0.

p-value : 0

P-value is smaller than the significance level of 0.05,so null hypothesis is rejected,it is highly likely that the the true correlation coefficient value for negemo and anger is greater than 0.

****

Hypothesis testing for ppron and i:

Null hypothesis: The true correlation coefficient value for ppron and i is 0.

Alternative hypothesis: The true correlation coefficient value for negemo and anger is greater than 0.

p-value : 0

P-value is smaller than the significance level of 0.05,so null hypothesis is rejected,it is highly likely that the the true correlation coefficient value for ppron and i is greater than 0.

****

Question(b)

The relevant linguistic variables I used in order to see whether or not particular threads are happier or more optimistic than other threads are “posemo”, ”negemo”, ”anger” and “sad”. It is because in order to know whether a particular thread is happier or more optimistic, the “posemo” should be high,”negemo”, ”anger” and “sad” should be 0 as they are negative feeling.

I analysed it with 2 different approaches. For the first approach, I analysed the data in a more general way, I found out the particular threads which are happier and more optimistic than other threads over the entire period from 2002 to 2012. Firstlty, I filtered out all the posts which has less than 100 words as they are not so significant. Then, I aggregate the data with the same threadID and get the mean of their corresponding relevant linguistic variables. Next, I did a summary on the relevant linguistic variables.

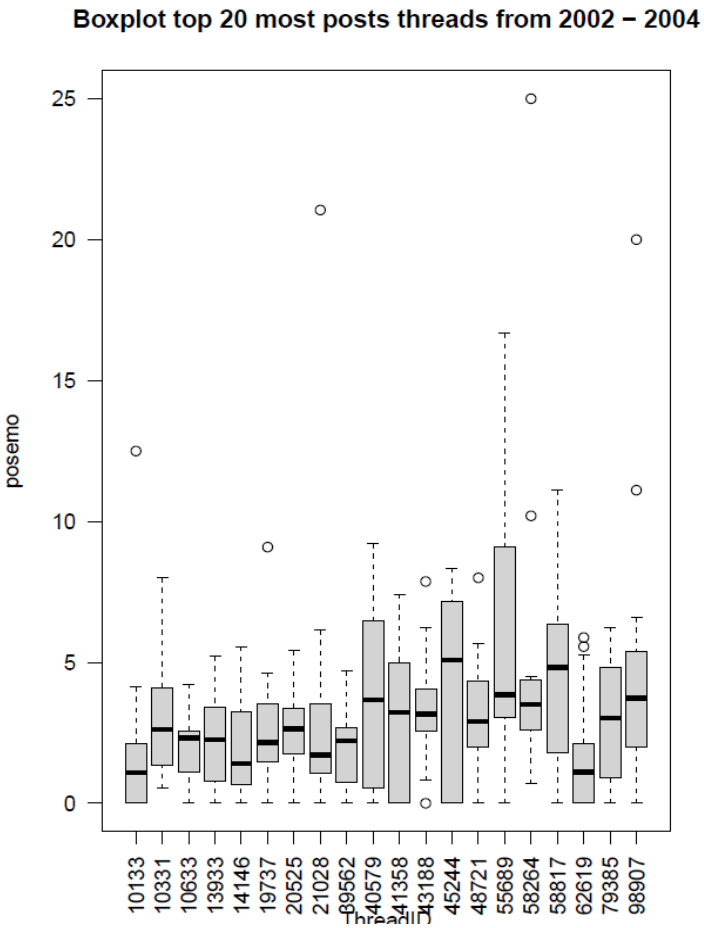
```
     posemo           negemo          anger            sad        
 Min.   : 0.000   Min.   :0.000   Min.   :0.0000   Min.   :0.0000  
 1st Qu.: 1.707   1st Qu.:1.200   1st Qu.:0.2800   1st Qu.:0.0000  
 Median : 2.470   Median :1.980   Median :0.7200   Median :0.1380  
 Mean   : 2.606   Mean   :2.156   Mean   :0.9136   Mean   :0.2866  
 3rd Qu.: 3.310   3rd Qu.:2.868   3rd Qu.:1.3007   3rd Qu.:0.4504  
 Max.   :11.430   Max.   :8.650   Max.   :6.0000   Max.   :4.2700  
```

Next, I filtered out the threads their negemo, anger and sad equal to 0.Then I sorted the threads with their posemo values in descending order. Therefore,I found the top 10 threads which are happier and more optimistic than other threads from year 2002-2012.The top 10 threads are shown below.

