FIT3152 Data analytics: Assignment 1

Ritvik Mahadeshwar | FIT3152 |10/05/2020

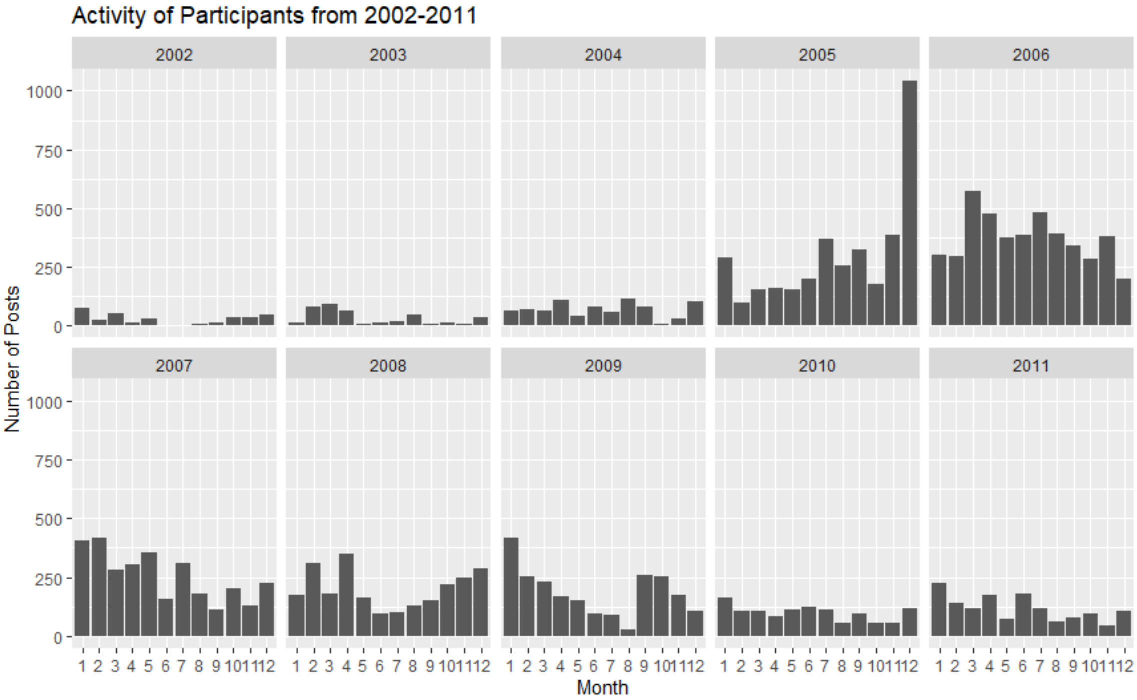
Introduction:

This report analyses metadata and linguistic summary from a real on-line forum. The data is contained in the file webforum.csv and consists of the metadata and linguistic analysis of posts over the years 2002 to 2011. The linguistic analysis was conducted using Linguistic Inquiry and Word Count (LIWC), which assesses the prevalence of certain thoughts, feelings and motivations by calculating the proportion of key words used in communication. A sample of 20,000 posts was taken to be analyzed, visualized and cross compared to each other to find trends and connections with the authors who wrote in it, the thread it is associated with and timestamps indicating when these posts were written.

Section A

Describe your data: How active are participants, and are there periods where this increases or decreases? Is there a trend over time?

