MONASH University

FIT3152 Data Analytics: Assignment 1

Tutorial group: Monday, 12pm – 2pm

Monash ID: 29634431

Name: Lai Wei Jian

Introduction

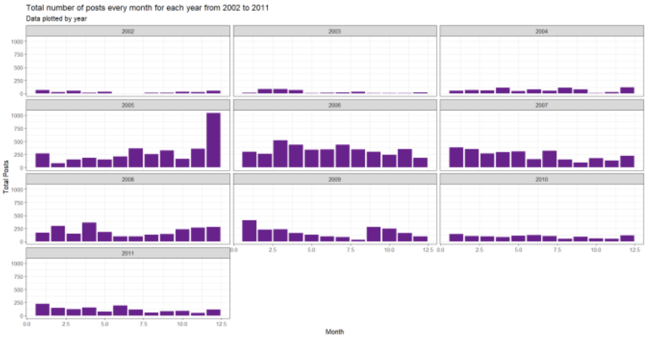
The webforum dataset contains the data of Thread ID, Author ID, Date, Time, Word Count, Linguistic variables such as Analytic, Clout, Authentic, Tone, ppron, I, We, You, Shehe, They, Affect, Posemo, Negemo and Words such as Anx, Anger, Social Family, Friend, Leisure, Money, Relig, Swear and QMark. Before doing anything to the webforum dataset, it has 20,000 rows of data with 29 variables. The type of the variables Thread ID, Author ID and Word Count (WC) are Integer type, Date and Time are Factor type and the remaining variables are Number value type. I have used a few libraries to analyze the data such as library(ggplot2) which is used for data visualization, library(dplyr) is used for data wrangling, library(reshape2) is used to melt table, library(igraph) is used to plot graph with nodes and edges and library(igraphdata) is used for decompose time series. After tidying the webforum dataset such as removing the anonymous authors for example, remove the Author ID with -1 and remove the posts with no words. The webforum dataset is left with 19,023 rows of data with 29 variables.

Please do Zoom in the graph for a clearer look

Task-A Analysis

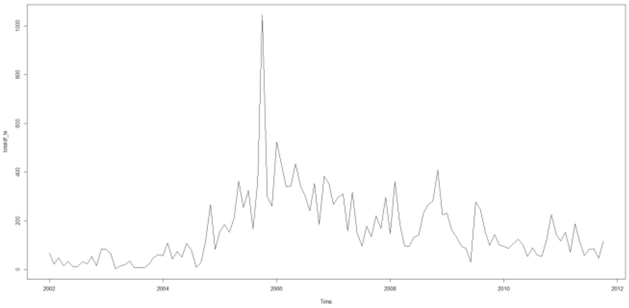
**** We first look at how active the participants are and is there a trend over time in the graph.**

Bar chart for total number of posts every month for each year from 2002 to 2011



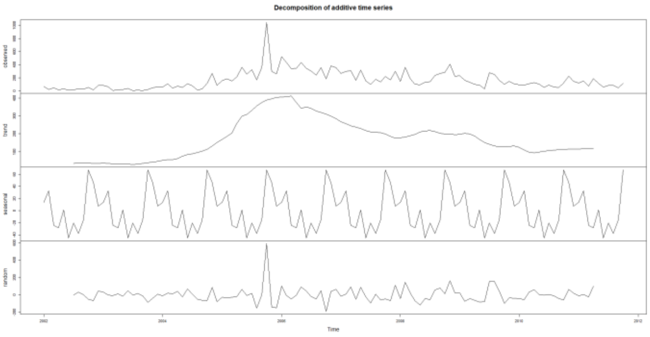
The graph above is a bar chart showing every month starting from 2002 January to 2011 December. The bar chart graph is produced by getting the total number of posts for each month of the year and plot it out according for each year. Each bar in the facet is representing the total number of posts each month for that year so from the bar chart graph, we can see that December 2005 have the highest number of posts compared to the years from 2002 to 2011. 2002 do not have any posts posted by the user which means that the forum is not that active at that time. However, the number of posts in the forum slowly increases from 2002 to 2006. The total number of posts decreases from 2008 to 2011.

Time Series Graph for total number of posts in the forum from 2002 to 2011



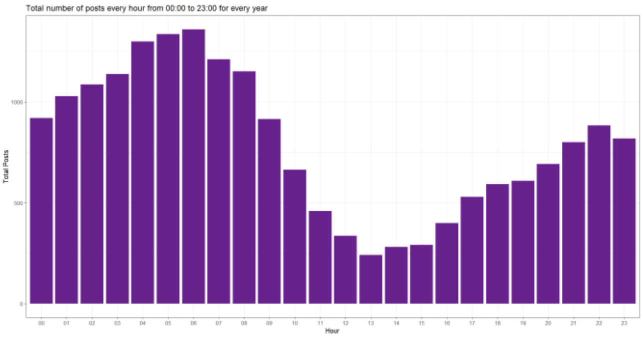
Next, we look at the time series of the graph from 2002 to 2011. This time series measures how active the participants are throughout every year starting from 2002 to 2011. The x-axis will be measuring the Date and the y-axis will measure the total number of posts posted by all the users in that one-year time. This time series graph is produced by getting the total number of posts in the forum each month of the year from 2002 to 2011 and plot it out in time series form. We can see that the time series graph increases from 2002 to the end of 2005 especially we can see there is a spike increase during the year 2005 and the peak is at the end of 2005. This shows that many people are using the forum to post in 2005. The total number starts to drop at the start of 2006 and it just slowly decreases until 2011. There is a small increase at the start of 2008, but it decreases back at the start of 2009. From this time series graph, we can observe an uptrend from 2002 to 2006 and a downtrend until 2011.

Decomposed Time Series Graph for the total number of posts in the forum every year from 2002 to 2011



The graph above showed that the trend happening in the forum over time by decomposing the time series for the total number of posts in the forum. The main observation that we can see from the graph above is that there is a linear trend and peak at the end of 2005, then we can see an obvious downtrend until 2011. This shows that the number of posts increases from 2002 to the start of 2006 and decreases until 2011. The seasonal peak at the end of the year between October to January of every year. The seasonal trough (lowest point) happens between May to August of every year. By looking at the seasonal graph above, we can see that there is always a downtrend at the start of every year and it will have an uptrend around June or July of every year and the graph will always peak at the end of the year. This shows that user do not post much at the start of the year and tend to post more at the end of the year between the month September to January.

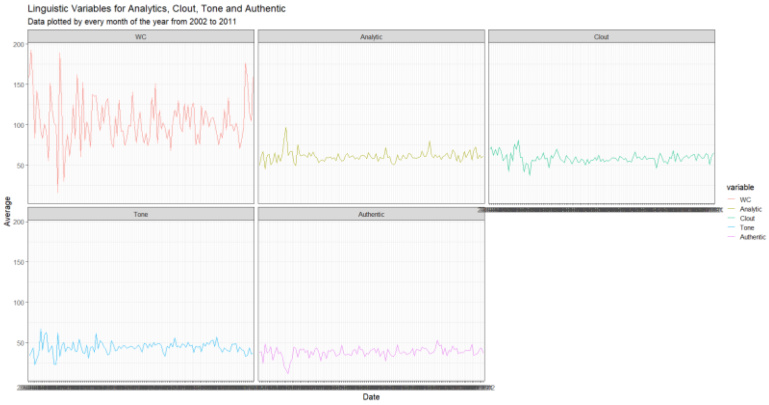
Bar chart for the total number of posts in the forum every hour from 00:00 to 23:00



The graph above showed that the total number of posts posted in the forum for 24 hours. From the graph above, we will be able to know at what time the user will be most active in the forum. From the graph, we can see that the average number of posts in the forum changes every hour and there is an increase in total number of posts which giving an uptrend starting from 00:00 am until 6:00 am. The bar chart graph peak at 6:00 am giving the greatest number of posts posted in the forum throughout the whole 24 hours. The total number of posts decreases from 6:00 am onwards until 13:00 pm which reaches the lowest number of posts posted in the forum. Then, the graph increases back until night which is 22:00 pm. This showed that most of the users tend to post on the forum between night to midnight time which is from 01:00 am to 06:00 am. User do not post much on the forum during afternoon. We can conclude that the users are most active during midnight and least active during afternoon time.

