Assignment 1

FIT3152

Rui Qin | 30874157

Introduction

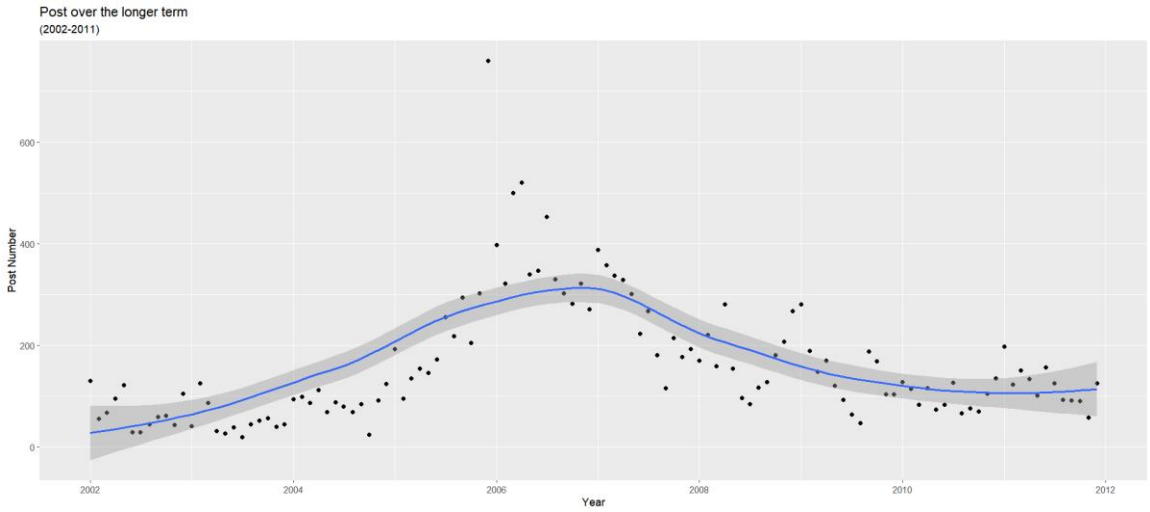
This report is the analysis of big data about linguistic variables on a forum. The data contain thread id and author id with different emotions and words usage analysis in his post, the data is between 2002 to 2011 and it can show the feeling and trends of that time.

The process of analysis is divided into 3 parts. The first is analyzing the active of users in forum and language in general. The second part is analysis of the linguistic variables by thread. There is also an analysis of social network in the end.

Section A

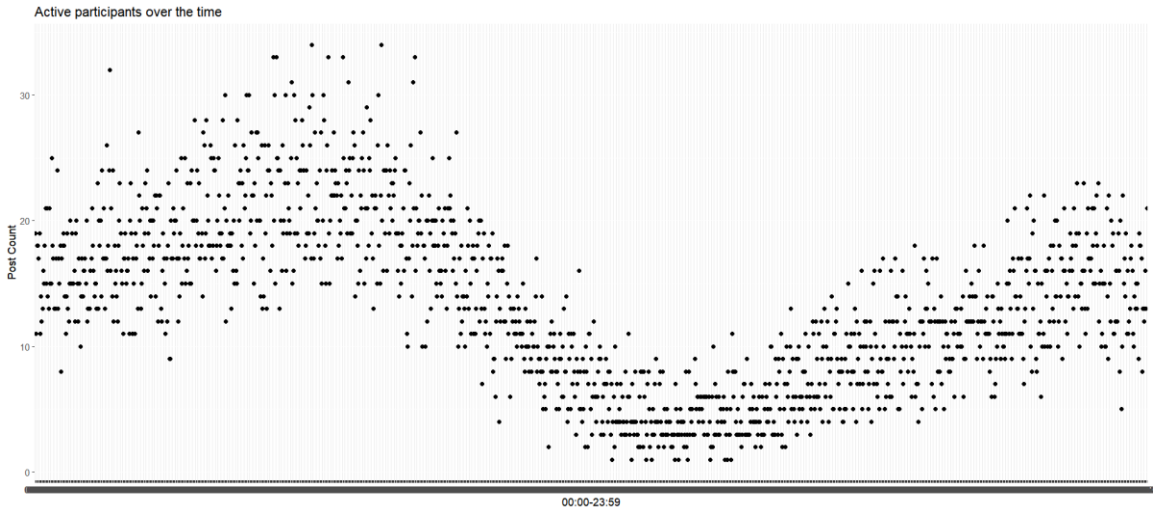
How active are participants over the longer-term How active are participants over the longer term? Are there periods where activity increases or decreases? Is there a trend over time?

Figure 1



In figure 1, we create a diagram that includes points and a line. The points are the number of posts in each month per year, and the line shows how active the user is over time. We can easily notice that the activity trend is different in different periods. the number of threads increase from 2002 to 2007 and reach its peak in 2007, which is about 300 posts. Then the number of users posting decreased each year, but the rate of decline is getting slower. In 2010, the number of users posting almost remained the same, then after 2011, the number slightly raise.

Figure 2



Based on figure 2, we can notice the trend of post-update during the time. The number of post increase during the mid-night, and start dropping and reaching the bottom in the afternoon. Then the trend starts climbing up again till the next day midnight.

Looking at the linguistic variables, do the levels of these change over the duration of the forum?

Figure 3

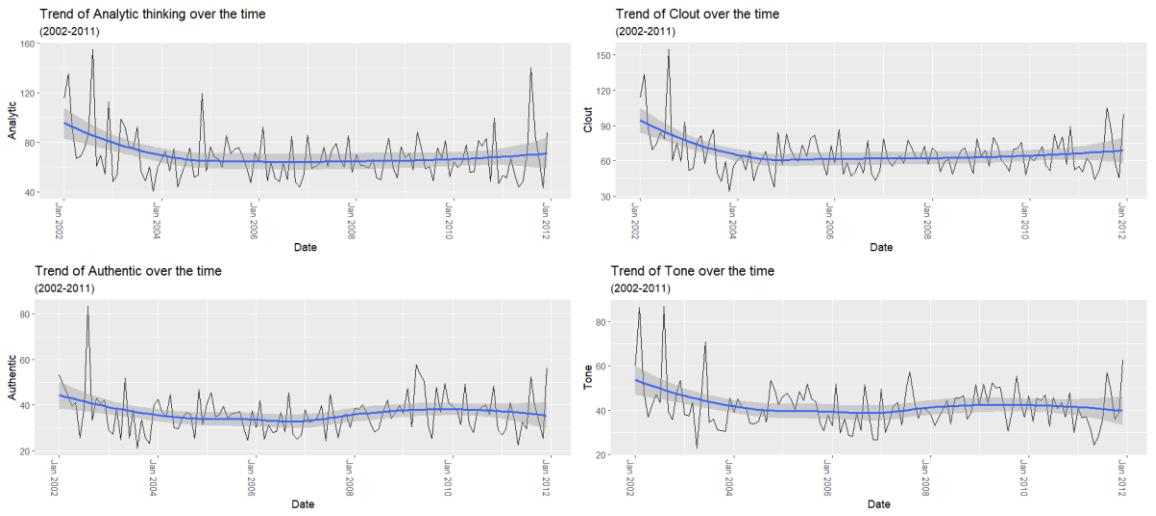


Figure 3 shows the trend of analytic thinking, clout, authentic and tone over time change in the forum. What we find is that the trend all started at the peak and then went down and remained stable until 2004 or 2005. Among them, analytic thinking and Clout have slightly increased after 2011, while Authentic and Tone have slightly increased after 2008 and decreased after 2010. The fluctuation intensity of their data is not strong, and the fluctuation of tone is extremely slight.