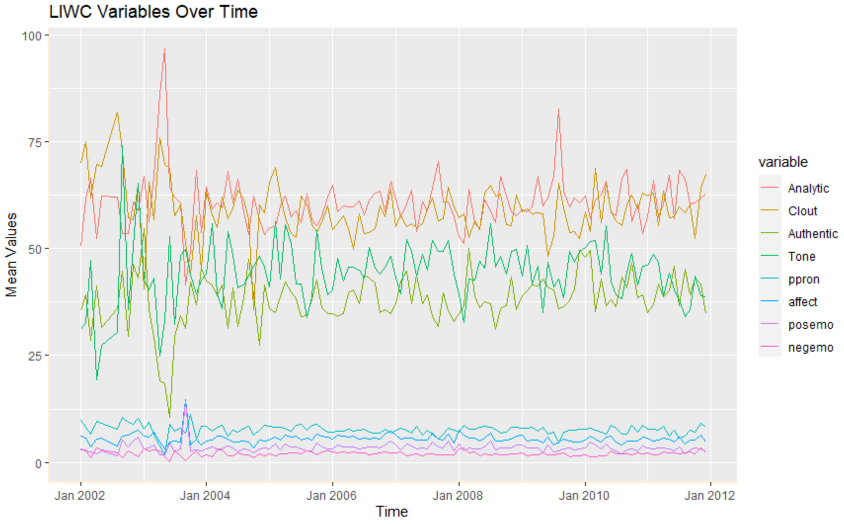
Figure a.1



Let us first describe figure 1.a. As you can see we have created a column graph that represents number of posts made by all authors in our dataset on our x-axis and our y axis containing the months from 1-12 as per a normal year. The beauty of this graph is that it has been subdivided into 10 separate plots each showcasing activity of authors in each year (2002 to 2011) throughout the year's months. This allows us to clearly see how active participants have been between years and if we were to draw comparisons from year to year or even holistically, it is now possible. So, let's dive into some analysis to look at some trends and answer our question. If we closely examine the first three years **2002, 2003, 2004,** we can see level of activity has been relatively low with none of the months recording anywhere more than 125 posts. The reasons for this are quite broad and there could be several influences impinging the data, but we can make an educated assumption. We have to realise the early 2000's is where we saw movements into technological advancements and accessibility. The internet was created in the 1980's but with all things it takes time for public acceptance and use. We see this lag in our data. Of course, there were online posts being made but due to probable high costs with accessing the Internet (i.e needing a laptop) and generational norms of using physical journals and newsletters, it could explain this relatively low activity. However, 2005 is a very interesting and exciting proportion of our data. January 2005 straight away records a new high for no. of posts and for most of the year we see a gradual rise with December 2005 recording a mammoth 1000+ posts. This spike shouldn't be unexpected because like mentioned before the 2000's were changing times where digitization was a core driver. Millennials often associated with birth years between 1990 to 2000 was a demographic cohort that were widely accepting of technology and the use of it. In terms of our timeline it fits well possibly explaining this sudden rise. Moving on to 2006, we do see a decline in activity compared to 2005 but still relatively high in contrast with the other years. However, the end of 2006 is a significant point as our activity level essentially begins to decline from here. Declines in 2007-2009 may be associated with the GFC. The GFC was a global crisis that affected millions of lives around the world and it is quite understandable with people losing homes and jobs posting on online web forums wouldn't be the first thing the public would be thinking of. However, in saying that the end of 2008 and start of 2009 did see a slight increase in activity and this could be due to people sharing personal experiences on hwo they were affected and wanting to share with others to spread positivity and inclusiveness. Unfortunately, the decline continues into years 2010 and 2011 where activity of participants can be described being similar to what we saw the first three years. Overall the trend of activity is quite interesting, and we can conclude the trend to be positively skewed. Major takeaway is year 2005 and the month December in that year particularly, where we saw a huge increase. *[For graph that shows posts over just years – simpler idea, check the appendix]*

Looking at the linguistic variables, do these change over time? Is there a relationship between them?

Figure a.2

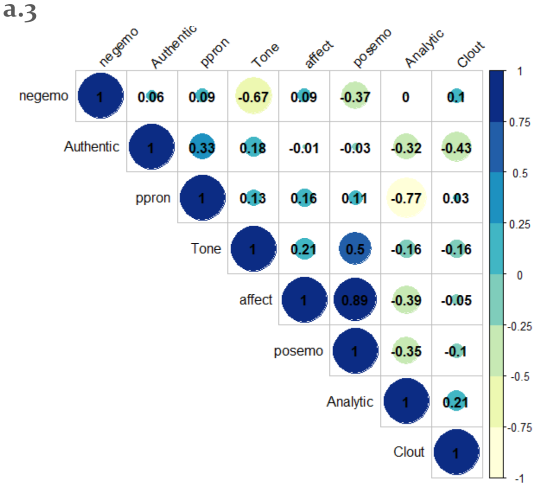


Once again let's talk through what this graph is representing and how we came about creating it. We have plotted the x-axis as our timeline taking January intervals for years between 2002-2011. Our y-axis are our aggregated mean values of our linguistic variables. The reason for this is remember we have 20,000 observations for each linguistic variable as it relates to a post made by an author in a particular thread. It would be impossible to graph all values so we have aggerated a mean of all LIWC variables for each month in each year, hence getting a nice continuous line graph which, we can interpret do describe if linguistic variables do actually change over time. What does the mean value of linguistic variables tell us? Well it tells us the proportion of a single post that uses words expressing a particular sentiment (linguistic variable) and it is important to keep this in mind when analyzing our graph. Notice if you take one point in the graph and draw a line all the way down, all the variables accumulated together will not add to 100%. This is because there are possible overlaps with words that can be classified as being associated with more than one variable. Finally, we have our variables plotted on our graph with different colors distinguishing each. The graph does not depict all of the linguistic variables, instead the most important ones which we feel are the LIWC summary variables (Analytic, Clout, Authentic and Tone), ppron (personal pronouns), affect (expressing sentiment), posemo (positive emotions) and finally negemo (negative emotions). It was not necessary to add varibales such as **i, we, you, shehe and they** as these variables are personal pronouns and can be assumed to have positive relationships with ppron variable.

This graph helps us answer the prompt whether linguistic variables change over time and concluding from the graph there are few that surely do over the months and year. Let's first look at the bottom of our graph. Linguistic variables such as ppron, affect, negemo and posemo do not change much and remain constant over our timeline. The mean values they operate in are between 0 and 12.5% and never really leave this bandwidth, indicating little or no change of authors using these types of languages over the 10 years. However, it should be noted that around the end of 2003 we see a sudden simultaneous spike in posemo and affect linguistic variables which could explain a relationship between them (but let us worry about that in our correlation plot explanation which is our next graph). On the flip side our summary variables Analytic, Clout, Authentic and Tone tell us a completely different story. We see a wide variety of fluctuations for these variable. For example, the use of analytic language can increase to 80-90% (recorded mid 2003 and Jan 2010) but then also drop down to around 40% (recorded end of 2003). Use of authentic language generally stay around 30-40%, but we saw a sudden drop to around 10% mid 2003. These are few examples of the drastic changes in language use but even the constant regular change between 40 to 70% seen from our summary linguistic over our 10 year timeline indicates significant variation caused by time.