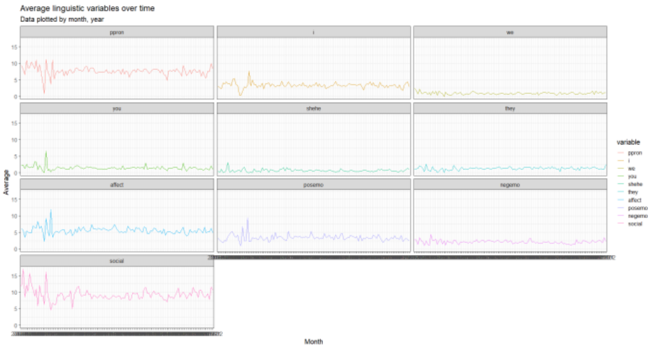
**** We will now look at the linguistic variables over different period of time and see if it changes over time. We will also be looking at if there's any relationship between the variables.**

Graph for the top 5 variables over the time period of every month of every year



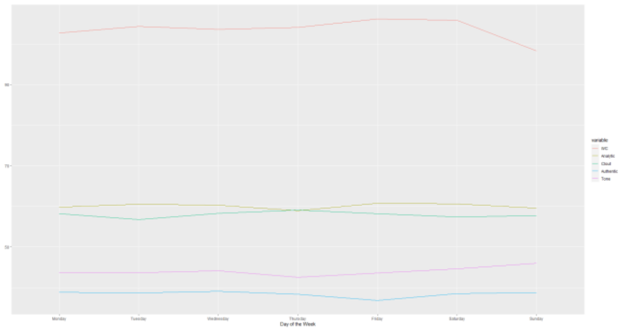
From the graph above, we get the top 5 variables, WC, Analytics, Clout, Authentic and Tone over the time period of every month for every year. We can see that there is an obvious downtrend line in WC at the start of the year and followed by a drastic and obvious uptrend line. We can say that there is an obvious change of the WC variable over the time period as the line in the graph keeps having a drastic increase and decrease overtime. Next, for the variable Analytics, we can see that there is an obvious increase in the Analytic variable at the start of the graph and it starts dropping drastically right after it reaches it's peak. This means that the user's post have an increase in analytical thinking at the start of the graph and decreases after it reaches the peak. Then, it does not have much changes after the drastic decrease in the value of Analytics variable. For the remaining 3 variables Clout, Tone and Authentic, do not have an obvious trend or changes over the time period.

Graph for the selected linguistic variables over the time period of every month of every year



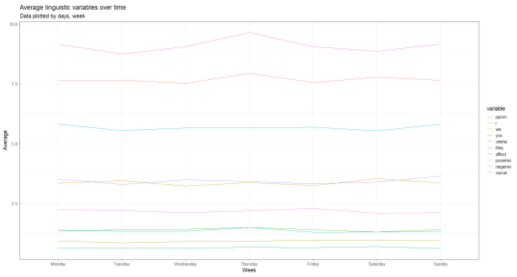
From the graph above, we get the selected 10 variables over the time period of each month for the whole year from 2002 to 2011. The reason why other linguistic variables are not used is because it does not have a trend over time and the changes over time is not that obvious enough to be observed. Based on the graph above, there are not much of an obvious change in the trend for some of the variables such as we, you, shehe, they and negemo over the period of time. We can see that the ppron variable and I variable in the graph above has decreased at the start of the graph and followed by an increase in the average value and did not change much over the remaining time period which means that user tend to use lesser personal pronouns and first person singular word at the start of the time period and it increases by a little throughtout the graph and did not change much after the increase. We can see that there is a small changes for affect and posemo variable at the start of the graph. Both the variables have a short increase in it's value and then an obvious decrease in the graph followed by an obvious increase in the value in the first few years. Next, we can see an obvious change in social variable which has an overall of downtrend over the first few years and a sharp increase until it reaches it's peak and decrease back. After it decreases, the average value of words referring to social processes slowly increases. This said that at the start of the few years, many users use words that referred to social processes and afterward it has a huge drop for few months and it went back up.

Graph for the top 5 variables over the time period of every day of the Week



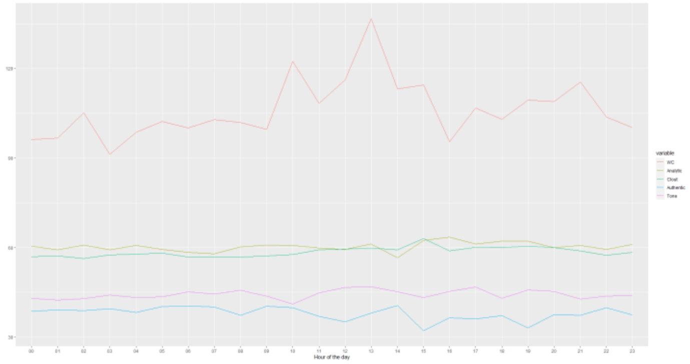
From the graph above, we get the top 5 variables, WC, Analytics, Clout, Authentic and Tone over the time period of every day for every week. We can see that there is not much changes in WC variable from Monday to Saturday and an obvious downtrend line on Sunday. This shows that user do not write many words in their posts during Sunday. For the remaining 4 variables Analytics, Clout, Tone and Authentic, do not have an obvious changes over the time period. For the Analytics variable, we can observe the the graph has a slight downtrend from Monday to Thursday and an uptrend from Thursday to Sunday. From the graph, we can see that when Clout variable reaches it's peak on Thursday, Analytic variable reaches it's lowest point too on that day. This means that user uses more power, impact and force words in their post on Thursday than analytical thinking words. For both the Tone and Authentic variables decreases over the weekday and start to increase when it's near the weekend.

Graph for the selected linguistic variables over the time period of the Day of the Week



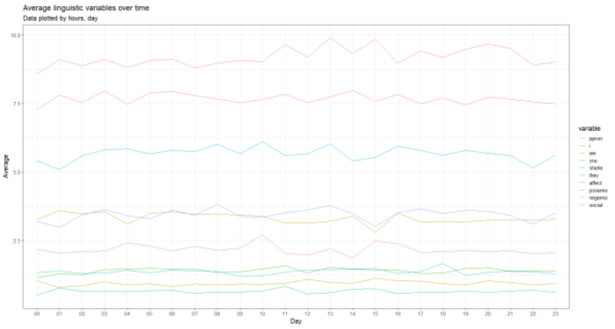
From the graph above, we get the remaining 20 variables time period of every day for every week. There are not much of an obvious change in the trend for most of the variables. And therefore I have chose a few linguistic variables to analyse because the other variables do not have a trend over time and the changes over time is not that obvious enough to be observed.We can see there are some obvious changes in social and ppron variable as the value keeps varying everyday for the whole week. We can say that user uses the most words referring to social processes and personal pronouns on Thursday for both the social and ppron variable. We can also observed some changes in I and posemo variable as the value changes every day. The highest peak for user to use words with positive emotions are on Sunday and Saturday for the words with first person singular. The remaining variables do not have an obvious trend to be observed over the whole week. The I and posemo variable fluctuate a little over the whole week where it keeps increasing and decreasing by a little everyday. However, it does not change much overtime although the average value changes everyday within the week.

Graph for the top 5 variables over the time period of Hour of the Day



From the graph above, we get the top 5 variables, WC, Analytics, Clout, Authentic and Tone over the time period of hour of the whole day. We can see that the word count of the text of the post increases from 00:00 am to 02:00 am which means that people tend to type more texts in their post during that time. We can see an uptrend of the graph from 03:00 am to 13:00 pm and the peak is at 13:00 pm which means that the word count of the text of the post is highest at 13:00 pm. During this period, the word count of the text in the post increases. After 13:00pm, the word count starts to drop which means that people do not write as much word in the post as before. For the other 4 variables, we can see that there are not much obvious changes in Analytic, Clout, Authentic and Tone throughout the whole 24 hours time period. However, we can see that there is an obvious change at 14:00 pm where the Analytic variable and Clout variable increased till 15:00 pm and Authentic variable has an obvious decrease from 14:00 pm to 15:00 pm and followed by an increase on the following hour.