ThreadID	posemo	negemo	anger	sad
515618	8.740	0	0	0
52165	8.330	0	0	0
829409	7.480	0	0	0
457980	6.360	0	0	0
325895	6.000	0	0	0
364011	5.825	0	0	0
302700	5.740	0	0	0
242306	5.690	0	0	0
782807	5.630	0	0	0
398879	5.470	0	0	0

For the second approach, I filtered out the top 20 threads with most posts posted from 2002-2004, 2004-2006, 2006-2008, 2008-2010 and 2010-2012. Then I plotted a boxplot for each of the interval to compare their posemo values. The reason for choosing only top 20 threads with most posts posted is analysing the threads with more posts is more likely to be more accurate and we can avoid the anomalies. The reason for comparing their posemo values is because the threads with high posemo values means that the authors were expressing positive emotions hence the threads should be happier and more optimistic and vice versa.

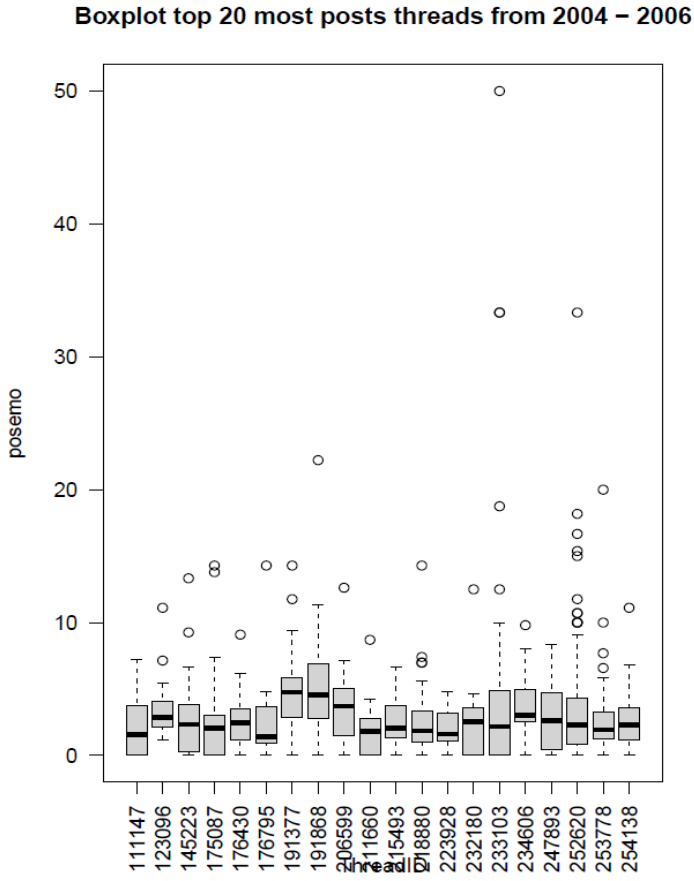
The boxplots plotted are shown below.



From the boxplot,we can see that the average of posmo of each thread is range between 0-5,and we can see that thread 45244 and 58817 have the relatively higher average posemo,which is very close to 5.0.

```
     posemo
Min.    :  0.000
1st Qu.:  1.025
Median :  2.630
Mean   :  3.813
3rd Qu.:  4.415
Max.   :100.000
```

The mean of the posemo of the top 20 most posts posted thread from 2002-2004 is 3.81.

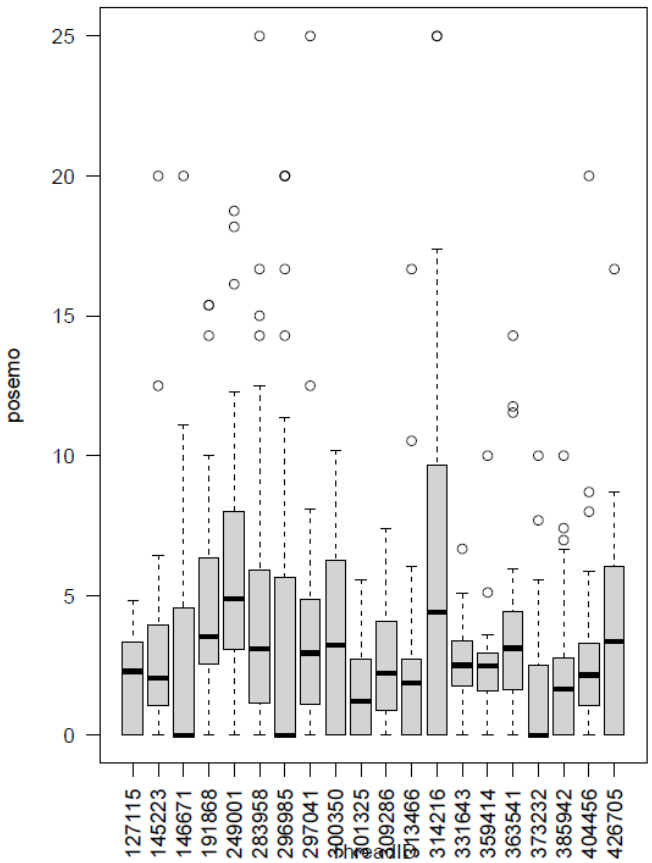


From the boxplot plotted,we can see that the average of posmo of each thread is range between 0-5,the threads with ThreadID 191377 and ThreadID 191868 have the relatively high posemo which are close to 5.0.

```
     posemo
Min.    :  0.000
1st Qu.:  0.390
Median :  2.380
Mean   :  3.532
3rd Qu.:  4.350
Max.   :100.000
```

The mean of the posemo of the top 20 thread from 2004-2006 is 3.53.