To provide logical rationale behind why we see these changes occurring at different points in time is quite difficult due to the unpredictability of these fluctuations. The fact that we don't know exactly the origins of the posts and what particular topic they address makes it even hard for us to associate with factors that could have caused changes in the use of linguistic variables. We can assume and make our own opiniated guesstimates with relationships we see within our graph. For example, in the spring of 2003 we saw a rise in affect and posemo variables and a subsequent decline in both as well start of 2004. This could suggest maybe a political campaign or change in government regulations that benefited the society which would imply more positive and sentimental language being used within posts made on web forums. Coinciding with the spring of 2003 a political campaign around this timeline would also explain the sudden exponential rise in analytical language as well, as more people dissect and decipher the meanings behind the political movement.

Figure a.3

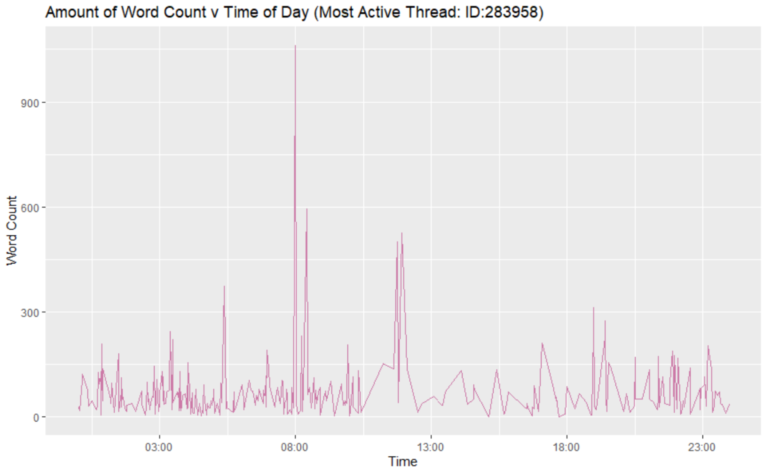


While figure a.2 focused more on time affecting linguistic variables, figure a.3 depicts a correlation plot explaining the relationship between each variable we analyzed before. Remember positive numbers indicate that the variables move in the same direction, while a negative number indicates that the variables move in opposite directions. For the sake of a good analysis let's look at a few specific correlations that are closest to 1 and -1. The variable posemo (positive emotion) is a very interesting one. Like mentioned before variables posemo and affect are highly positively linked. When authors express higher positive emotions they tend to also express more sentiment which logically makes sense. Posemo also has fairly positive impact on our Tone summary variable. We find that generally when authors are writing with emotional tones they tend to be filled with words that display positive emotions as well. Now looking at some of our negative relationships, Tone + negemo and Anlaytic + ppron are two correlations that are interesting. Firstly, we see when authors tend be more analytic in their thinking when writing there is a lesser, in fact the use of personal pronouns reduces drastically. When we think of analytical writing it does tend to be more third person views stating facts and involvement of numbers and can explain the correlation of -0.77 between the variables. Additionally, we see a similar picture with Tone and negemo. A lot of the emotional tones posts that authors write on this web forum are positive and it backs up what we talked before with posemo and Tone.

Section B

Threads indicate groups of participants communicating on the same topic. Describe the threads present in your data.

Figure b.1

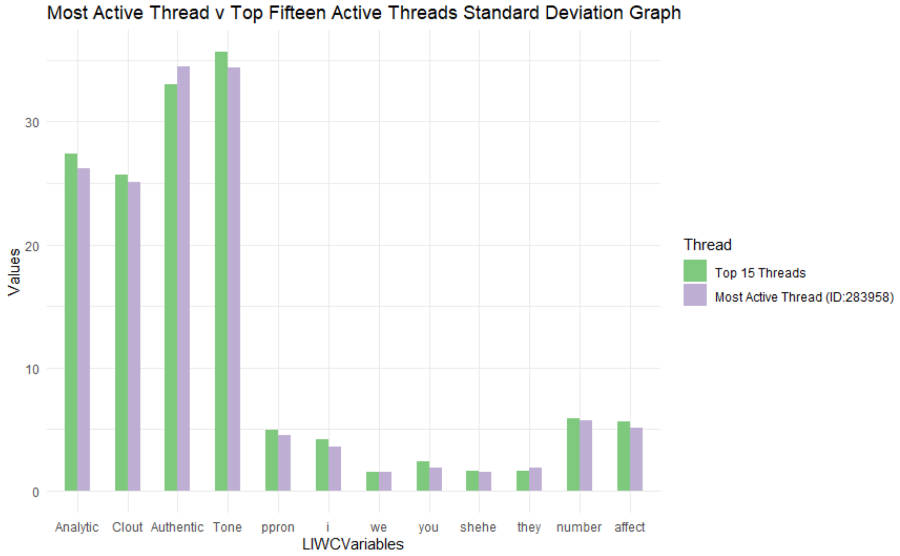


In this graph we have once again stuck to our popularity method instead of analyzing all threads. Analyzing all threads is a difficult task and visually representing can become cluttered. So, we picked the most popular thread to give us a better idea and control when analyzing. Our y-axis once again the aggregated mean of word count and we are using the x-axis as a timeframe to describe our thread. We can clearly see from the density of the graph between midday till 6pm there are not a lot of authors making a post around this time, which makes sense as most people would be at work. However, we see a higher density occurring after 10pm with more authors frequently posting currently. What is interesting to see is that there is a sudden massive increase in word count around 8-9am. Generally, we can say that the time of a day dictates when posts are made in threads and this happens to be at night when people are likely to have access to more free time after their commitments.