Graph for the selected variables over the time period of Hour of the Day

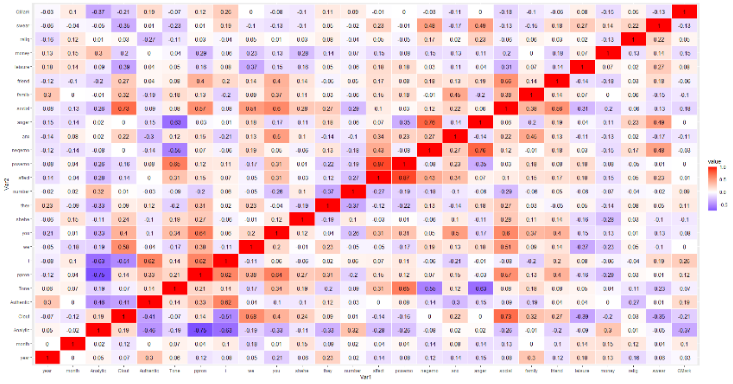


From the graph above, we get some of the selected variables over the time period of hour of the whole day. The reason why I didn't choose the remaining variable is because there are not much of an obvious change in the trend for most of the variables and some of the value of the variables are too small and di not change much over time. We can see that the social variable in the graph has slowly incraesed over the time period from 00:00 am to 15:00 pm which means that user tend to use more social word over time. There is a drastic decrease at 15:00 pm for social variable until 16:00 pm and the trend until 23:00 pm is overall a downtrend. We can see that the affect variable decreases from 00:00 am to 01:00 am and have an uptrend line until 10:00 am which means during this period, the user tend to express their sentiments at that period of time. There's not much change can be observed after 10:00 am from the graph for affect variable. We can also see the posemo variable decrease from 00:00 am to 01:00 am and have an uptrend line until 13:00 pm which means the user are expressing more words with positive emotions during this period of time. There's not much change can be observed after 13:00 pm from the graph for posemo variable. For the remaining of the variables, there are no obvious change can be seen from the graph as most of them does not vary much over the time period of 24 hours.

Conclusion for Task A

In conclusion, we can say that the participants are most active in the year 2005 and they are more active during midnight time from 00:00 am to 6:00 am morning. The variable social, affect, posemo, ppron,WC and i do have an obvious change over time as it can be seen from the graph that it changes for different period of time such as in every month of the year, every day of the week and every hour of the day graph.

Heatmap for all the variables in the dataset



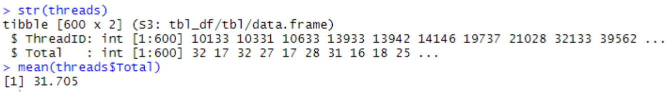
We can use heatmap to find the correlation between all the different variables. From the heatmap graph above, the boxes with the colour closest to red or value of 1 means that they have a strong relationship between the two variables. Vice versa, the boxes that is closest to blue colour or value of 0 means that it has no relationship between the two variables. By looking at the graph above, we can see that there are a couple of variables have a strong to moderate relationship between each other. The values below shows how strong the relationship of the variables between them are:

1. Posemo & Affect: 0.87
2. Anger & Negemo: 0.76
3. Social & Clout: 0.73
4. Posemo & Tone:0.65
5. You & Ppron: 0.64
6. Authentic & I: 0.62
7. Ppron & I: 0.62
8. You & Social: 0.60

As all the values of the paired variables above are having a value closer to 1. Therefore, we can conclude that these pair of variables have a relationship between them.

Task B Analysis

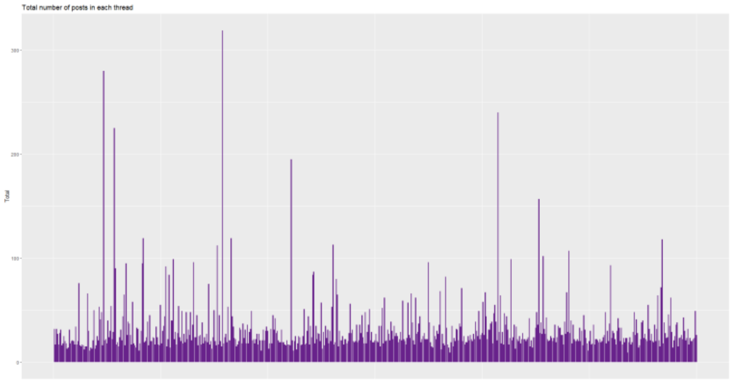
**** Next, we will move on and focus more on the thread by analysing the thread ID.**

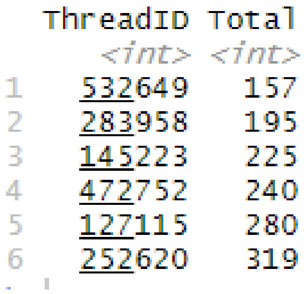


```
> str(threads)
tibble [600 x 2] (S3: tbl_df/tbl/data.frame)
 $ ThreadID: int [1:600] 10133 10331 10633 13933 13942 14146 19737 21028 32133 39562 ...
 $ Total   : int [1:600] 32 17 32 27 17 28 31 16 18 25 ...
> mean(threads$Total)
[1] 31.705
```

By performing some basic analysis of the thread present in the data, I created a new dataset and store the total number of posts in one unique thread by grouping the Thread ID together. There are total of 600 unique different thread ID in the dataset. The average posts in each thread is 31.705 that rounds up to 32 posts. The average posts are gotten by using the mean function which sums up the total number of posts in each thread and divide by the total number of unique thread giving 31.705.

Bar chart for total number of posts in every thread

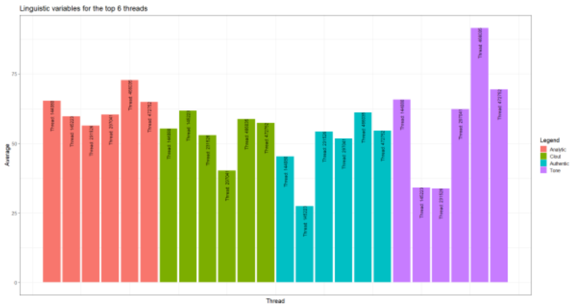


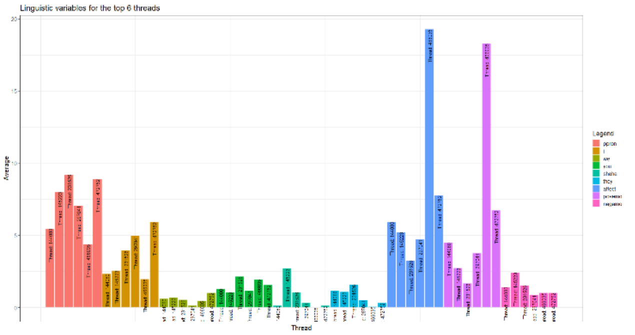

```
  ThreadID Total
     <int> <int>
1   532649   157
2   283958   195
3   145223   225
4   472752   240
5   127115   280
6   252620   319
```

From the bar chart above, we can't see that there is an ovboius trend in the total number of posts for every thread as some of the threads increases and some decreases without having a trend. However, we can see that there are a few threads that have really long lines in the bar chart which means that it has a high number of posts in the thread. Therefore, I analysed the top 6 threads in the forum for further analysis to see if there's trend within the top 6 threads. I sort the data set in ascending order for the Total number of Posts. From the table above, we can see that the top 6 Threads with it's thread id and the total (total number of posts) in ascending order. We will know that the thread id 252620 has the most number of posts compared to other threads in the forum.