Figure 4

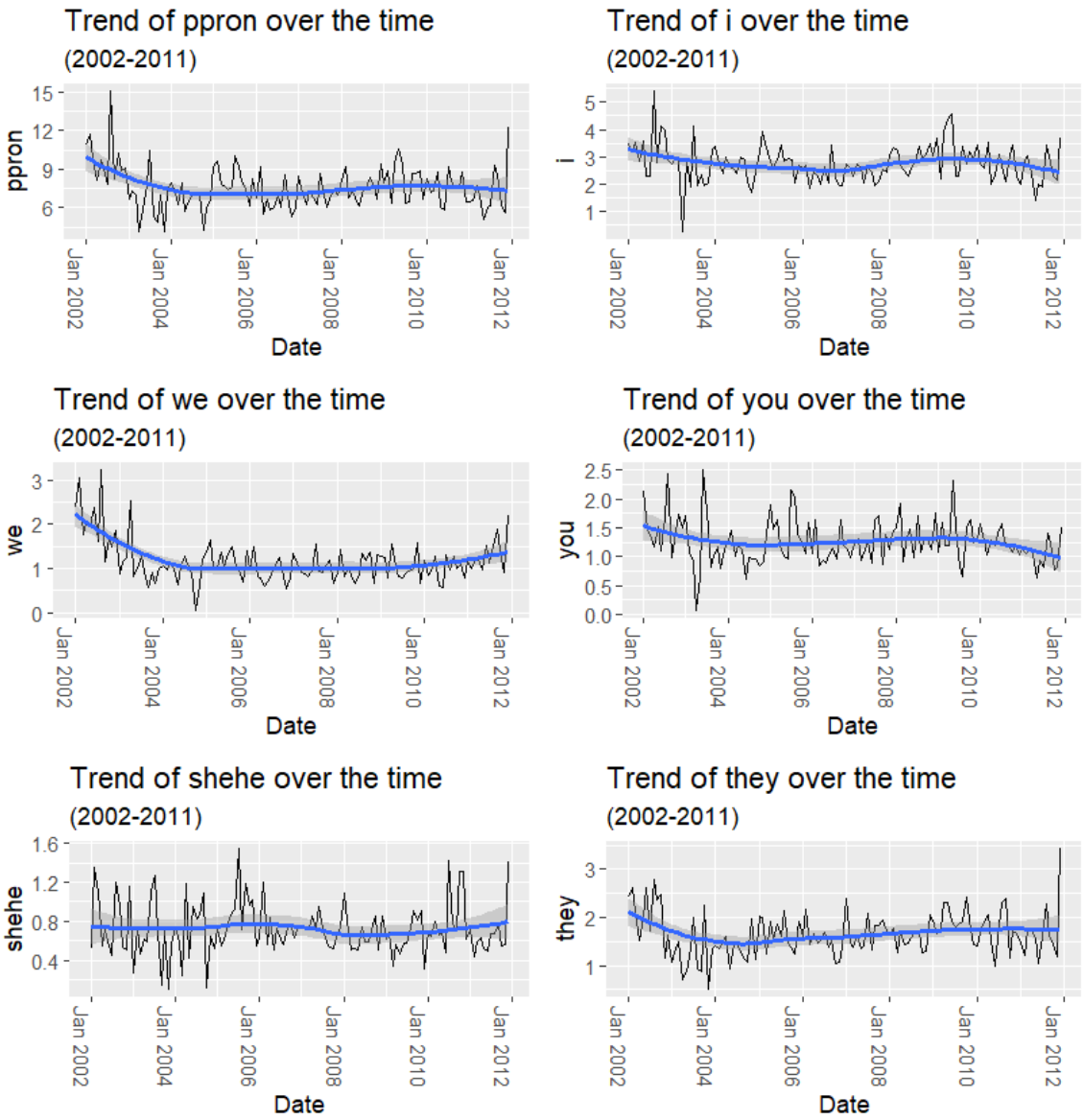


Figure 4 shows the trend in personal pronouns. The first diagram about ppron is the trend of total personal pronouns, the personal pronouns usage peak in the beginning, then dropped to the bottom in 2004 and remained the same till 2007, after 2007 the usage slightly up and down. The usage of shehe is different, it did not change too much during the time change.