[For an extra graph that shows top 10 threads and what proportion of our summary variables make up the thread language, check the appendix]

By analyzing the linguistic variables for all or some of the threads, is it possible to see a difference in the language used by these different groups?

Figure b.2



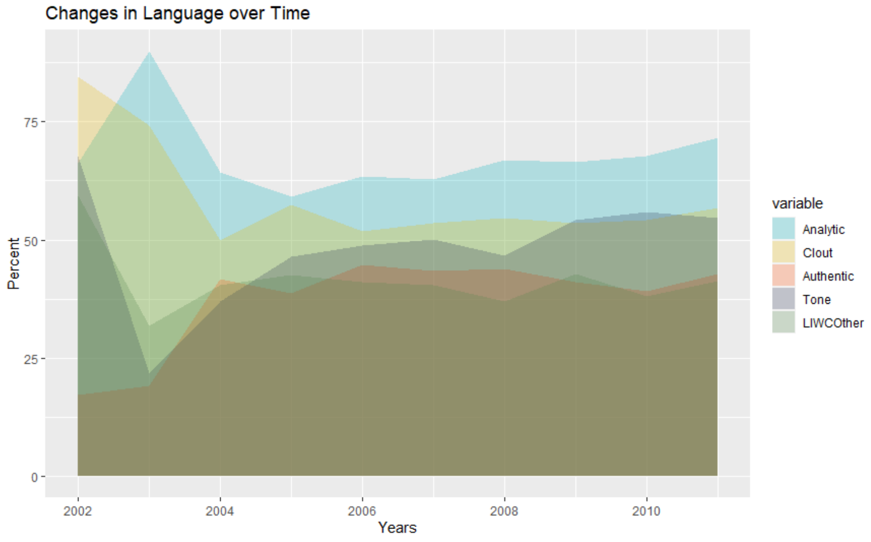
This test was conducted to prove that language does vary between different groups of people. The results above were obtained through aggregating the data by thread ID and calculating the mean of each of the LIWC variables, note we have not included specific words such as anger, social, family etc. as analysing linguistic variables will cover all of the other separate words. After aggregating we then found the most active thread and separated into its own data frame. The same was done with the top fifteen threads. Consequently, the standard deviation of both sets was found and plotted onto a histogram. The reason we have used the standard deviation function is because like mentioned before we want to see the difference in the language used by these different groups. It is logically impossible to analyze all threads so we took the most active one which we feel would incorporate a variety of linguistic variables (our control) and compared it to the top fifteen threads. The histogram shows that the standard deviation is higher in most attributes tested in the fifteen ten threads when compared against the most active thread in the forum. This suggests that there is a difference between threads in that individuals participating in a single thread will adopt different language patterns to those in another thread.

```
Welch Two Sample t-test

data:  Analytic and Authentic
t = 13.738, df = 1623.6, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 16.89441 22.52205
sample estimates:
mean of x mean of y 
 59.36609  39.65786 
```

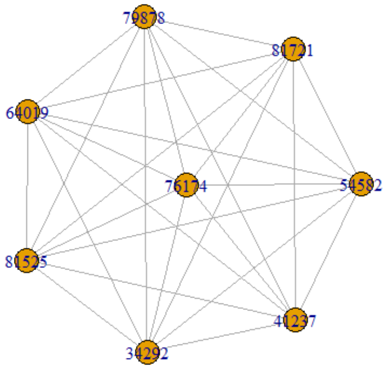
We can use a t-test to prove that there is a significant difference between the language used. The t-test has analysed Analytic and Authentic linguistic variables in 2004 and has given us a t value of 13.738. The greater the magnitude of **T**, the greater the evidence against the null hypothesis. This **means** there is greater evidence that there is a significant difference.

Does the language used within threads change over time?
Figure b.3

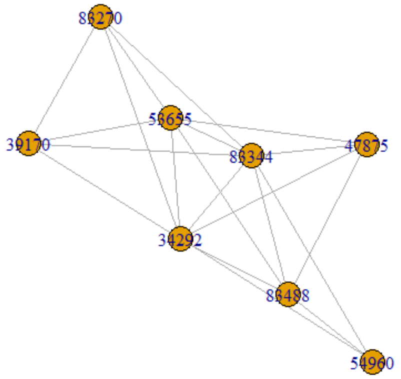


To analyse the change in the language used within threads over the 10 years, only the 4 summary linguistic variables have been plotted and the rest including the words, numbers and question mark characteristics have been combined into an Other category. The reason for this is that the summary variables incorporate most of the specific variables and are far more important to examine as it gives a better and holistic view of language. Once again this graph does not incorporate all threads as analyzing each and every one of them is not a logical approach. We have taken the top 40 threads as a suitable sample to avoid errors and variation and allow for a better comparability, rather than just selecting randomly. Our y -axis is self-explanatory indicating our timeline and our x-axis indicates the proportion of the language used in post within our 40 threads. From the plot we can clearly see a decrease in the level of Analytic and Clout language used in our 40 threads, whereas Authentic use of language has increased. Tone seems to fluctuate through the years but we do see a spike in the variable in 2011.