****Next, we will move on to analyse the linguistic variables for all and some of the selected threads and see if the language changes in different groups of thread.**

Bar chart for the top 6 threads in the forum with all the linguistic variables





Based on the graph above, we are comparing the top 6 threads with all the linguistic variables to see if there's a change in the language in different groups. Each bar with the same represents a different thread. The reason why I chose these variables to analyse if there's a change in the language becaues first, all these variables have a higher value compared the remaining variables that I did not choose. I have choesn all the linguistic variables in the dataset except for those words that indicate some expression or emotions because the remaining variables do not have much changes between the 6 threads group and therefore I exclude it from the graph. I got this graph by first finding the top author in the dataset which means that this author will be psoting the most number of times in the forum. Next, I will find the thread where this author posted the most in and get all the thread ID in this thread. I will then take the top 6 threads from this thread dataset and group the original dataset with these top 6 threads. Next, I will get the average mean of all the variables in the dataset by grouping the thread id together. Finally, I just plot the 14 linguistic variables which is the selected variables for the top 6 threads on the bar chart producing the graph shown above. From the graph above, let's start bycomparing the third thread 231526 and fifth thread 466035. There is a really huge difference in Analytics, tone, ppron, affect and posemo where we can see that the Thread 466035 have a really high average value in tone, ppron, affect and posemo compared to Thread 231526. Therefore, we can say that there's a difference in the language in different groups of thread as the people in Thread 466035 uses a lot of words with analytical thinking, emotional tone, words with expressing sentiments and positive emotions compared to groups of people in Thread 231526 who did not use those word that often. On the other hand, the group of people in Thread 231526 uses more personal pronouns, First person singular, First person plural, Second person, Third person singular, Third person plural and words with negative emotions. This proves that the group of people in Thread 466035 are more analytical and emotional and also more positive than the groups of people in Thread 466035 because the people in thread 231526 uses many negative emotions to convey their message in the forum. People in the thread 466035 tend to use words with power, force and impact and also speak with an authentic tone of voice in the forum compared to the groups in thread 231526 because it has higher average mean value in Clout and Authentic variable than thread 466035. Therefore, we can conclude that the thread 231526 and 466035 have difference in the language used as it can be seeing clearly in the graph that most of the variables have a huge difference between the two threads.

In order to prove that the analysis above is true, I have performed hypothesis testing with all the variables that I used to plot the bar chart above. The hypothesis testing will be using the t test function to check if the groups are equal to each other.

Hypothesis testing: $H_o : u_1 - u_2 = 0$

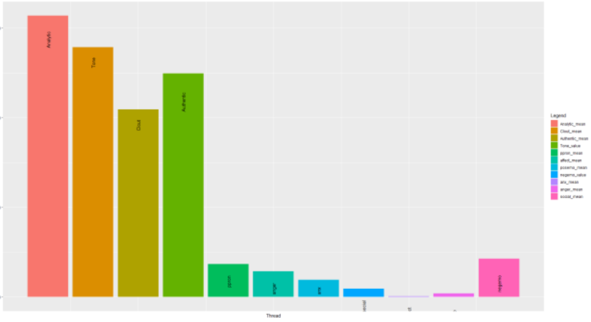
$$\text{vs}$$

$$H_A : u_1 - u_2 \neq 0$$

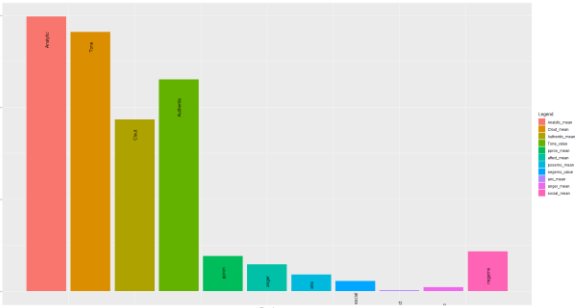
Linguistic variables	P-value
Analytic	0.01069
Clout	0.31403
Authentic	0.37355
Tone	1.32209e-12
Ppron	4.10514e-05
I	0.01574
We	0.77143
You	0.76153
Shehe	0.00814
They	2.80356e-05
Affect	1.07960e-12
Posemo	2.40537e-14
Negemo	0.418414

From the p-value table above, I used t test function to find the mean difference of all the linguistic va riables between the two different threads. The smaller the P-value, suggests that it has a strong evid ence against the null hypothesis which is saying that there is a difference in mean. When the P-value is smaller than 0.05, it indicates a strong evidence against null hypothesis which means that the null hypothesis can be rejected. In the table above, the variable Analytic, Tone, Ppron, I, Shehe, They and Posemo has a value of 0.01069, 1.32209e-12, 4.10514e-05, 0.01574, 0.00814, 2.80356e-05, 1.07960 e-12 and 2.40537e-14 respectively which is lesser than 0.05. This has a strong evidence against the null hypothesis which means that it supports the difference in variables hypothesis. This shows that there is a high probability of having a difference in means. As most of the variables have strong evidence against the null hypothesis compared to only 5 variables, Clout, Authentic, We, You and negemo have a weak evidence against the null hypothesis and therefore we can conclude that the two threads 231526 and 466035 have different language used.

<u>Barchart for the mean value of the selected linguistic variables between the top 20 threads and the bottom 20 threads:</u>



Top 20 Threads



Bottom 20 Threads

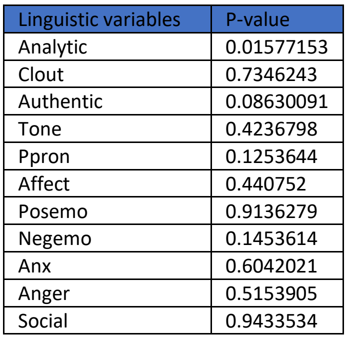
Based on the graph above, the left side of the graph showed the mean value of the selected linguistic variables for the top 20 threads and the right side of the graph showed the mean value for the bottom 20 threads in the dataset. Each bar represents the average mean value for the variable for each 20 threads for both top and bottom threads. The reason why I didn't choose all the variables are because some of the variables gives a really small mean value which does not show anything on the graph and therefore I have picked those variables that can see a difference in the mean value. The reason why I choose posemo, negemo, anx and anger variable although the mean value is small too is because by looking at how each thread reacted to positive or negative emotions by using words indicating anxiety and anger can tell the difference between the language used. First, I analyse the top 20 and bottom 20 threads in the dataset and get the mean value of all the selected variables I have chosen by grouping the top 20 thread ID together and store it intoa dataset. I will do the same thing again but this time I group the bottom 20 thread ID instead of the top 20 thread ID. Finally, I will plot the graph for both the mean value of the selected variables for both the top and bottom 20 threads and compare the value of the selected variables to see if there's any difference between these two groups. Let's compare the mean value of the both the top and bottom 20 threads graph. We can see that both the graphs are kind of similar as all the mean values in the variables are almos the same. Most of the variables only change with the mean values of plus/minus of 5. Therefore, we can conclude that the top 20 and bottom 20 threads used similar language for these two different groups of thread group because the mean value or the graph did not varied that much.

In order to prove that the analysis above is true, I have performed hypothesis testing with all the variables that I used to plot the bar chart above. The hypothesis testing will be using the t test function to check if the groups are equal to each other.

Hypothesis testing: $H_o : u_1 - u_2 = 0$

vs

$H_A : u_1 - u_2 \neq 0$

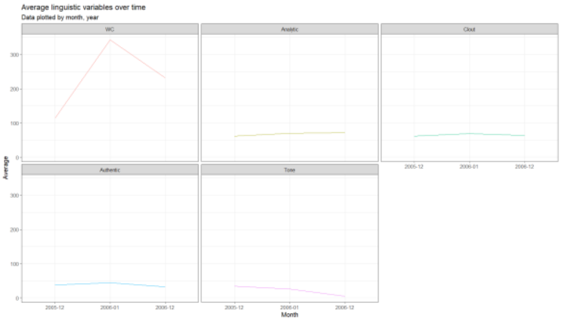


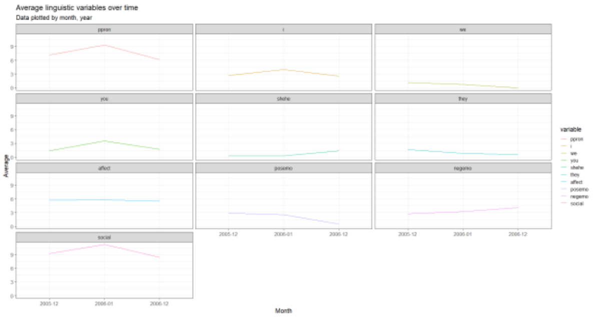
Linguistic variables	P-value
Analytic	0.01577153
Clout	0.7346243
Authentic	0.08630091
Tone	0.4236798
Ppron	0.1253644
Affect	0.440752
Posemo	0.9136279
Negemo	0.1453614
Anx	0.6042021
Anger	0.5153905
Social	0.9433534