Figure 5

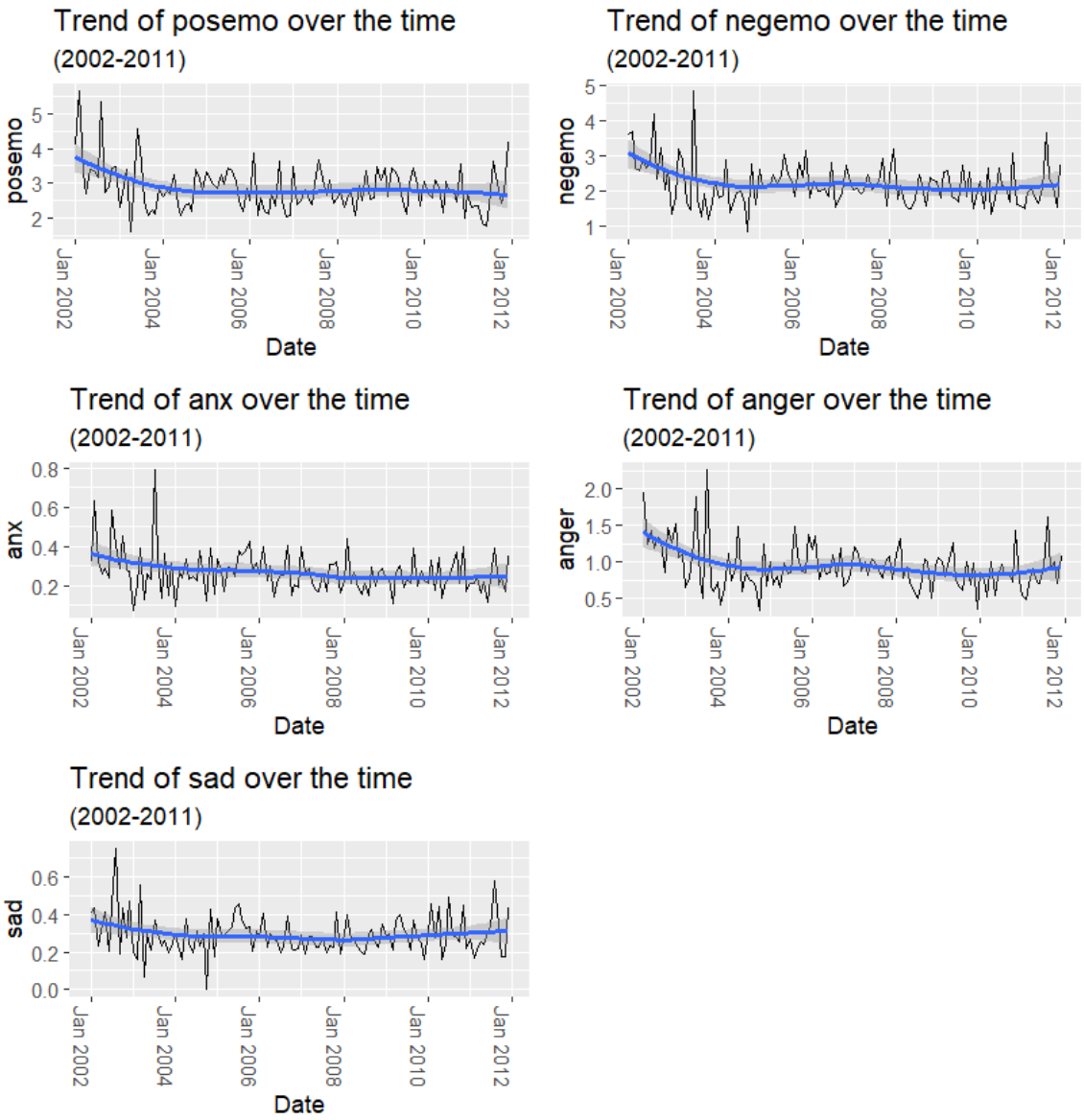


Figure 5 is about the emotional aspect, we can see that sadness and anxiety always stay around 0.2-0.4, and anxiety started around 0.4 and dropped around 0.2 in 2011. Positive emotion dropped at its peak in 2002 and remained the same from 2004 to 2011. Negative emotion and anger Both had minor ups and downs in 2008.

Figure 6

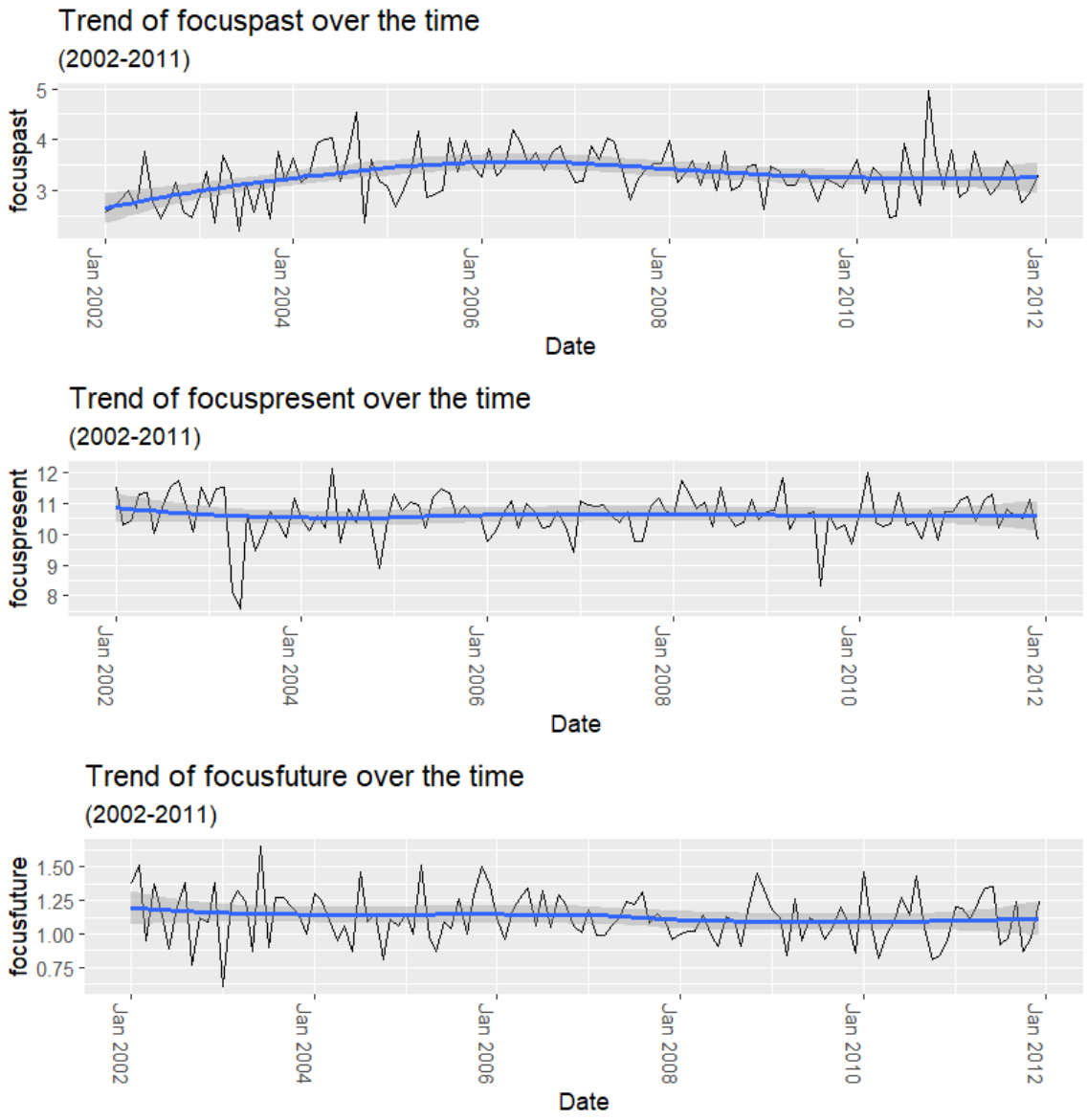


Figure 6 shows the trend of users talking about the future, present and past. The trend of focus past started low in the beginning and peak in 2006, the slightly low down. The other almost remained the same during the time change.