Section C
Can you define, graph and describe the social network that exists at a particular point in time, for example over one month? How does this change in the following months?



This network depicts the 8 most popular authors in the month December and year 2005. This period of 1 month was the most popular in terms of number of posts made and our graphs from section A support this. This network diagram shows all the connections between authors in terms of the threads they are communicating within and we can see all authors in one way or another have communicated in a similar thread as others, hence a complete clique network. There is no one most popular author as all of them are posting different posts in different threads, hence establishing the connections. This seems probable because we are analyzing the most popular month which recorded over a 1000 posts in the web forum, so it is highly likely that a post made by say author 76174 in a particular thread has also seen the other 7 authors make a post in that same thread.



However, in Jan 2006, which is the month subsequently straight after the one we analyzed before explains a different story. We can see how time has changed our social network. Jan 2006 obviously recorded fewer overall posts over multiple threads. Once again, we have taken our top 8 authors in this period and looked at any similarities in the threads they have posted. Author 83488 is our central author who relates to all other authors. However, we have quite a few authors in our network who aren't not related indicating they did not make posts in the same threads, indicating that maybe this period didn't see authors willing to communicate on similar topics Author 54960 and 39170 are examples.

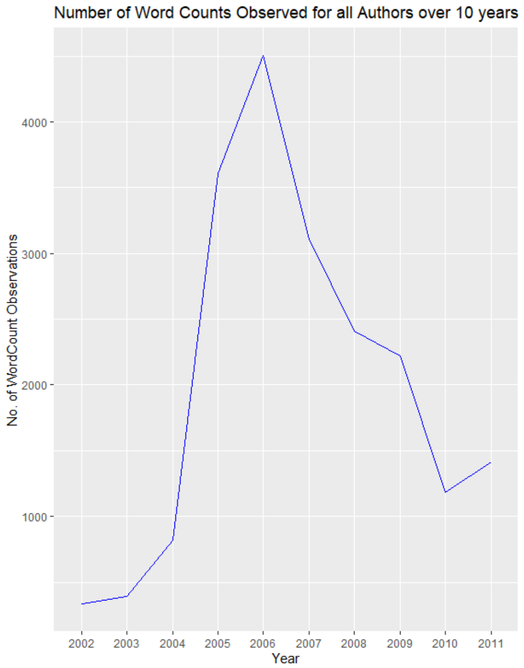
Appendix:

Creating Individual Data

```
install.packages("readr")
library(readr)
rm(list = ls())
set.seed(29677068)
webforum = read_csv("webforum.csv")
webforum = webforum [sample(nrow(webforum), 20000), ]
```

Question a

```
#Describe your data: How active are participants, and are there periods where this
increases
#or decreases? Is there a trend over time?
install.packages("ggplot2")
library(ggplot2)
WF = webforum
WF = subset(webforum, WC != 0)
WF$tempdate = as.Date(WF$Date, "%d/%m/%Y") # Create a temporary table for year
WF$year = as.numeric(format(WF$tempdate, "%Y")) # Extract the year
WF$tempdate = NULL
WF$tempdate2 = as.Date(WF$Date, "%d/%m/%Y" )
WF$month = as.numeric(format(WF$tempdate2, "%m"))# extract the month
WF$tempdate2 = NULL
attach(WF)
#Second Trial
AuthMonthYear = as.table(by(WC, list(year, month), length))
AuthMonthYear = as.data.frame(AuthMonthYear)
View(AuthMonthYear)
colnames(AuthMonthYear) = c("Year", "Month", "WordCount")
#Creating the graph
r = ggplot(data = AuthMonthYear)
r = r + geom_col(mapping = aes(x = Month, y = WordCount))
r = r + scale_x_discrete(breaks = c(1, 2, 3, 4,
    5, 6, 7, 8, 9, 10, 11, 12))
r  = r + facet_wrap(~Year, nrow = 2)
r  = r + ylab("Number of Posts")
r = r + ggtitle("Activity of Participants from 2002-2011")
r
```