From the p-value table above, I used t test function to find the mean difference of all the linguistic variables between the two different threads. The smaller the P-value, suggests that it has a strong evidence against the null hypothesis which is saying that there is a difference in mean. When the P-value is smaller than 0.05, it indicates a strong evidence against null hypothesis which means that the null hypothesis can be rejected. In the table above, the variable Clout, Authentic, Tone, Ppron, Affect, Posemo, negemo, anx, anger and social has a value of 0.7346243, 0.08630091, 0.4236798, 0.1253644, 0. 440752, 0.9136279, 0.1453614, 0.6042021, 0.5153905 and 0.9433534 respectively which is more than 0.05. This has a weak evidence against the null hypothesis which means that it does not support the difference in variables hypothesis. This shows that there is a high probability of having a same in means. As most of the variables have weak evidence against the null hypothesis compared to only 1 variable, Analytic has a strong evidence against the null hypothesis which is 0.01577153 which is lesser than 0.05. This shows that Analytic variable is the only one has a strong evidence against the null hypothesis that shows there is a high probability of having a difference in means. Therefore we can conclude that the top 20 and bottom 20 threads used similar languages as most of the variables analysed do not have strong evidence against the hypothesis proving that the two different threads, one with the top 20 threads and another one with the bottom 20 threads do not have different language

**** Next, we will be looking at the the language used in the thread and see if it changes over time**

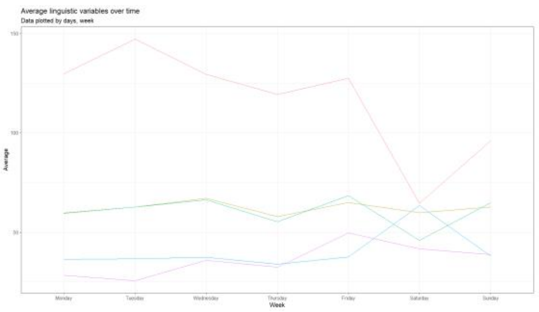
<u>Graph for the mean value of the selected linguistic variables for the top thread over the time period of month, year</u>

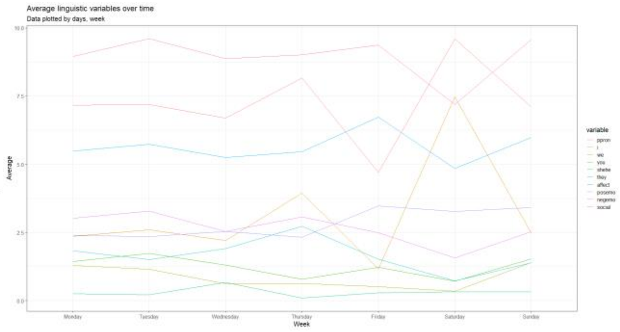




From the graph above, we are looking at the selected variables for the top thread in the dataset. This graph above analyzes the variables in the thread over every month for a year from 2002 to 2011. The reason why I didn’t choose the remaining variables is because there’s not much trend in change in language in the thread for those variables. Some of the mean value of those variables are too small to analyze too. Therefore, I only picked these variables, WC, Analytic, Clout, Tone, Authentic, Ppron, I, we, you, shehe, they, affect, posemo, negemo and social. From the graph above, we can see that WC variable has a huge increase and decrease back down when it reaches the peak. We can say that the word count of the texts in the thread increases over time and slowly decrease back. There is not much trend in the Analytic, Clout, Authentic and affect as it doesn’t change much over time. We can clearly see that the mean value of Tone, we, they and posemo decreases overtime as it is a downtrend in the graph. This means that people started to use lesser words with emotional tone, first person plural, third person plural and positive emotions over time in the thread. On the other hand, we can see that the variable negemo and shehe has a uptrend in the graph which means that the mean value increases and more people uses words with third person singular and negative emotions word in the thread. Next, for the variable I, you and social increases first in December 2005 and decreases back on the following month. In conclusion, we can say that although some of the variables changes over the time period of the two months but there’s no strong enough evidence to show that the language changes overtime.

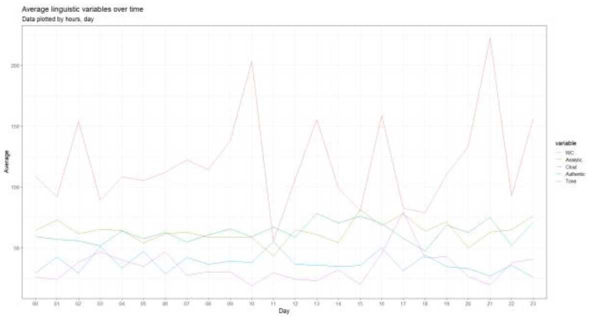
Graph for the mean value of the selected linguistic variables for the top thread over the time period of day, week

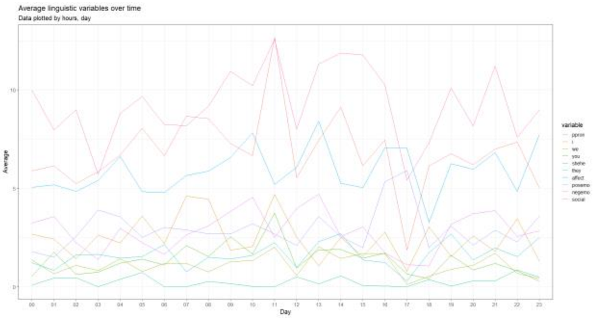




From the graph above, we are looking at the selected variables for the top thread in the dataset. This graph above analyzes the variables in the thread over every day for a week from 2002 to 2011. The reason why I didn’t choose the remaining variables is because there’s not much trend in change in language in the thread for those variables. Some of the mean value of those variables are too small to analyze too. Therefore, I only picked these variables, WC, Analytic, Clout, Tone, Authentic, Ppron, I, we, you, shehe, they, affect, posemo, negemo and social. From the graph above, we can see that the WC variable has the highest mean value on Tuesday and lowest on Saturday which means that people tend to type more words in the thread post on Tuesday and least words on Saturday. We can see an obvious trend that the mean value of the WC variable to increase from Monday to Tuesday and slowly decreases until Saturday which can see an obvious downtrend in the graph. We can see that Clout, Authentic and Tone increases over time where it slowly increases throughout the whole week and until Friday or Saturday, it starts to decrease in the mean value of Clout, Authentic and Tone. For Authentic variable, there’s a spike increase from Friday to Saturday which means that many people use analytical thinking words in the posts in the thread. In overall, we can see that there’s a change in language overtime for example, at the start of the week, people tend to speak with less analytical thinking, power, force, impact, authentic tone of voice, emotional tone, personal pronouns, words with social processes, positive emotions and words with expressing sentiments which means that they are more chill and talks with less positive emotions. As the time passes by, the mean of those variables increases throughout the week until Thursday or Friday and it starts to decrease back. This can say that people will get more serious as the week passes and used more authentic voice of tone, analytical thinking, use more words with positive emotions and when it reaches the end of the week, they will start to use back chiller words and lesser analytical thinking and power words in the posts. The negative emotions in the posts decreases over the whole week which means that people tend to be more negative at the start of the week and start to feel happier at the end of the week as the negative emotions in the posts decreases. Therefore, this proves that there’s a change in the language used over time as in the top thread, the variables do change as the day passes such as author’s tone, emotions and the words used changes which lead to change in language used.