Conclusion of this part

Most linguistic variables usage showed a decreasing trend in the beginning except she/he, focus past, focus present and focus future. They all stop decreasing in the middle of 2004.

Is there a relationship between linguistic variables over the longer term?

Figure 7

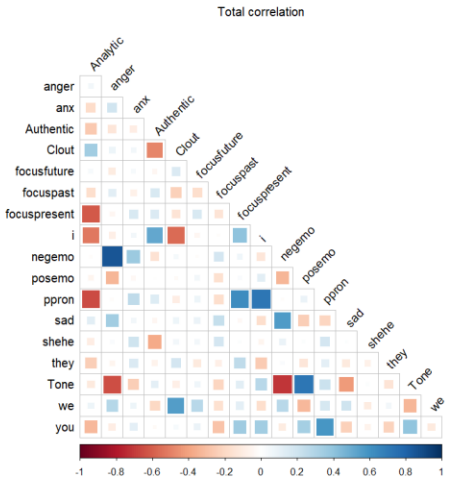


Figure 7 is the correlation all the time, we can see that anger - negative emotion has a strong correlation which is almost 1, and I – personal pronouns, positive emotion-tone have 0.8 strong relations. There is also a 0.7-0.6 high correlation between sad and negative emotion, personal pronouns and you, focus present and personal pronouns, clout and we, I and authentic.

Figure 8

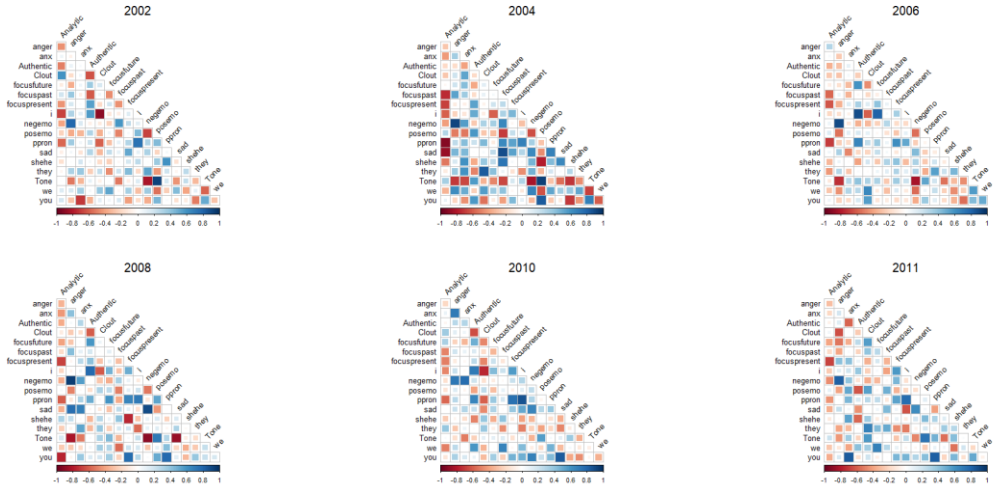


Figure 8 is the correlation changing with time, we can notice that in 2004 the positive correlation went strong and the negative correlation went weak.

Section B

Using the relevant linguistic variables, is it possible to see whether or not particular threads are happier or more optimistic than other threads, or the forum in general, at different periods in time

Figure 9

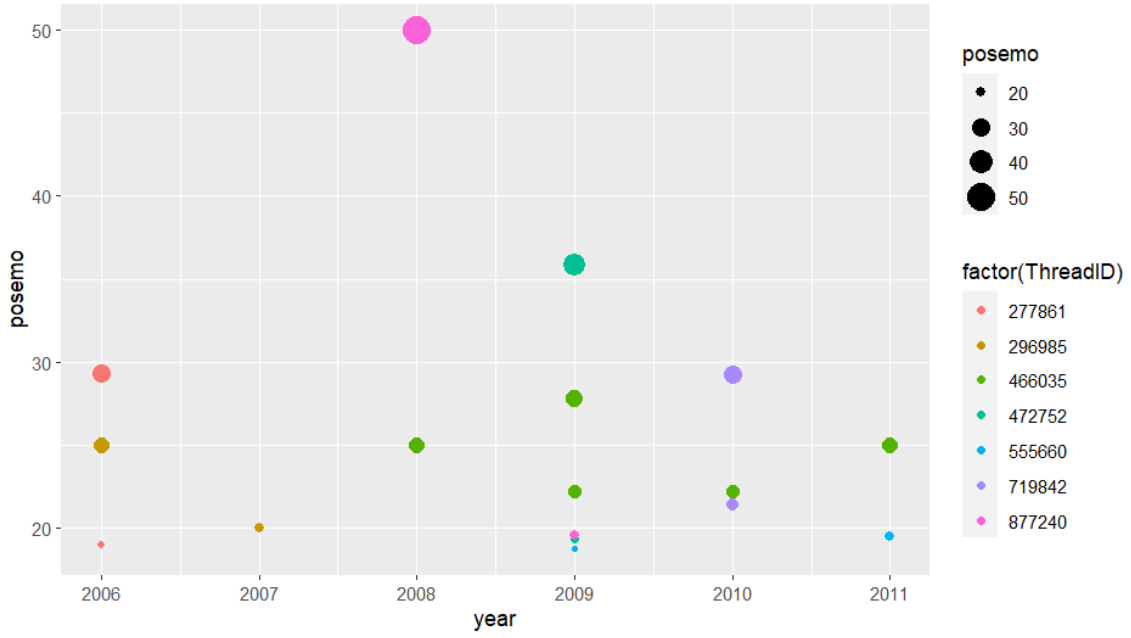
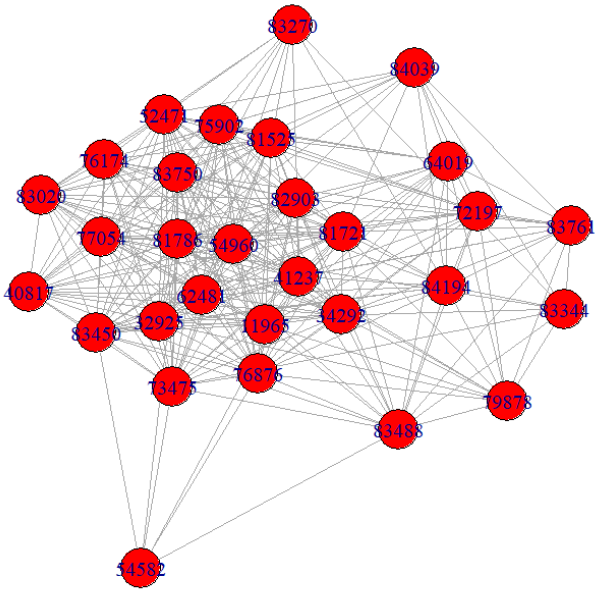


Figure 9 shows the threads which have higher positive emotion than average during the period, and they were all active for more than 1 year. The normal average number of threads is 9.92, after 3 round filtering, the lowest average positive emotion index in figure 9 is 20.