Boxplot top 20 most posts threads from 2006 – 2008

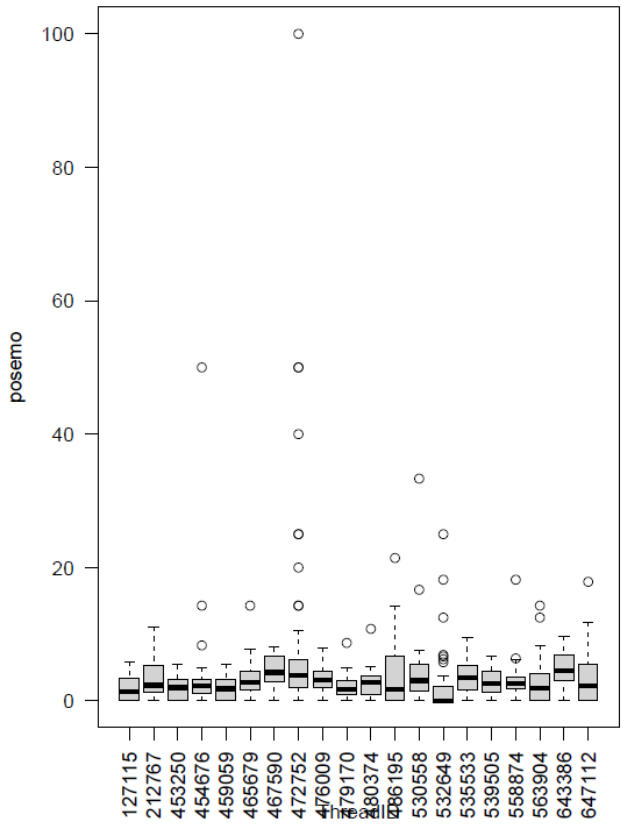


From the boxplot plotted,we can see that the average of posmo of each thread is range between 0-5,and the threads with threadID 249001 and threadID 314216 have the relatively high average posemo which are close to 5.0.

```
     posemo
Min.    :  0.000
1st Qu.:  0.000
Median :  2.440
Mean   :  3.774
3rd Qu.:  4.550
Max.   :100.000
```

The mean of the posemo of the top 20 most posts posted thread from 2006-2008 is 3.77.

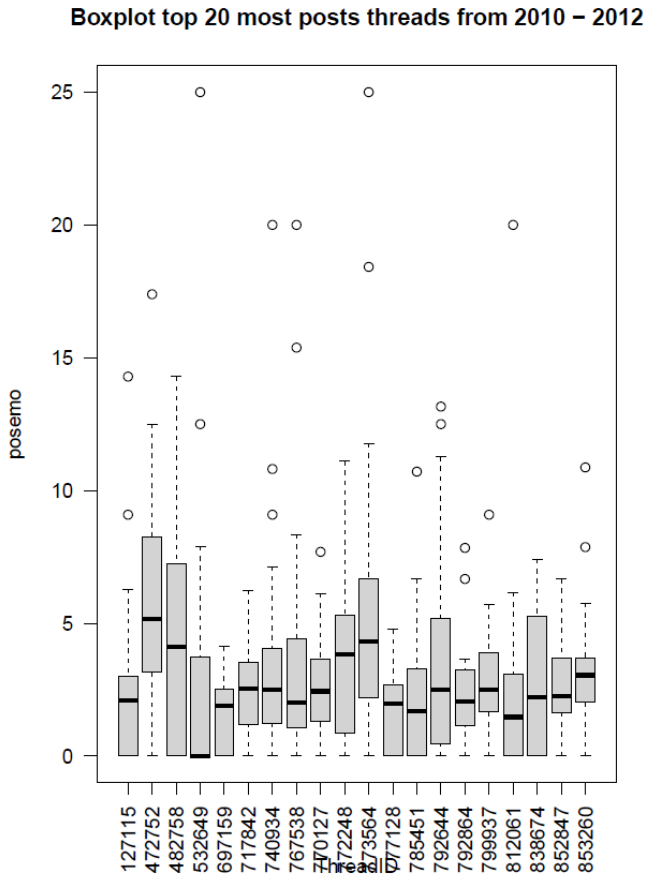
Boxplot top 20 most posts threads from 2008 – 2010



From the boxplot plotted,we can see that the average of posmo of each thread is range between 0-5.The threads with ThreadID 467590, 472752 and 643386 have relatively higher average posemo which is close to 5.0.

```
     posemo
Min.    :  0.0000
1st Qu.:  0.7025
Median :  2.4400
Mean   :  3.5243
3rd Qu.:  4.3500
Max.   :100.0000
```

The mean of the posemo of the top 20 most posts posted thread from 2008-2010 is 3.52



From the boxplot plotted,we can see that the average of posmo of each thread is range between 0-5 except the thread with 472752 is slightly greater than 5.The threads with ThreadID 482758,772248 and 773564 have relatively high average posemo which are close to 5.0.