Extra graph for question a part 1

```
#First Trial
AuthWordCount = as.table(by(WC, year, length)) # Using length get the no of observations of word
count in each yea
AuthWordCount = as.data.frame(AuthWordCount)
colnames(AuthWordCount) = c("Year", "WordCount")
AuthWordCount$Year = as.numeric(AuthWordCount$Year)
View(AuthWordCount)
View(WF)

m = ggplot(AuthWordCount, aes(x = Year, y = WordCount, group = 1))
m = m + geom_line(color ="blue")
m =  m + ylab("No. of WordCount Observations")
m = m + ggtitle("Number of Word Counts Observed for all Authors over 10 years")
m
```

```
#Question a (Second Dot Point) : Looking at the linguistic variables, do these change over
time? Is there a relationship
#between them?
install.packages("zoo")
install.packages("dplyr")
install.packages("reshape2")
library(zoo)
library(reshape2)
library(dplyr)
library(ggplot2)
WFliwcsumm = WF[c(3,6,7,8,9,10,17,18,19)]
WFliwcsumm$Date = as.Date(WF$Date, "%d/%m/%Y")
```

```
WFliwcsumm$Date = as.yearmon(WFliwcsumm$Date, "%Y-%m")
ling_month_year = aggregate(WFliwcsumm, by = list(WFliwcsumm$Date), mean)
ling_month_year$Group.1 = NULL
lingmelt = melt(ling_month_year, id.vars = "Date")
colnames(lingmelt) = c("Date", "variable", "Avgvalue")
#Creating the graph
library(ggplot2)
lg = ggplot(data = lingmelt, aes(x = as.yearmon(Date), y =Avgvalue, colour = variable, group
=variable))
lgline = lgline + geom_line()
lgline = lgline + theme(axis.line = element_line(colour = "maroon", size =2, linetype = "solid"))
lgline = lgline + xlab("Time")
lgline = lgline + ylab("Mean Values")
lgline = lgline + ggtitle("LIWC Variables Over Time")
lgline
#Creating a correlation plot
WFCorrelation = WFliwcsumm
View(WFCorrelation)
WFCorrelation$Date = as.Date(WF$Date, "%d/%m/%Y") # Create a temporary table for year
WFCorrelation$Date = as.numeric(format(WFCorrelation$Date, "%Y"))
CorrAvg = aggregate(WFCorrelation, by = list(WFliwcsumm$Date), mean)
View(CorrAvg)
CorrAvg$Group.1 = NULL
by(CorrAvg[2:9], factor(CorrAvg$Date), cor)
install.packages("corrplot")
library(corrplot)
library(RColorBrewer)
Z = cor(CorrAvg[2:9])
corrplot(Z,method = "circle", addCoef.col = "black" ,type="upper", order="hclust", tl.col= "black",
tl.srt= 45, col=brewer.pal(n=8, name="YlGnBu"))
```

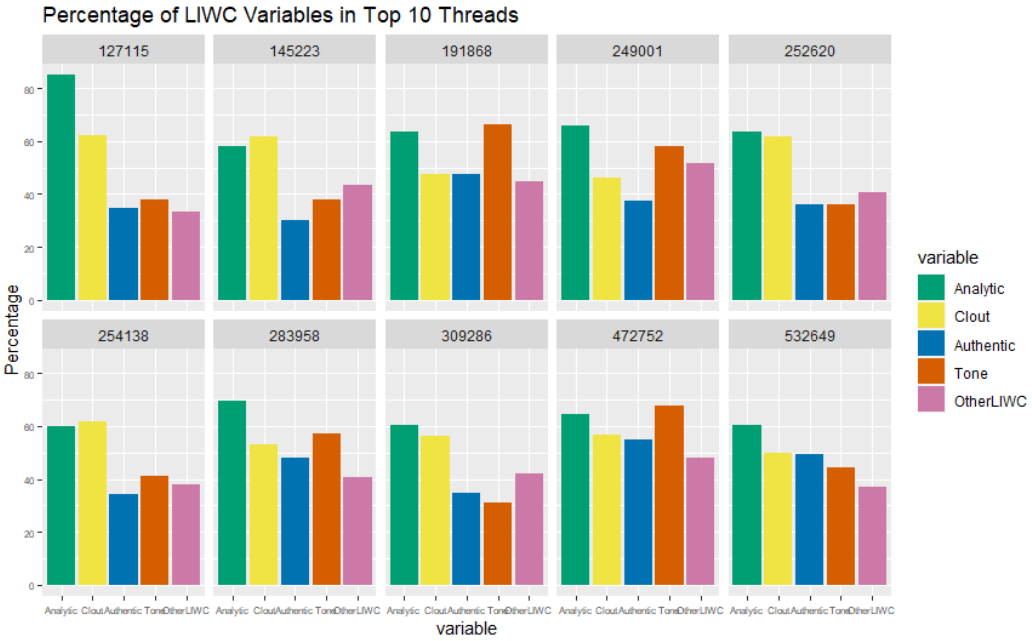
Question b

```
#Part B (First Dot point) : Analyse the language used by groups.
# 1. Threads indicate groups of participants communicating on the same topic. Describe the
#    threads present in your data.
attach(webforum)
cleanerWF = webforum[!(webforum$WC==0),]
attach(cleanerWF)
install.packages("plyr")
install.packages("scales")
library(plyr)
library(ggplot2)
library(reshape2)
library(dplyr)
library(scales)
groupThread = aggregate(cleanerWF[5:29], cleanerWF[1], mean)
freqCount= data.frame(count(cleanerWF,c("ThreadID")))
```

```
freqCount[which.max(freqCount$freq),] #this line of code tells us which ThreadID has the highest
frequency
maxthread = subset(cleanerWF, cleanerWF$ThreadID == "283958")
maxthread$Time <- as.POSIXct(maxthread$Time, format="%H:%M")

ggplot(data=maxthread, aes(x=maxthread$Time,y=maxthread$WC)) +
 labs(title="Amount of Word Count v Time of Day (Most Active Thread: ID:283958)", x="Time",
y="Word Count") +
 geom_line(colour="#CC79A7") +
 scale_x_datetime(date_breaks = "5 hour", labels = date_format("%H:%M"))
```