Graph for the mean value of the selected linguistic variables for the top thread over the time period of hour, day

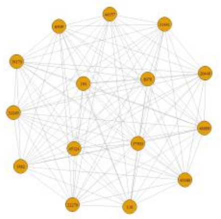




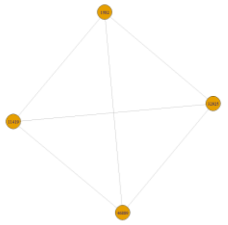
From the graph above, we are looking at the selected variables for the top thread in the dataset. This graph above analyzes the variables in the thread over every hour for a day from 2002 to 2011. The reason why I didn't choose the remaining variables is because there's not much trend in change in language in the thread for those variables. Some of the mean value of those variables are too small to analyze too. Therefore, I only picked these variables, WC, Analytic, Clout, Tone, Authentic, Ppron, I, we, you, shehe, they, affect, posemo, negemo and social. From the graph above, we can clearly see that the word count of the words in the posts increases overtime until 10:00 am and it starts to decrease and increase for a few consecutive hours. The word count of the words in the posts peak at 21:00 pm which means that people tend to type more at night than during afternoon. We can also clearly see that many variables such as social, I and posemo decrease at the start of the first few hours from 00:00 am to 03:00 am and it increase back until 11:00am. We can see that many variables peaked at 11:00 am such as social, ppron, I and you. After this, these variables start to keep decreasing in the mean value. The mean value for negemo and affect starts to increase from the start of the hour 00:00 am to 13:00 pm although it increases and decrease over that period but it's an overall of uptrend curve. After 13:00 pm, the men value for both the negemo and affect starts to decrease until 18:00 pm to increase back till midnight. Therefore, we can say that the language within the threads do change overtime such as at 11:00 am, authors in the thread tends to use more words that express sentiments, personal pronouns, first person singular and first-person plural compared to other time period which have a lower mean in those variables. Therefore, we can conclude that the language used at the start, middle and the end of the 24 hours period are different proves that the language used within the threads change over time.

Task C Analysis

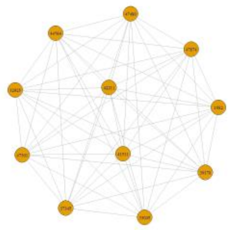
**** We now look at the groups of participants in a thread and group them together to form a social network.**



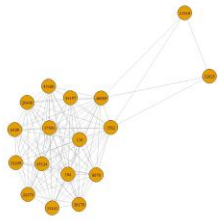
Graph 1: Thread 145223



Graph 2: Thread 104567

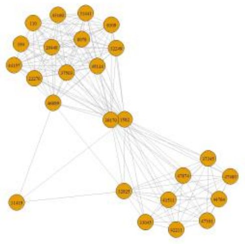


Graph 3: Thread 150445



Merge graph 1 & graph 2:

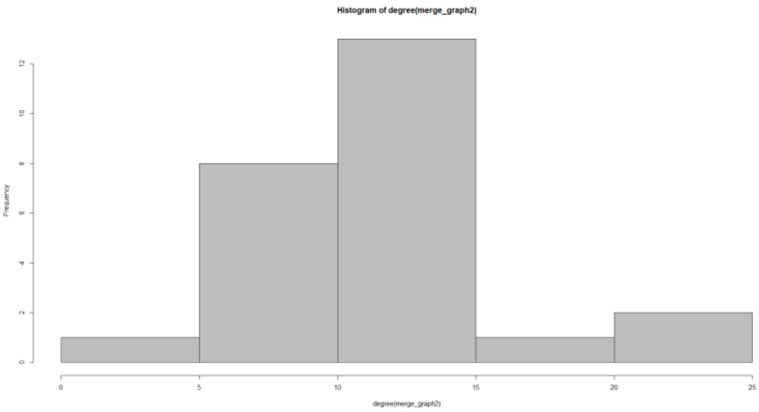
I get the top author from the dataset which is the author that posted the greatest number of times in the forum and find the thread where the author posted the most in. Next, I will group the dataset with thread id and month and year. I choose the month and year in the dataset that have the greatest number of threads in it. Finally, I chose 3 threads from the dataset to plot a graph to form a social network. These authors will be group together when the same author posted on another thread and it will extend the social network. I plot all the graphs by using adjacency matrix function. I plot thread 145223 in graph 1 and thread 104567 in graph 2 and thread 150445 in graph 3. The fourth graph is a merge graph of graph 1 and graph 2. The author 46889 and 1582 appear in both the thread and therefore, it connects the two graphs together which extends the social network.



Merge 1 & 2 & 3

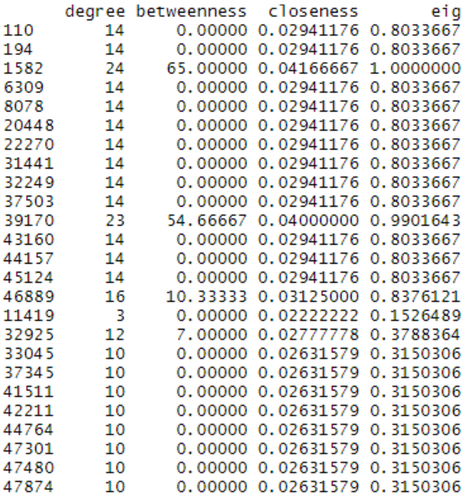
The graph above shows the merge graph of graph 1, graph 2 and graph 3. From the above graph, we can see that the thread 39170 and 1582 have a connection in all the 3 threads and the thread 32925, 46889 have connection with two graphs but on different thread group. In conclusion, we can see that the thread 1582 has connection with the other two threads causing it to be the most important person in the social network.

The diagram below shows the histogram of the degree distribution



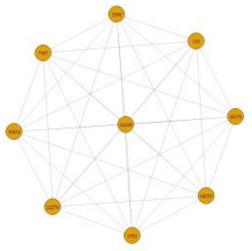
The histogram above shows the degree distribution of the merge graph for the 3 combined threads together. It shows that the degree from 10 to 15 have the highest frequency of degree distribution and both the degree from 0 to 5 and 15 o 20 have the lowest frequency of degree distribution.

The table showing the degree, betweenness, closeness and eigenvector for all

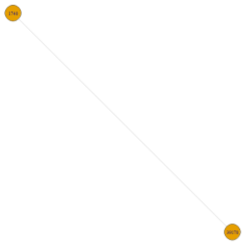


```
      degree betweenness  closeness       eig
110       14     0.00000 0.02941176 0.8033667
194       14     0.00000 0.02941176 0.8033667
1582      24    65.00000 0.04166667 1.0000000
6309      14     0.00000 0.02941176 0.8033667
8078      14     0.00000 0.02941176 0.8033667
20448     14     0.00000 0.02941176 0.8033667
22270     14     0.00000 0.02941176 0.8033667
31441     14     0.00000 0.02941176 0.8033667
32249     14     0.00000 0.02941176 0.8033667
37503     14     0.00000 0.02941176 0.8033667
39170     23    54.66667 0.04000000 0.9901643
43160     14     0.00000 0.02941176 0.8033667
44157     14     0.00000 0.02941176 0.8033667
45124     14     0.00000 0.02941176 0.8033667
46889     16    10.33333 0.03125000 0.8376121
11419      3     0.00000 0.02222222 0.1526489
32925     12     7.00000 0.02777778 0.3788364
33045     10     0.00000 0.02631579 0.3150306
37345     10     0.00000 0.02631579 0.3150306
41511     10     0.00000 0.02631579 0.3150306
42211     10     0.00000 0.02631579 0.3150306
44764     10     0.00000 0.02631579 0.3150306
47301     10     0.00000 0.02631579 0.3150306
47480     10     0.00000 0.02631579 0.3150306
47874     10     0.00000 0.02631579 0.3150306
```

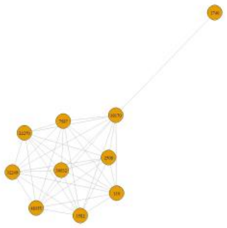
From the table above, we can see that the top 4 most important authors are 1582, 39170, 46889 and 32925 because it has the highest value on all measures. The author 1582 has degree of 24, betweenness of 65, closeness of 0.04166667 and eigenvector of 1 which will be the most important person in the network. The second most important person is author 39170 has degree of 23, betweenness of 54.66667, closeness of 0.04 and eigenvector of 0.9901643 followed by author 46889 has degree of 16, betweenness of 10.33333, closeness of 0.03125 and eigenvector of 10.8376121. It can be seen from the graph above that the author 1582 has the most connections with the other authors in the other threads.



Graph 1: Thread 145223



Graph 3: Thread 150445



Merge 1 & 2 & 3

We used the same threads and analyze it for the next month. I plot thread 145223 in graph 1 and thread 150445 in graph 2. The reason why I didn't plot Graph 2 for Thread 104567 is because there's no thread 104567 in the following month. The third graph is a merge graph of graph 1 and graph 3. The author 39170 appears in both the thread and therefore, it connects the two graphs together which extends the social network. We can say that the author 39170 is the most important person in the network.