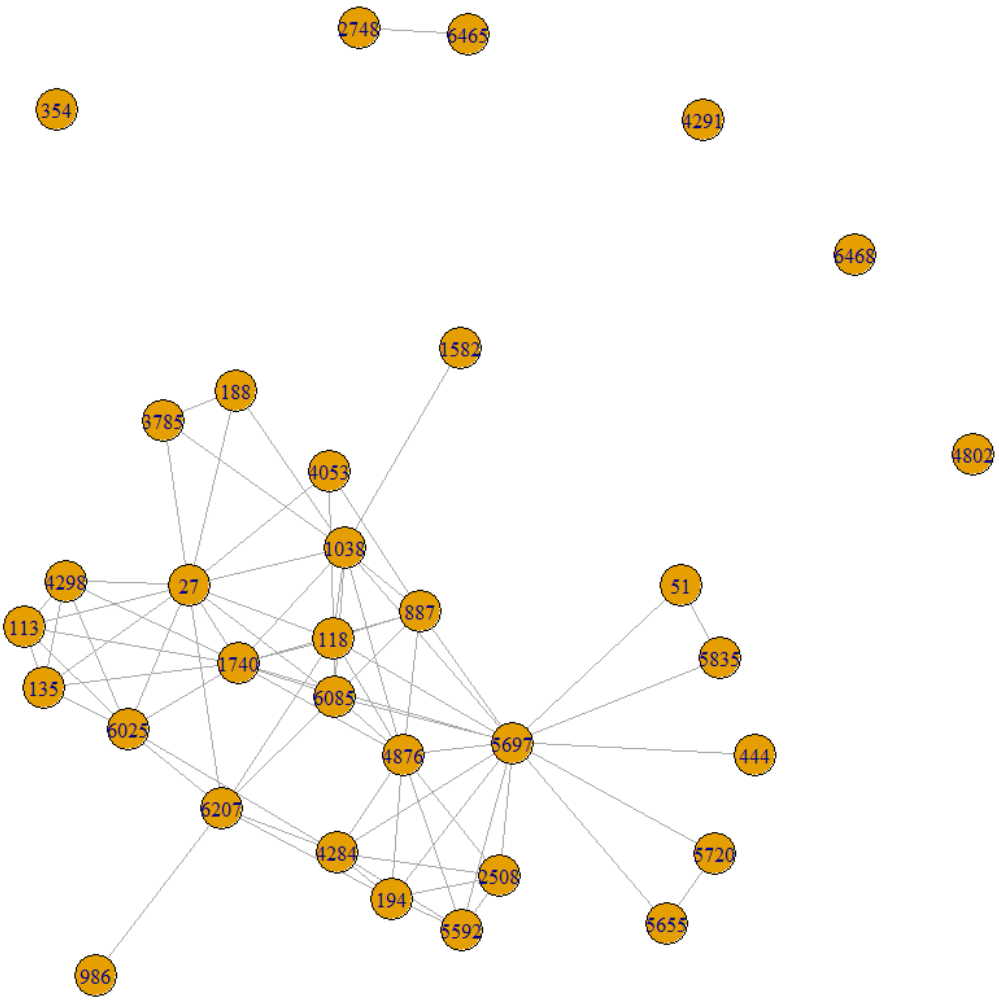
```
     posemo
Min.    :  0.000
1st Qu.:  0.605
Median :  2.360
Mean   :  3.458
3rd Qu.:  4.280
Max.   :100.000
```

The mean of the posemo of the top 20 most posts posted thread from 2010-2012 is 3.46.

In conclusion, the mean of the posemo of the top 20 most posts posted threads for each of the two years interval from 2002-2012 range between 3.46-3.81, the highest average posemo is from 2002-2004 which is 3.81,so we can conclude that in 2002-2004,the top 20 most posts posted threads are relatively happier and more optimistic.The lowest average posemo is from 2010-2012,which is 3.46 hence we can conclude that in 2010-2012,the top 20 most posts posted threads are relatively less happier and less optimistic

Question (c)(1)

I have created a non-trivial social network of all authors who are posting over a month in march 2002 which has a total of 80 posts posted on that month. The network graph created is shown below.