Section C

Create a non-trivial social network of all authors who are posting over a particular time period

Figure 10



This network is based on the month that had the largest number of postings: 2015-12. After being deleted the author only posted once time and the thread only showed once time, there are 589 histories of posts. In this figure, they are the top 30 highest active authors this month, and they are connected if they showed up in the same thread. There are about 293 edges in this network, and author ID 34292 has the most-posted 22 times.

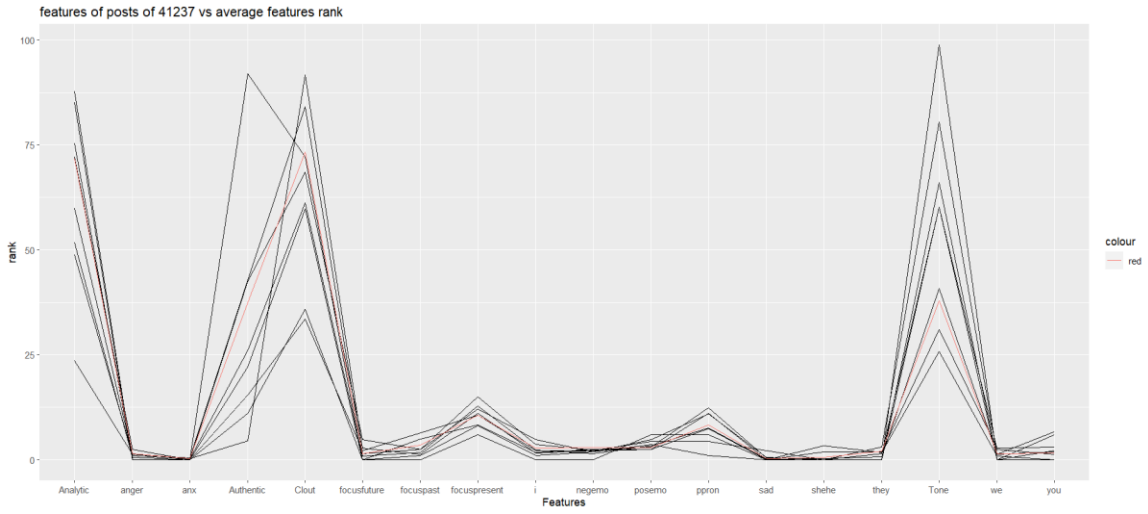
Identify the most important author in the social network you created. Looking at the language they use, can you observe any difference between them and other members of their social network?

Figure 11

```
> closeness(g)
     11965      32925      34292      40817      41237      52471      54582      54960      62481      64019
0.02857143 0.02777778 0.03125000 0.02564103 0.03225806 0.02857143 0.01923077 0.02777778 0.02777778 0.02439024
     72197      73475      75902      76174      76876      77054      79878      81525      81721      81786
0.02380952 0.02777778 0.02857143 0.02702703 0.03030303 0.02702703 0.02222222 0.02857143 0.03125000 0.03030303
     82903      83020      83270      83344      83450      83488      83750      83761      84039      84194
0.02857143 0.02631579 0.02083333 0.02040816 0.02777778 0.02380952 0.02702703 0.02222222 0.02222222 0.02439024
> # Compare individual vertex based on betweenness centrality
> betweenness(g)
     11965      32925      34292      40817      41237      52471      54582      54960      62481      64019      72197
 4.186167   3.832848  10.480114   0.000000  14.324559   4.710961   0.000000   3.126066   4.271725   2.977328   2.724081
     73475      75902      76174      76876      77054      79878      81525      81721      81786      82903      83020
 7.743434   4.710961   2.171701   9.082459   4.694444   2.262648   4.710961   8.341237  11.030586   4.439230   1.377778
     83270      83344      83450      83488      83750      83761      84039      84194
 1.801440   2.581757   5.123060   5.832562   2.923063   2.754326   3.323082   6.461419
>
> eigen_centrality(g)
$vector
     11965      32925      34292      40817      41237      52471      54582      54960      62481      64019      72197
0.9253119  0.8914286  0.9833198  0.8231875  1.0000000  0.9178379  0.2480477  0.9001317  0.8866451  0.6399043  0.5980996
     73475      75902      76174      76876      77054      79878      81525      81721      81786      82903      83020
0.8716083  0.9178379  0.8705357  0.9627514  0.8526575  0.4614593  0.9178379  0.9924489  0.9542486  0.9233375  0.8406473
     83270      83344      83450      83488      83750      83761      84039      84194
0.3856347  0.2652676  0.8858072  0.5561528  0.8641154  0.4370728  0.4669009  0.5788779
```

Based on figure 11, we calculate the closeness centrality, betweenness centrality and eigen centrality of each node in the network. We can easily notice that author 41237 has the highest numbers, which means he is the most important author during this month.

Figure 12



According to figure 12, the negative emotion of his posts is slightly lower than the average level, and his positive emotion is a little bit higher than average. There is a large part of posts have a very high tone.

Appendix

A1

```
library(dplyr)

library(ggplot2)

library(tidyverse)

library(lubridate)

library(zoo)

library(gridExtra)

library(grid)

library(lattice)

library(reshape2)

library(corrplot)

library(RColorBrewer)

library(igraph)

library(igraphdata)

library(data.table)

#set up path

#setwd("/Users/mac/My Drive/Documents/Assignment/2-SEM_1/FIT3152/A1 (20%)")

setwd("C:/Users/aud/My Drive/Documents/Assignment/2-SEM_1/FIT3152/A1 (20%)")

#read data

rm(list = ls())

set.seed(30874157)

webforum <- read.csv("webforum.csv")

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

webforum$Year_and_month <- format(as.Date(webforum$Date), "%y-%m")
```

```
str(webforum)

#data tidy

#check if NA value and datatype

webforum<-na.omit(webforum)

webforum<-webforum%>%distinct()

str(webforum)

#turn date from char to date

webforum$Date <- as.Date(webforum$Date)

#make new columns with year, month and day

webforum_with_date<-webforum%>%
  mutate(
    year = year(Date),
    month = month(Date),
    day = day(Date)
  )

# clean Author ID is -1

webforum_with_date <- webforum_with_date[!(webforum_with_date$AuthorID == -1),]

# Clean post which word count is 0

webforum_with_date <- webforum_with_date[!(webforum_with_date$WC == 0),]
```