Extra graph



```
####(Still First Dot Point)
#2.Threads indicate groups of participants communicating on the same topic. Describe the
# threads present in your data.
install.packages("plyr")
library(plyr)
v = webforum[!(webforum$WC==0),]
View(v)
View(u)
u = aggregate(v[5:29], v[1], mean)
count1 = data.frame(count(v, c("ThreadID")))
t10 = as.data.frame(head(count1[order(-count1[2]),],10))
t10I=subset(v,ThreadID%in%t10$ThreadID)
attach(t10I)
q = as.data.frame(as.table(by(Analytic, ThreadID, mean)))
```

```
colnames(q) = c("ThreadID", "Analytic")
View(q)
j = as.data.frame(as.table(by(Clout, ThreadID, mean)))
colnames(j) = c("ThreadID", "Clout")
View(j)
y = as.data.frame(as.table(by(Authentic, ThreadID, mean)))
colnames(y) = c("ThreadID", "Authentic")
View(y)
d = as.data.frame(as.table(by(Tone, ThreadID, mean)))
colnames(d) = c("ThreadID", "Tone")
View(d)
SummBgraph = merge(q, j, by = "ThreadID")
SummBgraph = merge(SummBgraph, y, by = "ThreadID")
SummBgraph = merge(SummBgraph, d, by = "ThreadID")
library(dplyr)
mydata = t10I
View(mydata)
mydata = aggregate((ppron + i + we + you + shehe + they + number + affect + posemo + negemo+
anx + anger + social + family + friend + leisure + money + relig + swear + QMark)~ThreadID, t10I,
mean)
colnames(mydata) = c("ThreadID", "OtherLIWC")
SummBgraph = merge(SummBgraph, mydata, by ="ThreadID")
summary(SummBgraph)
library(plyr)
library(dplyr)
library(reshape2)
MeltSumm = melt(data = SummBgraph, id.vars = "ThreadID")
library(ggplot2)
r = ggplot(data = MeltSumm)
r = r + geom_col(mapping = aes(x = variable, y = value, fill = variable))
r = r + scale_fill_manual(breaks = c("Analytic", "Clout",
"Authentic", "Tone", "OtherLIWC"), values = c("#009E73", "#F0E442", "#0072B2",
"#D55E00", "#CC79A7"))
r = r + facet_wrap(~ThreadID, nrow = 2)
r = r + ylab("Percentage")
r = r + ggtitle("Percentage of LIWC Variables in Top 10 Threads")
r = r + theme(axis.text = element_text(size = 7.45))
r
```

#Second dot point (Part B) By analysing the linguistic variables for all or some of the threads, is it possible to see a
#difference in the language used by these different groups?

```
######Single Thread v Multiple Threads – SD
cleanerWF = webforum[!(webforum$WC==0),]
#Group Data by Thread
```

```
groupT = aggregate(cleanerWF[5:29], cleanerWF[1], mean)
attach(groupT)
#Find the most active thread and subset
freqC= data.frame(count(cleanerWF, c("ThreadID")))
View(maxthread)
freqC[which.max(freqC$freq),]#Most active thread is 283958ID
#Find the top 15 active threads and subset
topFifteen=as.data.frame(head(freqC[order(-freqC[2]),],15))
topFifteenInfo=subset(cleanerWF,ThreadID%in%topFifteen$ThreadID)#Each threads info
#Finding SD of max thread
SDmaxthread=(apply(maxthread[6:17],2,sd))
SDmaxthread=as.data.frame(SDmaxthread)
#Finding SD of top 15 threads
SDtopFifteen=apply(topFifteenInfo[6:17],2,sd)
SDmaxthread$SDtopFifteen=as.data.frame(SDtopFifteen)
View(SDmaxthread)
#Binding and transposing attributes in order to melt data
SDmaxthread2=t(rbind(t(SDmaxthread[,2]),t(SDmaxthread[,1])))
#melting data and renaming columns
SDThreadMelt=melt(SDmaxthread2)
colnames(SDThreadMelt)=c("LIWCVariables","Thread","Values")


ggplot(SDThreadMelt, aes(x=LIWCVariables,y=Values,fill=Thread))+
  geom_bar(stat="identity", width=.5,position="dodge")+
  ggtitle("Most Active Thread v Top Fifteen Active Threads Standard Deviation Graph")+
  scale_fill_brewer(palette = "Accent", labels=c("Top 15 Threads", "Most Active Thread
(ID:283958)"))+theme_minimal()
```