There are 33 authors(nodes) in total. We can see that the authors with AuthorID 354,4291,6468 and 4802 are isolated.

Question (c)(2)

In order to find the most important author in the network, I have to calculate and analyse the degree and centrality of each node(author). After getting the values for degree and centrality of each node, I selected the node with the highest Degree centrality, Betweenness centrality, Closeness centrality and Eigenvector centrality. It is because the author with higher centrality means that he/she is closer to the network hence he/she may be more powerful and more influential in the social network.

The node selected is 5697, so the most important author in the social network I created is the author with AuthorID 5697.

	degree	betweenness	closeness	eig
5697	15	132.279654	0.0042372881	1.00000000
1038	10	52.646970	0.0041493776	0.87168051
1740	11	43.680303	0.0041666667	0.99771413
27	12	38.552381	0.0040816327	0.85030911
6207	7	32.349351	0.0040160643	0.53721557

Picture above is showing the centrality of top 5 nodes sorted by betweenness centrality in descending order.

Author 5697 is selected as the most important author as node 5697 has the greatest degree centrality (15) so it has the highest node connectivity. Author 5697 has connection to the most number of authors in the network hence he/she has connection with wider network.

Secondly, node 5697 has the greatest betweenness centrality, which is 132.28, so it indicates that node 5697 has the greatest number of times which lies on the shortest path between other nodes. Therefore, Author 5697 is likely to be the most powerful influencer who influences the flow around the system.

Furthermore, node 5697 has the greatest closeness centrality, which is 0.0042, means that node 5697 has close relationships with many nodes.It also indicates that node 5697 has the greatest number of shortest path between all the nodes. Therefore, Author 5697 is likely to be best placed to influence the whole social network in the shortest time.

Lastly, node 5697 has the greatest Eigenvector centrality, which is 1.0, means that node 5697 is a node which is connected to the most number of nodes who themselves have high scores and well connected. Therefore, Author 5697 is likely to be the most influential Author over the entire social network.

Question c(2)

To observe any difference between the language used by author 5697 and other authors, I have calculated the mean of the values of linguistic variables of all the posts posted by author 5697 and other authors. I grouped up all the authors other than the top author(5697) as “OTHERS” so that I can compare the top author with other authors in the network. I have also filtered out a few linguistic variables like “i”, ”ppron” as I am not interested at comparing them as they are just personal pronouns so I think they are less relevant to compare with.

The picture below shows the top author and other authors in the social network with their corresponding mean linguistic variables

	Group.1	WC	Analytic	Clout	Authentic	Tone	posemo	negemo	anx	anger	sad	focuspast	focuspresent	focusfuture
1	5697	559.1111	78.20889	60.30667	23.84333	39.11556	2.634444	1.953333	0.2877778	0.7655556	0.3888889	3.638889	7.52000	0.7911111
2	OTHERS	109.8732	64.60366	65.11225	31.69183	37.00704	3.034366	2.654789	0.2070423	1.7594366	0.1866197	3.032394	10.25901	0.8678873

From the comparison above, we can see that author 5697 has higher Analytic than others. Next, we can see that author 5697 has lower anger and lower focuspresent than other authors. In order to have confirmation on the statements above, I have performed several hypothesis testing.

1)

Null hypothesis: True mean of Analytic(5697) – True mean of Analytic(OTHERS) = 0

Alternative hypothesis: True mean of Analytic(5697) – True mean of Analytic(OTHERS) > 0

t = 2.43, p-value = 0.014

p-value is smaller than the significance level of 0.05, so we have sufficient evidence to reject the null hypothesis, so we can conclude that the true mean of Analytic of the posts posted by author 5697 is greater than the true mean of word count of the posts posted by others.

2)

Null hypothesis: True mean of anger(5697) – True mean of anger(OTHERS) = 0

Alternative hypothesis: True mean of anger(5697) – True mean of anger(OTHERS) < 0

t = -2.53, p-value = 0.0069

p-value is smaller than the significance level of 0.05, so we have sufficient evidence to reject the null hypothesis, so we can conclude that the true mean of anger of the posts posted by author 5697 is lesser than the true mean of anger of the posts posted by others.

3)

Null hypothesis: True mean of focuspresent (5697) – True mean of focuspresent (OTHERS) = 0

Alternative hypothesis: True mean of focuspresent(5697) – True mean of focuspresent (OTHERS) < 0

t = -2.53, p-value = 0.0069

p-value is smaller than the significance level of 0.05, so we have sufficient evidence to reject the null hypothesis, so we can conclude that the true mean of focuspresent of the posts posted by author 5697 is lesser than the true mean of focuspresent of the posts posted by others.

In conclusion, we can conclude that in average, the posts posted by author 5697 is more analytical thinking, less indicating anger and less expressing a focus in the present than the other authors.