AQ1

```
#make a data frame with year column and month column

AQ1_dataframe<-webforum_with_date%>%group_by(year,month)%>%summarise(count
=n())

#make a data frame about the year and month

long_term_dataframe<-AQ1_dataframe%>%mutate(date = make_date(year, month))

#graph create
```

```
ggplot(long_term_dataframe,
       aes(x <- date,
           y <- count)) +
  labs(
    title = "Post over the longer term",
    subtitle = "(2002-2011)",
    x = "Year",
    y = "Post Number"
  )+
  geom_point() +
  geom_smooth(method = "loess", formula = y ~ x)


#Time
AQ1_time<-webforum_with_date%>%group_by(Time)%>%summarise(count =n())
ggplot(AQ1_time,
       aes(x <- Time,
           y <- count)) +
  labs(
    title = "Active participants over the time",
    x = "Time",
    y = "Post Count"
  )+
  geom_point() +
  geom_smooth(method = "loess", formula = y ~ x)
```

AQ2

```
#AQ2
```

```
AQ2_dataframe = webforum_with_date%>%group_by(year,month)%>%
  summarise(count=n(),

WC,Analytic,Clout,Authentic,Tone,ppron,i,we,you,shehe,they,posemo,negemo,anx,
          anger,sad,focuspast,focuspresent,focusfuture
  )
AQ2_dataframe[,5:19] <- AQ2_dataframe[,5:19]*(AQ2_dataframe$WC/100)
AQ2_dataframe$Date= as.yearmon(paste(AQ2_dataframe$year, AQ2_dataframe$month),
"%Y %m")
AQ2_dataframe<-AQ2_dataframe%>%group_by(Date)%>%summarise(
  WC = mean(WC,na.rm = TRUE),
  Analytic = mean(Analytic,na.rm = TRUE),
  Clout = mean(Clout,na.rm = TRUE),
  Authentic = mean(Authentic,na.rm = TRUE),
  Tone = mean(Tone, na.rm = TRUE),
  ppron = mean(ppron, na.rm = TRUE),
  i = mean(i,na.rm = TRUE),
  we = mean(we,na.rm = TRUE),
  you = mean(you,na.rm = TRUE),
  shehe = mean(shehe,na.rm = TRUE),
  they = mean(they,na.rm = TRUE),
  posemo = mean(posemo,na.rm = TRUE),
  negemo = mean(negemo,na.rm = TRUE),
  anx = mean(anx,na.rm = TRUE),
  anger = mean(anger,na.rm = TRUE),
  sad = mean(sad,na.rm = TRUE),
```

```
  focuspast = mean(focuspast,na.rm = TRUE),
  focuspresent = mean(focuspresent,na.rm = TRUE),
  focusfuture = mean(focusfuture,na.rm = TRUE),
)



#Analytic
# Calculate the Graph
Analytic_plot=ggplot(
  AQ2_dataframe,
  aes(x=Date,
      y=Analytic)) +
  labs(
    title = "Trend of Analytic thinking over the time",
    subtitle = "(2002-2011)",
    x = "Date",
    y = "Analytic")+
  geom_line(color = 'black') +
  geom_smooth(method = "loess", formula = y ~ x)+
  theme(axis.text.x = element_text(angle = 270))



#Clout
# Calculate the Graph
Clout_plot=ggplot(
  AQ2_dataframe,
```

```
  aes(x=Date,
      y=Clout)) +
  labs(
    title = "Trend of Clout over the time",
    subtitle = "(2002-2011)",
    x = "Date",
    y = "Clout")+
  geom_line(color = 'black') +
  geom_smooth(method = "loess", formula = y ~ x)+
  theme(axis.text.x = element_text(angle = 270))

#Authentic
# Calculate the Graph
Authentic_plot=ggplot(
  AQ2_dataframe,
  aes(x=Date,
      y=Authentic)) +
  labs(
    title = "Trend of Authentic over the time",
    subtitle = "(2002-2011)",
    x = "Date",
    y = "Authentic")+
  geom_line(color = 'black') +
  geom_smooth(method = "loess", formula = y ~ x)+
  theme(axis.text.x = element_text(angle = 270))
```

```
#Tone
# Calculate the Graph
Tone_plot=ggplot(
  AQ2_dataframe,
  aes(x=Date,
      y=Tone)) +
  labs(
    title = "Trend of Tone over the time",
    subtitle = "(2002-2011)",
    x = "Date",
    y = "Tone")+
  geom_line(color = 'black') +
  geom_smooth(method = "loess", formula = y ~ x)+
  theme(axis.text.x = element_text(angle = 270))
grid.arrange(Analytic_plot, Clout_plot, Authentic_plot, Tone_plot)

#ppron
# Calculate the Graph
ppron_plot=ggplot(
  AQ2_dataframe,
  aes(x=Date,
      y=ppron)) +
  labs(
    title = "Trend of ppron over the time",
    subtitle = "(2002-2011)",
    x = "Date",
```

```
    y = "ppron")+
  geom_line(color = 'black') +
  geom_smooth(method = "loess", formula = y ~ x)+
  theme(axis.text.x = element_text(angle = 270))

i_plot=ggplot(
  AQ2_dataframe,
  aes(x=Date,
      y=i)) +
  labs(
    title = "Trend of i over the time",
    subtitle = "(2002-2011)",
    x = "Date",
    y = "i")+
  geom_line(color = 'black') +
  geom_smooth(method = "loess", formula = y ~ x)+
  theme(axis.text.x = element_text(angle = 270))

we_plot=ggplot(
  AQ2_dataframe,
  aes(x=Date,
      y=we)) +
  labs(
    title = "Trend of we over the time",
    subtitle = "(2002-2011)",
    x = "Date",
```

```
    y = "we")+
  geom_line(color = 'black') +
  geom_smooth(method = "loess", formula = y ~ x)+
  theme(axis.text.x = element_text(angle = 270))

you_plot=ggplot(
  AQ2_dataframe,
  aes(x=Date,
      y=you)) +
  labs(
    title = "Trend of you over the time",
    subtitle = "(2002-2011)",
    x = "Date",
    y = "you")+
  geom_line(color = 'black') +
  geom_smooth(method = "loess", formula = y ~ x)+
  theme(axis.text.x = element_text(angle = 270))

shehe_plot=ggplot(
  AQ2_dataframe,
  aes(x=Date,
      y=shehe)) +
  labs(
    title = "Trend of shehe over the time",
    subtitle = "(2002-2011)",
    x = "Date",
```

```
    y = "shehe")+
  geom_line(color = 'black') +
  geom_smooth(method = "loess", formula = y ~ x)+
  theme(axis.text.x = element_text(angle = 270))


they_plot=ggplot(
  AQ2_dataframe,
  aes(x=Date,
      y=they)) +
  labs(
    title = "Trend of they over the time",
    subtitle = "(2002-2011)",
    x = "Date",
    y = "they")+
  geom_line(color = 'black') +
  geom_smooth(method = "loess", formula = y ~ x)+
  theme(axis.text.x = element_text(angle = 270))
grid.arrange(ppron_plot,i_plot,we_plot, you_plot, shehe_plot, they_plot)
#emotion
posemo_plot=ggplot(
  AQ2_dataframe,
  aes(x=Date,
      y=posemo)) +
  labs(
    title = "Trend of posemo over the time",
    subtitle = "(2002-2011)",
```

```
    x = "Date",
    y = "posemo")+
  geom_line(color = 'black') +
  geom_smooth(method = "loess", formula = y ~ x)+