Conclusion

In conclusion, I have analyzed the webforum data and found out some trends and answers from the graphs and data. For example, I have found out that at which time of the day or which day of the week users are most active in. I also analyzed that the types of words user like to use at different time. Next, I have found out more information about threads and does the language change over different group of threads. I also look at how the language in the thread changes over time by performing some analysis. Finally, I grouped the threads together by plotting a graph and analyze how each participants extend their social network when they communicate on another threads by using social network analysis.

Appendix

R-Code

```
##################### Library

library(ggplot2) #data visualization such as plotting the graphs

library(dplyr) #data wrangling

#FOr it to use melt function

library(reshape2)

library(igraphdata)

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


#create our individual data and tidy the data by removing unnecessary data.

rm(list = ls())

set.seed(29634431) # 29634431 = your student ID

webforum = read.csv("webforum.csv")

webforum = webforum [sample(nrow(webforum), 20000), ] # 20000 rows

# Remove anonymous authors

webforum = subset(webforum, AuthorID != -1)

# Remove posts with no words

webforum = subset(webforum, WC != 0)


####Task A part1


##Analyse the forum over time by looking at how active the participants is over time.


attach(webforum)


# Format date string to a Date data type

webforum$Date = as.Date(webforum$Date)

webforum$my = format(webforum$Date, "%Y-%m")

#extract the year

webforum$year = as.numeric(format(webforum$Date, "%Y"))
```

```
#extract the month
webforum$month = as.numeric(format(webforum$Date, "%m"))
webforum$day = weekdays(webforum$Date)
#extract time for hours
Time  = factor(webforum$Time)
Time2 = strptime(Time, "%H:%M")
webforum$hour = format(Time2, "%H")

#create a dataset that group the webforum data by month, year and Author ID with an extra column
of length of Author ID
# which means that it stores total number of posts on that month and year with the same Author ID.
Adf1 = webforum %>% group_by(month,year,AuthorID) %>% summarise(Total = length(AuthorID))

#Create a dataset that stores the total number of posts within the time frame of hours by grouping
with hour variables.
timedf = webforum %>% group_by(hour) %>% summarise(Total = length(AuthorID))

#Plot the barchart for the number of posts in the forum for year from 2002 to 2011.
ggplot(timedf, aes(x = hour, y = Total)) + geom_bar(stat = "identity", fill = "darkorchid4")


#Get the total number of posts in the forum and the total number of users using the forum for each
month of the year from 2002 to 2011.
Adf2 = webforum %>% group_by(month,year) %>% summarise(TotalPost = length(AuthorID))
#Sort the dataset in years starting from 2002 to 2011.
Adf2 = Adf2[order(Adf2$year),]


#Plot the barchat for the number of posts in the forum for every month of the year from 2002 to
2011.
ggplot(Adf2, aes(x = month, y = TotalPost)) + geom_bar(stat = "identity", fill = "darkorchid4") +
facet_wrap(~ year, ncol = 3) +
  labs(title = "Total number of posts every month for each year from 2002 to 2011", subtitle = "Data plotted by year", y = "Total Posts",
       x = "Month") + theme_bw(base_size = 15)

#Plot the time series of the total number of posts posted by the participants on the forum over the period of time from 2002 to 2011.

Adf2_ts = ts(Adf2$TotalPost, frequency = 12, start = c(2002,1))

plot(Adf2_ts)



#decompose the time series for the total number of posts posted by the participants on the forum over the time from 2002 to 2011

decompts = decompose(Adf2_ts)

plot(decompts)




####Task A part2


##Analyse the selected linguistic variables over time and analyse the relationship between them


summary(webforum)


#Month, Year function for the top 5 variables. MONTHLY TREND FOR A YEAR

get_my_graph1 = function(data) {

  #take out the variables/columns that we want to look at which is the top 5 variables in the data

  data1.1 = data[ , (names(data) %in% c("Group.1","Analytic", "WC", "Clout", "Authentic", "Tone"))]

  #melt the table

  melted_data = melt(data1.1, id="Group.1")


  final_data = ggplot(melted_data, aes(x = Group.1, y = value, colour=variable, group=variable)) + geom_line() + facet_wrap(~ variable, ncol = 3) +

    labs(title = "Average linguistic variables over time", subtitle = "Data plotted by month, year", y = "Average",
        x = "Month") + theme_bw(base_size = 15)
  return(final_data)
}



#Month, Year function for the remaining variables. MONTHLY TREND FOR A YEAR
get_my_graph2 = function(data) {
  data1.1 = data[ , (names(data) %in%
c("Group.1","ppron","i","we","you","they","shehe","affect","posemo","negemo","social"))]
  #melt the table
  melted_data = melt(data1.1, id="Group.1")
  final_data = ggplot(melted_data, aes(x = Group.1, y = value, colour=variable, group=variable)) +
geom_line() + facet_wrap(~ variable, ncol = 3) +
    labs(title = "Average linguistic variables over time", subtitle = "Data plotted by month, year", y =
"Average",
        x = "Month") + theme_bw(base_size = 15)
  return(final_data)
}



#Day,Week function for the top 5 variables. DAILY TREND FOR A WEEK
get_day_graph1 = function(data) {
  #take out the variables/columns that we want to look at which is the top 5 variables in the data
  data1.1 = data[ , (names(data) %in% c("Group.1","Analytic", "WC", "Clout", "Authentic", "Tone"))]
  #melt the table
  melted_data = melt(data1.1, id="Group.1")
  #Plot the barchat for the number of posts in the forum for every day of the week from 2002 to
2011.
  final_data = ggplot(melted_data, aes(x = Group.1, y = value, colour=variable, group=variable)) +
geom_line() +
    labs(title = "Average linguistic variables over time", subtitle = "Data plotted by days, week", y =
"Average",
        x = "Week") + theme_bw(base_size = 15)
  return(final_data)

}



#Day,Week function for the remaining variables. DAILY TREND FOR A WEEK

get_day_graph2 = function(data) {

  data1.1 = data[ , (names(data) %in%
c("Group.1","ppron","i","we","you","they","shehe","affect","posemo","negemo","social"))]

  #melt the table

  melted_data = melt(data1.1, id="Group.1")

  #Plot the barchat for the number of posts in the forum for every day of the week from 2002 to
2011.

  final_data = ggplot(melted_data, aes(x = Group.1, y = value, colour=variable, group=variable)) +
geom_line() +

    labs(title = "Average linguistic variables over time", subtitle = "Data plotted by days, week", y =
"Average",

         x = "Week") + theme_bw(base_size = 15)

  return(final_data)

}



#Hour,Day function for the top 5 variables. HOURLY TREND FOR A DAY

get_hour_graph1 = function(data) {

  #take out the variables/columns that we want to look at which is the top 5 variables in the data

  data1.1 = data[ , (names(data) %in% c("Group.1","Analytic", "WC", "Clout", "Authentic", "Tone"))]

  #melt the table

  melted_data = melt(data1.1, id="Group.1")

  #Plot the barchat for the number of posts in the forum for every day of the week r from 2002 to
2011.

  final_data = ggplot(melted_data, aes(x = Group.1, y = value, colour=variable, group=variable)) +
geom_line() +

    labs(title = "Average linguistic variables over time", subtitle = "Data plotted by hours, day", y =
"Average",
        x = "Day") + theme_bw(base_size = 15)
  return(final_data)
}



#Hour,Day function for the remaining variables. HOURLY TREND FOR A DAY
get_hour_graph2 = function(data) {
  data1.1 = data[ , (names(data) %in% c("Group.1","ppron","i","we","you","they","shehe","affect","posemo","negemo","social"))]
  #melt the table
  melted_data = melt(data1.1, id="Group.1")
  #Plot the barchat for the number of posts in the forum for every day of the week from 2002 to 2011.
  final_data = ggplot(melted_data, aes(x = Group.1, y = value, colour=variable, group=variable)) + geom_line() +
    labs(title = "Average linguistic variables over time", subtitle = "Data plotted by hours, day", y = "Average",
        x = "Day") + theme_bw(base_size = 15)
  return(final_data)
}