APPENDIX

```
rm(list = ls())

set.seed(32156944)

webforum <- read.csv("webforum.csv") # Read the data

webforum <- webforum [sample(nrow(webforum), 20000), ] # 20000 rows


# Library used,there's another igraph library will be attached in question(c) later,if I attach it here,

# there will be some errors in Question(a)(1).


library(dplyr)

library(ggplot2)

library(tidyverse)

library(correlation)

library(corrplot)


#Question (a)(1)

#Need to detach the igraph package or it may have errors.

#Convert the dates in Date column to date objects with correct format.

webforum$Date <- as.Date(webforum$Date, format = "%Y-%m-%d")


#Sort the data by date in ascending order.

webforum <- webforum[order(as.Date(webforum$Date, format="%Y-%m-%d")),]


#To get the frequency of posts posted in every month of each year,I have to get the frequency of
every month of each year.

tab <- table(cut(webforum$Date, 'month'))

tab2 <- table(cut(webforum$Date, 'year'))


#Then,I make it into a data frame recording the date and its corresponding frequency.

date_frequency <- data.frame(Date=format(as.Date(names(tab)),
'%m-%Y'),Frequency=as.vector(tab))

time.series2 <- data.frame(Date=format(as.Date(names(tab2)), '%Y'),Frequency=as.vector(tab2))
```

```
#Create the time_series object

monthly_webforum_ts <- ts(date_frequency$Frequency, start =c(2002,1), freq = 12)


#Decompose the time series into seasonal, trend and irregular components

monthly_webforum_decomposed <- decompose(monthly_webforum_ts)


#Plot the decomposed time series

plot(monthly_webforum_decomposed)


#Question(a)(2), to observe the levels of these linguistic variables over the duration of the forum


#I split the webforum into webforum 2 and webforum 3,each of those contain Date and different linguistic variables.

webforum2 <- webforum[,c(3,5:14)]

webforum3 <- webforum[,c(3,15:23)]


#Convert the date to year-month format and add the day of 01 to each date so that it can be converted to Date object in R

webforum2$Date <- format(as.Date(webforum2$Date), "%Y-%m")

webforum2$Date <- as.Date(paste(webforum2$Date,"-01",sep=""))


#Convert the date to year-month format and add the day of 01 to each date so that it can be converted to Date object in R

webforum3$Date <- format(as.Date(webforum3$Date), "%Y-%m")

webforum3$Date <- as.Date(paste(webforum3$Date,"-01",sep=""))


#Then,for all the posts posted on each year-month,I group them up by using the group_by(Date) function and used

#summarise_all(.funs = mean) to get the mean value of all their linguistic variables on that particular year-month.

average_webforum2 <- webforum2 %>% group_by(Date) %>% summarise_all(.funs = mean)

average_webforum3 <- webforum3 %>% group_by(Date) %>% summarise_all(.funs = mean)
```

```
#Then,I reformat the data from "wide" to "long" format and plot each of the linguistic variable over years with facet_wrap.

formatted <- average_webforum2 %>% pivot_longer(2:11, names_to = "linguistic_variable", values_to = "values")

ggplot(formatted, aes(Date, values)) + geom_line() + facet_wrap(~linguistic_variable, scales = "free_y", ncol = 1)


#Then,I reformat the data from "wide" to "long" format and plot each of the linguistic variable over years with facet_wrap.

formatted2 <- average_webforum3 %>% pivot_longer(2:10, names_to = "linguistic_variable", values_to = "values")

ggplot(formatted2, aes(Date, values)) + geom_line() + facet_wrap(~linguistic_variable, scales = "free_y", ncol = 1)


clout_analytic <- formatted[formatted$linguistic_variable %in% c("Clout","Analytic"),]

t.test(values ~ linguistic_variable,data = clout_analytic)


#Question(a)(2) Part 2

#To filter out all the linguistic variables

linguistic_variables <- webforum[,5:23]


#correlation is the function from correlation package

#Performs a correlation analysis on linguistic variables,it will also output the p-values for each correlation coefficient .

corr <- as.data.frame(correlation::correlation(linguistic_variables,include_factors = TRUE, method = "auto"))


#Filter out the relevant columns

corr <- corr[,c(1,2,3,9)]


#Filter out the correlation coefficient values with p-value < 0.05

corr <- filter(corr,p < 0.05)
```

```
#Find the strongly correlated linguistic variables

strong_relationship <- filter(corr, r >= 0.6 | r <= -0.6)