  theme(axis.text.x = element_text(angle = 270))

negemo_plot=ggplot(
  AQ2_dataframe,
  aes(x=Date,
      y=negemo)) +
  labs(
    title = "Trend of negemo over the time",
    subtitle = "(2002-2011)",
    x = "Date",
    y = "negemo")+
  geom_line(color = 'black') +
  geom_smooth(method = "loess", formula = y ~ x)+
  theme(axis.text.x = element_text(angle = 270))

anx_plot=ggplot(
  AQ2_dataframe,
  aes(x=Date,
      y=anx)) +
  labs(
    title = "Trend of anx over the time",
    subtitle = "(2002-2011)",
```

```
    x = "Date",
    y = "anx")+
  geom_line(color = 'black') +
  geom_smooth(method = "loess", formula = y ~ x)+
  theme(axis.text.x = element_text(angle = 270))

anger_plot=ggplot(
  AQ2_dataframe,
  aes(x=Date,
      y=anger)) +
  labs(
    title = "Trend of anger over the time",
    subtitle = "(2002-2011)",
    x = "Date",
    y = "anger")+
  geom_line(color = 'black') +
  geom_smooth(method = "loess", formula = y ~ x)+
  theme(axis.text.x = element_text(angle = 270))

sad_plot=ggplot(
  AQ2_dataframe,
  aes(x=Date,
      y=sad)) +
  labs(
    title = "Trend of sad over the time",
    subtitle = "(2002-2011)",
```

```
    x = "Date",
    y = "sad")+
  geom_line(color = 'black') +
  geom_smooth(method = "loess", formula = y ~ x)+
  theme(axis.text.x = element_text(angle = 270))
grid.arrange(posemo_plot,negemo_plot, anx_plot, anger_plot, sad_plot)
#focus
focuspast_plot=ggplot(
  AQ2_dataframe,
  aes(x=Date,
      y=focuspast)) +
  labs(
    title = "Trend of focuspast over the time",
    subtitle = "(2002-2011)",
    x = "Date",
    y = "focuspast")+
  geom_line(color = 'black') +
  geom_smooth(method = "loess", formula = y ~ x)+
  theme(axis.text.x = element_text(angle = 270))

focuspresent_plot=ggplot(
  AQ2_dataframe,
  aes(x=Date,
      y=focuspresent)) +
  labs(
    title = "Trend of focuspresent over the time",
```

```
    subtitle = "(2002-2011)",
    x = "Date",
    y = "focuspresent")+
  geom_line(color = 'black') +
  geom_smooth(method = "loess", formula = y ~ x)+
  theme(axis.text.x = element_text(angle = 270))

focusfuture_plot=ggplot(
  AQ2_dataframe,
  aes(x=Date,
      y=focusfuture)) +
  labs(
    title = "Trend of focusfuture over the time",
    subtitle = "(2002-2011)",
    x = "Date",
    y = "focusfuture")+
  geom_line(color = 'black') +
  geom_smooth(method = "loess", formula = y ~ x)+
  theme(axis.text.x = element_text(angle = 270))
grid.arrange(focuspast_plot, focuspresent_plot, focusfuture_plot)

#correlation
AQ2_Correlation <- webforum_with_date[c(3,6:23)]
AQ2_Correlation$Date <- as.numeric(format(AQ2_Correlation$Date, "%Y"))
webforum_with_date_with_month<-webforum_with_date
webforum_with_date_with_month$Date<-as.yearmon(webforum_with_date$Date)
```

```
webforum_with_date_with_month <- webforum_with_date_with_month[c(3,6:23)]
Average_correlation <- aggregate(AQ2_Correlation,
                                 by = list(webforum_with_date_with_month$Date),
                                 mean)
by(Average_correlation[3:20],
   factor(Average_correlation$Date),
   cor)
str(Average_correlation)
corrplot(cor(Average_correlation[3:20]),
         method = 'square',
         order = 'alphabet',
         type = 'lower',
         tl.col= "black",
         tl.srt= 45,
         diag = FALSE)+
  mtext("Total correlation", at=9, line=2, cex=1)
par(mfrow=c(2,3))
#2002
Average_correlation_2002 = Average_correlation %>% group_by(Date) %>% filter(Date
== 2002)
plot_2002=corrplot(cor(Average_correlation_2002[3:20]),
                   method = 'square',
                   order = 'alphabet',
                   type = 'lower',
                   tl.col= "black",
                   tl.srt= 45,
```

```
                diag = FALSE)+
                mtext("2002", at=9, line=0.2, cex=1)

#2004
Average_correlation_2004 = Average_correlation %>% group_by(Date) %>% filter(Date == 2004)
plot_2004=corrplot(cor(Average_correlation_2004[3:20]),
                method = 'square',
                order = 'alphabet',
                type = 'lower',
                tl.col= "black",
                tl.srt= 45,
                diag = FALSE)+
                mtext("2004", at=9, line=0.2, cex=1)

#2006
Average_correlation_2006 = Average_correlation %>% group_by(Date) %>% filter(Date == 2006)
plot_2006=corrplot(cor(Average_correlation_2006[3:20]),
                method = 'square',
                order = 'alphabet',
                type = 'lower',
                tl.col= "black",
                tl.srt= 45,
                diag = FALSE)+
                mtext("2006", at=9, line=0.2, cex=1)
```

```
#2008

Average_correlation_2008 = Average_correlation %>% group_by(Date) %>% filter(Date
== 2008)

plot_2008=corrplot(cor(Average_correlation_2008[3:20]),

                   method = 'square',

                   order = 'alphabet',

                   type = 'lower',

                   tl.col= "black",

                   tl.srt= 45,

                   diag = FALSE)+

                   mtext("2008", at=9, line=0.2, cex=1)

#2010

Average_correlation_2010 = Average_correlation %>% group_by(Date) %>% filter(Date ==
2010)

plot_2010=corrplot(cor(Average_correlation_2010[3:20]),

                   method = 'square',

                   order = 'alphabet',

                   type = 'lower',

                   tl.col= "black",

                   tl.srt= 45,

                   diag = FALSE)+

                   mtext("2010", at=9, line=0.2, cex=1)

#2011

Average_correlation_2011 = Average_correlation %>% group_by(Date) %>% filter(Date ==
2011)

plot_2011=corrplot(cor(Average_correlation_2011[3:20]),

                   method = 'square',
```

```
order = 'alphabet',

type = 'lower',

tl.col= "black",

tl.srt= 45,

diag = FALSE)+

mtext("2011", at=9, line=0.2, cex=1)
```