#### Question b Part3 - Does the language used within threads change over time?

```
View(WF)
library(plyr)
cleanerWF = webforum[!(webforum$WC==0),]
ThreadGroup = aggregate(cleanerWF[6:29], cleanerWF[1], mean)
FrequencyCount= data.frame(count(cleanerWF, c("ThreadID")))
Top40Threads = as.data.frame(head(FrequencyCount[order(-FrequencyCount[2]),],40))
top40Info=subset(cleanerWF,ThreadID%in%Top40Threads$ThreadID)

A1 = aggregate(cbind(Analytic, Clout, Authentic, Tone)~ ThreadID + year, top30Info, mean)
A2 = aggregate((ppron + i + we + you + shehe + they + number + affect + posemo + negemo+ anx +
anger + social + family + friend + leisure + money + relig + swear + QMark)~ThreadID + year,
top30Info, mean)
colnames(A2) = c("ThreadID", "year", "LIWCOther")
A3 = merge(A1, A2, by = c("ThreadID", "year"))
A3$ThreadID = NULL
A4 = aggregate(A3[2:6], A3[1], mean)
Amelt = melt(data= A4, id.vars = "year")
```

```
ggplot(Amelt, aes(year, value, fill=variable)) +
  geom_area(aes(fill=variable),position="identity") +
  scale_fill_manual(values = alpha(c("#E7B800",  “#00AFBB” , "#293352", "#FC4E07", "#52854C"),
.25)) +
  labs(
    title="Changes in Language over Time",
    y="Percent", x="Years"
  )
install.packages("expss")
library(expss)
```

Question C

```
##Question c) Can you define, graph and describe the social network that exists at a
particular point in
#time, for example over one month? How does this change in the following mo nths?
library(zoo)
attach(webforum)
library(plyr)
install.packages("zoo")
install.packages("plyr")
NetThr = webforum[!(webforum$AuthorID==-1),]
NetThr = NetThr[!(NetThr$WC==0),]
NetThr$Date = as.Date(NetThr$Date, "%d/%m/%Y")
NetThr$Date = as.yearmon(NetThr$Date, "%Y-%m")

#Most Aactive Month (Dec 2005)

NetThread2 =aggregate(NetThr[5:29], NetThr[3], mean)
fqcNetwork = as.data.frame(count(NetThr, "Date"))
View(fqcNetwork)
fqcNetwork[which.max(fqcNetwork$freq),] #Month Dec 2005 is the most active
maxNetThr = subset(NetThr, NetThr$Date=="Dec 2005")
AuthNetworkFreq= data.frame(count(maxNetThr, c("AuthorID")))
TopEightAuthNetwork = as.data.frame(head(AuthNetworkFreq[order(-AuthNetworkFreq[2]),],8))
TopEightAuthNetworkinfo = subset(maxNetThr,AuthorID%in%TopEightAuthNetwork$AuthorID)
AdjacencyTable = TopEightAuthNetworkinfo[, -c(4:29)]
AdjacencyTable = AdjacencyTable[, -c(3)]
View(AdjacencyTable)
library(igraph)
library(igraphdata)
data(karate)
View(karate)
graph    <- graph.data.frame(AdjacencyTable,       directed=FALSE)
get.adjacency(graph)
plot(graph, vertex.colour = "red")
```

```
gg = graph_from_literal(76174 : 34292       : 64019 : 81721 : 81525 : 41237 : 54582 : 79878 -- 76174:
34292: 64019       : 81721 : 41237 : 54582 : 79878)
plot(gg)

#Next Month (Jan 2006)
NextSecNetThr = subset(NetThr, NetThr$Date=="Jan 2006")
Auth2Freq= data.frame(count(NextSecNetThr, c("AuthorID")))
TopEight2AthNet = as.data.frame(head(Auth2Freq[order(-Auth2Freq[2]),],8))
TopEight2Info = subset(NextSecNetThr,AuthorID%in%TopEight2AthNet$AuthorID)
AdjacencyTable2 = TopEight2Info[, -c(4:29)]
AdjacencyTable2 = AdjacencyTable2[, -c(3)]
View(AdjacencyTable2)
cc = graph.formula(34292-39170, 34292-53655, 34292-83270, 34292-83344, 34292-54960, 34292-
47875, 34292-83488, 39170-34292, 39170-53655, 39170-83270, 39170-83344, 47875-53655, 47875-34292,
47875-83344, 47875-83488, 53655-34292, 53655-39170, 53655-83270, 53655-83344, 53655-47875,
53655-83488, 54960-34292, 54960-83344, 54960-83488, 83270-34292, 83270-39170, 83270-53655,
83270-83344, 83344-34292, 83344-39170, 83344-53655, 83344-83270, 83344-47875, 83344-83488,
83344-54960, 83488-34292, 83488-47875, 83488-53655, 83488-83344, 83488-54960)
plot(cc)

#t.test for question b part 2
#Most popular thread in every year
attach(webforum)
#2004
ttest = webforum[!(webforum$WC==0),]
View(ttest)
ttest$Date = as.Date(ttest$Date, "%d/%m/%Y") # Create a temporary table for year
ttest$Date = as.numeric(format(ttest$Date, "%Y")) # Extract the year
#Pick year 2004
ttest2004 = subset(ttest, ttest$Date=="2004")
View(ttest2004)
attach(ttest2004)
t.test(Analytic, Authentic)
```