#Find the moderately correlated linguistic variables

moderate_relationship <- filter(corr, (r >= 0.4 & r < 0.6) | (r <= -0.4 & r > -0.6))


#Find the weakly correlated linguistic variables

weak_relationship <- filter(corr, (r >= 0.2 & r < 0.4) | (r <= -0.2 & r > -0.4))


#Find the very weakly correlated linguistic variables

very_weak_relationship <- filter(corr, (r >= 0.0 & r < 0.2) | (r <= -0.0 & r > -0.2))


#corrplot is the function from corrplot package which is used to plot correlogram.

corr_plot <- corrplot(cor(linguistic_variables),method = "number",type = "upper",sig.level =
0.05,insig = "blank")


#Question(b)

#Filter out the posts with WC >= 100

webforum_filtered_wc <- filter(webforum, WC >= 100)


#Filter out the columns with only relevant linguistic variables

emo <- webforum_filtered_wc[,c(1,16,17,19,20)]


#Aggregate by ThreadID and get the mean of each linguistic variable.

emo <- aggregate(emo,list(emo$ThreadID),mean)

emo <- emo[,2:6]


#Do the summary

summary(emo[,2:5])
```

```
#Filter out the threads with their negomo == 0  & anger == 0 & sad == 0
optimistic <- filter(emo,negemo == 0 & anger == 0 & sad == 0)


#Get the top 10 threads with greatest posemo
optimistic <- optimistic[order(-optimistic$posemo),]
optimistic <- optimistic[1:10,]
#Plot the boxplot of top 20 most posts posted threads with their posemo values from year 2002-2004
year02_04 <- filter(webforum, Date < as.Date("2004-01-01"))
summary(year02_04[,c("posemo","negemo","anger","sad")])
table(year02_04$ThreadID)
tab_threadID <- table(year02_04$ThreadID)
threadID_freq <- data.frame(ThreadID = names(tab_threadID),Frequency=as.vector(tab_threadID))
attach(threadID_freq)
threadID_freq <- threadID_freq[order(-Frequency),]
#Filter out top 20 most posts posted threads
filtered_0204 <- filter(year02_04,year02_04$ThreadID %in% threadID_freq[1:20,1])
boxplot_0204 <- filtered_0204[,c(1,16)]
boxplot(boxplot_0204$posemo ~ boxplot_0204$ThreadID, las = 2, xlab = "ThreadID", ylab = "posemo",
      main = "Boxplot top 20 most posts threads from 2002 - 2004"  )


#Plot the boxplot of top 20 most posts posted threads with their posemo values from year 2004-2006
year04_06 <- filter(webforum, Date > as.Date("2003-12-31") & Date < as.Date("2006-01-01"))
summary(year04_06[,c("posemo","negemo","anger","sad")])
table(year04_06$ThreadID)
tab_threadID <- table(year04_06$ThreadID)
threadID_freq <- data.frame(ThreadID = names(tab_threadID),Frequency=as.vector(tab_threadID))
attach(threadID_freq)
threadID_freq <- threadID_freq[order(-Frequency),]
filtered_year04_06 <- filter(year04_06,year04_06$ThreadID %in% threadID_freq[1:20,1])
```

```
boxplot_0406 <- filtered_year04_06[,c(1,16)]

boxplot(boxplot_0406$posemo ~ boxplot_0406$ThreadID, las = 2, xlab = "ThreadID", ylab =
"posemo", main =  "Boxplot top 20 most posts threads from 2004 - 2006")


#Plot the boxplot of top 20 most posts posted threads with their posemo values from year 2006-
2008

year06_08 <- filter(webforum, Date > as.Date("2005-12-31") & Date < as.Date("2008-01-01"))

summary(year06_08[,c("posemo","negemo","anger","sad")])

table(year06_08$ThreadID)

tab_threadID <- table(year06_08$ThreadID)

threadID_freq <- data.frame(ThreadID = names(tab_threadID),Frequency=as.vector(tab_threadID))

attach(threadID_freq)

threadID_freq <- threadID_freq[order(-Frequency),]

filtered_year06_08 <- filter(year06_08,year06_08$ThreadID %in% threadID_freq[1:20,1])

boxplot_0608 <- filtered_year06_08[,c(1,16)]

boxplot(boxplot_0608$posemo ~ boxplot_0608$ThreadID, las = 2, xlab = "ThreadID", ylab =
"posemo", main =  "Boxplot top 20 most posts threads from 2006 - 2008")


#Plot the boxplot of top 20 most posts posted threads with their posemo values from year 2008-
2010

year08_10 <- filter(webforum, Date > as.Date("2007-12-31") & Date < as.Date("2010-01-01"))

summary(year08_10[,c("posemo","negemo","anger","sad")])

table(year08_10$ThreadID)

tab_threadID <- table(year08_10$ThreadID)

threadID_freq <- data.frame(ThreadID = names(tab_threadID),Frequency=as.vector(tab_threadID))

attach(threadID_freq)

threadID_freq <- threadID_freq[order(-Frequency),]

filtered_year08_10 <- filter(year08_10,year08_10$ThreadID %in% threadID_freq[1:20,1])

boxplot_0810<- filtered_year08_10[,c(1,16)]