BQ1

```
BQ1_df <- webforum_with_date[!(webforum_with_date$posemo == 0),]

BQ1_df <- BQ1_df%>%group_by(ThreadID,year,month)%>%summarise(posemo = mean(posemo,na.rm = TRUE))


BQ1_df <- subset(BQ1_df, posemo >= mean(BQ1_df$posemo))

BQ1_df_without_repeat = BQ1_df %>% group_by(ThreadID) %>% mutate(n=n()) %>% filter(n==1) %>% select(-n)

BQ1_df <- setdiff(BQ1_df,BQ1_df_without_repeat)

average = mean(BQ1_df$posemo)


BQ1_df <- subset(BQ1_df, posemo >= mean(BQ1_df$posemo))

BQ1_df_without_repeat = BQ1_df %>% group_by(ThreadID) %>% mutate(n=n()) %>% filter(n==1) %>% select(-n)

BQ1_df <- setdiff(BQ1_df,BQ1_df_without_repeat)

average = mean(BQ1_df$posemo)



BQ1_df <- subset(BQ1_df, posemo >= mean(BQ1_df$posemo))

BQ1_df_without_repeat = BQ1_df %>% group_by(ThreadID) %>% mutate(n=n()) %>% filter(n==1) %>% select(-n)
```

```
BQ1_df <- setdiff(BQ1_df,BQ1_df_without_repeat)

average = mean(BQ1_df$posemo)


remove(BQ1_df_without_repeat)

ggplot(BQ1_df, aes(x=year, y=posemo)) +

  geom_point(aes(size = posemo, colour = factor(ThreadID)))
```

CQ1 AND CQ2

```
#find out the frequancy of post

frequancy_post = as.table(by(webforum_with_date,webforum_with_date$Year,nrow))

frequancy_post = as.data.frame(frequancy_post)

#choose the highest frequancy

webforum_05_12 = webforum_with_date[webforum_with_date$Year_and_month=="05-
12",]


#delete author only post once time

webforum_05_12_without_repeat = webforum_05_12 %>% group_by(AuthorID) %>%
mutate(n=n()) %>% filter(n==1) %>% select(-n)

webforum_05_12 = setdiff(webforum_05_12,webforum_05_12_without_repeat)

remove(webforum_05_12_without_repeat)


#delete thread only show once time

webforum_05_12_without_repeat = webforum_05_12 %>% group_by(ThreadID) %>%
mutate(n=n()) %>% filter(n==1) %>% select(-n)

webforum_05_12 = setdiff(webforum_05_12,webforum_05_12_without_repeat)

remove(webforum_05_12_without_repeat)


#find top 30 most posts author
```

```
top_authors = count(webforum_05_12)
top_authors = setDT(top_authors)[order(-n), .SD[1:30]]
sum(top_authors$n)

#merge
webforum_05_12 = merge(top_authors, webforum_05_12, by = "AuthorID")
webforum_05_12 = unique(webforum_05_12)
webforum_05_12 = select(webforum_05_12, AuthorID, ThreadID)

#delete threadID and unique graphdata
graphdata = dplyr::inner_join(webforum_05_12, webforum_05_12, by = "ThreadID")
graphdata = graphdata[graphdata$AuthorID.x!=graphdata$AuthorID.y]
graphdata$ThreadID=NULL
graphdata = unique(graphdata)

#draw graph
g = graph.data.frame(graphdata, directed=F)
#duplicate: 586
E(g)
g = simplify(g, remove.multiple = T, remove.loops = T)
#drop to 293
E(g)
plot(g, vertex.color = "red")
```

```
#CQ2

# Compare clustering coefficient of graphs

transitivity(g)

# Compare individual vertex based on closeness centrality

closeness(g)

# Compare individual vertex based on betweenness centrality

betweenness(g)

#eigen_centality

eigen_centrality(g)


#find the reason

#choose the highest frequancy

webforum_41237 = webforum_with_date[webforum_with_date$Year_and_month=="05-
12",]


#delete author only post once time

webforum_41237_without_repeat = webforum_41237 %>% group_by(AuthorID) %>%
mutate(n=n()) %>% filter(n==1) %>% select(-n)

webforum_41237 = setdiff(webforum_41237,webforum_41237_without_repeat)

remove(webforum_41237_without_repeat)


#delete thread only show once time

webforum_41237_without_repeat = webforum_41237 %>% group_by(ThreadID) %>%
mutate(n=n()) %>% filter(n==1) %>% select(-n)

webforum_41237 = setdiff(webforum_41237,webforum_41237_without_repeat)

remove(webforum_41237_without_repeat)
```

```
#make table 41237 2005-12 posting only
webforum_41237 = webforum_41237[webforum_41237$AuthorID=="41237",]
webforum_41237[ ,c(1,2,3,4,24,25,26,27)] <- list(NULL)
webforum_41237 = data.frame(t(webforum_41237[-1]))
#using AQ2 data frame get the general data
webform_average_05_12 = AQ2_dataframe[AQ2_dataframe$Date=="Dec 2005",]
webform_average_05_12$Date = NULL
webform_average_05_12 = data.frame(t(webform_average_05_12[-1]))
colnames(webform_average_05_12) <- "attribute"
#merge
test <- merge(webform_average_05_12, webforum_41237, by=0, all=TRUE)
rownames(test) <- test$Row.names
test$Row.names<-NULL
#draw
ggplot(test, aes(x = row.names(test))) +
  geom_line(aes(y = X1,group = 1)) +
  geom_line(aes(y = X2,group = 1)) +
  geom_line(aes(y = X3,group = 1)) +
  geom_line(aes(y = X4,group = 1)) +
  geom_line(aes(y = X5,group = 1)) +
  geom_line(aes(y = X6,group = 1)) +
  geom_line(aes(y = X7,group = 1)) +
  geom_line(aes(y = X8,group = 1)) +
  geom_line(aes(y = attribute,group = 1,color="red"))+
  labs(
    title = "features of posts of 41237 vs average features rank",
```

```
  x = "Features",
  y = "rank"
)
```