boxplot(boxplot_0810$posemo ~ boxplot_0810$ThreadID, las = 2, xlab = "ThreadID", ylab =
"posemo", main =  "Boxplot top 20 most posts threads from 2008 - 2010")
```

```
#Plot the boxplot of top 20 most posts posted threads with their posemo values from year 2010-2012

year10_12 <- filter(webforum, Date > as.Date("2009-12-31"))

summary(year10_12[,c("posemo","negemo","anger","sad")])

table(year10_12$ThreadID)

tab_threadID <- table(year10_12$ThreadID)

threadID_freq <- data.frame(ThreadID = names(tab_threadID),Frequency=as.vector(tab_threadID))

attach(threadID_freq)

threadID_freq <- threadID_freq[order(-Frequency),]

filtered_year10_12 <- filter(year10_12,year10_12$ThreadID %in% threadID_freq[1:20,1])

boxplot_1012 <- filtered_year10_12[,c(1,16)]

boxplot(boxplot_1012$posemo ~ boxplot_1012$ThreadID, las = 2, xlab = "ThreadID", ylab = "posemo", main =

      "Boxplot top 20 most posts threads from 2010 - 2012")


#Question(c)(1)


#Library of igraph imported here or the graph in (a)(1) cannot be plotted

library(igraph)

webforum_month_year <- webforum


#Get all the data from march 2002 only.

webforum_month_year$Date <- format(as.Date(webforum_month_year$Date), "%Y-%m")

webforum_month_year$Date <- as.Date(paste(webforum_month_year$Date,"-01",sep=""))

webforum_march_2002 <- filter(webforum_month_year,Date == as.Date("2002-03-01"))


#Get all the unique AuthorID in march 2002 so that can be used as nodes.

Author <- unique(webforum_march_2002$AuthorID)

Author <- as.data.frame(Author)


#Make an empty graph

social_network <- make_empty_graph(directed = FALSE)
```

```
#Add all the vertices with AuthorID as the name into the graph.
for(i in 1:nrow(Author)){
  social_network <- add_vertices(social_network,1,name = as.character(Author$Author[i]))
}

#Get only ThreadID and AuthorID
webforum_threads_author <- webforum_march_2002[,c(1,2)]

# loop through each thread
for(thread in unique(webforum_threads_author$ThreadID)){
  temp = webforum_threads_author[(webforum_threads_author$ThreadID == thread),]
  if(nrow(temp)>1){
    # Combine each pair of authors to make an edge list
    Edgelist = as.data.frame(t(combn(temp$AuthorID,2)))
    # loop through pairs of edges and add
    for(i in 1:nrow(Edgelist)){
      colnames(Edgelist) = c("Author1","Author2")
      social_network <-
add_edges(social_network,c(as.character(Edgelist$Author1[i]),as.character(Edgelist$Author2[i])))
    }
  }
}

#Simplify to remove duplicated edge
social_network = simplify(social_network)
layout <- layout.fruchterman.reingold(social_network)
#Plot the graph
plot(social_network,layout = layout, vertex.size = 9)
layout <- layout.circle(social_network)
plot(social_network,layout = layout, vertex.size = 9)
```

```
#Question (c)(2)

#get the degree centrality

degree = as.table(degree(social_network))

#get the betweenness centrality

betweenness = as.table(betweenness(social_network))

#get the closeness centrality

closeness = as.table(closeness(social_network))

#get the eigenvector centrality

eig = as.table(evcent(social_network)$vector)

#Combine the values and make them into a dataframe

vertex_analysis <- as.data.frame(rbind(degree, betweenness, closeness, eig))

vertex_analysis <- t(vertex_analysis)

#Sort the vertex by betweenness centrality in descending order

vertex_analysis <- vertex_analysis[order(-betweenness),]

top_author_others <- webforum_march_2002

top_author_others$AuthorID[top_author_others$AuthorID != 5697] <- "OTHERS"

top_author_others_avrg <-
as.data.frame(aggregate(top_author_others,list(top_author_others$AuthorID),mean))

top_author_others_avrg <- top_author_others_avrg[,c(1,6,7,8,9,10,17,18,19,20,21,22,23,24)]

#Hypothesis testing

#Test if True mean of WC(5697) – True mean of WC(OTHERS) > 0

t.test(WC ~ AuthorID,data = top_author_others,alternative = "greater")
```

```
#Test if True mean of Analytic(5697) – True mean of Analytic(OTHERS) > 0
t.test(Analytic ~ AuthorID,data = top_author_others,alternative = "greater")

#Test if True mean of anger(5697) – True mean of anger(OTHERS) < 0
t.test(anger ~ AuthorID,data = top_author_others,alternative = "less")

#Test if True mean of focuspresent(5697) – True mean of focuspresent (OTHERS) < 0
t.test(focuspresent ~ AuthorID,data = top_author_others,alternative = "less")
```