#create a dataset that stores all the mean of the variables in the dataset grouping by month and year
Adf3 = aggregate(webforum, by = list(webforum$my), mean)
Adf4 = aggregate(webforum, by = list(webforum$my), mean)

#Month,Year for the top 5 variables. The trend for the top 5 variables in 1 month for a year.
get_my_graph1(Adf3)
#Month,Year for the remaining variables. The trend for the remaining 20 variables in 1 month for a year.
get_my_graph2(Adf4)
```

```
#create a dataset that stores all the mean of the variables in the dataset grouping by month and year

Adf5 = aggregate(webforum, by = list(webforum$day), mean)

Adf6 = aggregate(webforum, by = list(webforum$day), mean)

Adf5$Group.1 = factor(Adf5$Group.1, levels = c("Monday","Tuesday","Wednesday","Thursday","Friday","Saturday","Sunday"))

Adf6$Group.1 = factor(Adf6$Group.1, levels = c("Monday","Tuesday","Wednesday","Thursday","Friday","Saturday","Sunday"))


#Day,Week for the top 5 variables. The trend for the top 5 variables in 1 day for a week.

get_day_graph1(Adf5)

#Day,Week for the remaining variables. The trend for the remaining 20 variables in 1 day for a week.

get_day_graph2(Adf6)


#create a dataset that stores all the mean of the variables in the dataset grouping by month and year

Adf7 = aggregate(webforum, by = list(webforum$hour), mean)

Adf8 = aggregate(webforum, by = list(webforum$hour), mean)


#Hour,Day for the top 5 variables. The trend for the top 5 variables in 24 hours for a day.

get_hour_graph1(Adf7)

#Hour,Day for the remaining variables. The trend for the remaining 20 variables in 24 hours for a day.

get_hour_graph2(Adf8)


#create a dataset that stores all the variables which needs to be used to analyse the correlations between two variables

df6 = webforum[ , !(names(webforum) %in% c("ThreadID", "AuthorID", "Date", "Time", "month","year","my","day","hour"))]

#create a heatmap to determine the correlations between different variables
```

```
matrix = round(cor(df6),2)

melted_matrix = melt(matrix)


#plot the heatmap out to visualise the correlations between two variables in colours and values.

#use value of 0 and 1 and colour blue, white and red to determine the strength of correlations between two variables.

ggplot(melted_matrix, aes(x = Var1, y = Var2, fill=value)) + geom_tile(colour="white") +

  geom_text(aes(Var1,Var2, label=value),colour= "black", size=4) +

  scale_fill_gradient2(low="blue", high="red", mid="white", midpoint=0)




####Task B

##Task B part 1


##Analyse the threads present in the data


#create a dataset that stores the total number of posts posted in each thread grouped by Thread ID

threads = webforum %>% group_by(ThreadID) %>% summarise(Total = length(ThreadID))

sorted_threads = threads[order(threads$Total),]

#check the type of the column and also how many rows and columns do the data have in total

str(threads)

#find the mean of the total number of posts in each unique thread using mean function

mean(threads$Total)


#to list out the number from 1 to the total number of rows of data so that it can be used to plot bar chart graph.

threads$Number = seq.int(1,nrow(threads))


#Plot the barchart for the number of posts in the forum for year from 2002 to 2011.

ggplot(threads, aes(x = Number, y = Total)) + geom_bar(stat = "identity", fill = "darkorchid4") +
  theme(axis.text.x = element_blank()) + labs(title = "Total number of posts in each thread", y = "Total Posts",
      x = "Threads") + theme_bw(base_size = 15)


#get the last 6 rows from the sorted_threads dataset to get the top 6 threads in the data.
top_thread = tail(sorted_threads, n=6)
top_thread



##TaskB part2


## We now look at the top 6 threads and compare some selected linguistic variables to compare with each other. We also analyse the mean of the selected
##linguistic variables in the top 20 threads with the selected one in the top 20 bottom threads and compare the differences.


#create a dataset to store all the mean values of the variables in the webforum by grouping Thread ID.
Bdf1 = aggregate(webforum, by=list(webforum$ThreadID),mean)
#take out the variables/columns that we want to look at which is the top 4 variables in the dataset
Bdf1.1 = Bdf1[ , (names(Bdf1) %in% c("Group.1","Analytic", "Clout", "Authentic", "Tone"))]
#melt the table
melted_bdf1 = melt(Bdf1.1, id="Group.1")


#plot the top 4 linguistic variables for every thread in the dataset with the mean values.
ggplot(melted_bdf1, aes(x = Group.1, y = value, colour=variable, group=variable)) + geom_line() + facet_wrap(~ variable, ncol = 3) +
  labs(title = "Top 4 linguistic variables Analytic, Clout, Tone and Authentic for every thread in the data", y = "Average",
      x = "Threads") + theme_bw(base_size = 15)
```

```
#take out the variables/columns that we want to look at which is the remaining variables in the dataset

Bdf1.2 = Bdf1[ , !(names(Bdf1) %in% c("Time","day","hour","Date","year","month","my","ThreadID","AuthorID","Analytic", "WC", "Clout", "Authentic", "Tone"))]

#melt the table

melted_bdf2 = melt(Bdf1.2, id="Group.1")


#plot the remaining variables for every thread in the dataset with the mean values.

ggplot(melted_bdf2, aes(x = Group.1, y = value, colour=variable, group=variable)) + geom_line() + facet_wrap(~ variable, ncol = 3) +
  labs(title = "Mean values for the remaining variables for every thread in the data", y = "Average",
      x = "Threads") + theme_bw(base_size = 15)
```

```
##Analyse the linguistic variables of the top 6 threads and see if there's a difference in language in different groups


#Find the top author in the forum which is the author that appeared in the forum for the most number of times

#by first getting the total number of users in the forum with their number of posts.

total_authors = as.data.frame(table(webforum$AuthorID))

total_authors = total_authors[order(total_authors$Freq), ]

top_author = as.numeric(as.character(tail(total_authors$Var1, n=1)))


#create a data set and group all the data with the top 5 authors that posted the most number of times in the forum.

#All the data in the data set will be the same author ID which is top 5 authors posted at different times and different forums.

top_author_threads = webforum[webforum$AuthorID %in% top_author,]

top_author_threads
```

```
#Get the top 6 threads in the data set and store it into the data set called top_6_thread_posts.
#From the top 1 author, get the top 6 threads with the most number of posts in the forum.
top_6_thread = as.data.frame(table(top_author_threads$ThreadID))
top_6_thread
top_6_thread = top_6_thread[order(top_6_thread$Freq), ]
top_6_thread = as.numeric(as.character(tail(top_6_thread$Var1, n=6)))
top_6_thread
top_6_thread_posts = webforum[webforum$ThreadID %in% top_6_thread,]


#Create a data set and store every variables with their mean values by grouping with the top 6
threads we got from
#the previous datatset top_6_thread_posts.
top_author_thread_averages = aggregate(top_6_thread_posts, list(top_6_thread_posts$ThreadID),
mean)
top_author_thread_averages
#take the variables/column that we want to analyse or visualise with to plot the graph out.
top_author_thread_averages1 = top_author_thread_averages[ ,
(names(top_author_thread_averages) %in%
                                                  c("Group.1", "Analytic", "Clout", "Tone", "Authentic"))]




top_author_thread_averages1 = melt(top_author_thread_averages1, id="Group.1")
top_author_thread_averages1


labels = c("Thread: 144080", "Thread: 145223", "Thread: 231526", "Thread: 297041", "Thread:
466035", "Thread: 472752",
        "Thread: 144080", "Thread: 145223", "Thread: 231526", "Thread: 297041", "Thread: 466035",
"Thread: 472752",
        "Thread: 144080", "Thread: 145223", "Thread: 231526", "Thread: 297041", "Thread: 466035",
"Thread: 472752",
          "Thread: 144080", "Thread: 145223", "Thread: 231526", "Thread: 297041", "Thread: 466035",
"Thread: 472752")

#plot a bar chart graph for the top 6 threads and analyse the trend. It compares the top 4 linguistic
variables

#against each other to analyse if there is a difference in the language.

ggplot(data = top_author_thread_averages1, aes(x = seq(1:length(value)), y = value, fill = variable)) +

  geom_bar(stat = 'identity', position = 'dodge') + labs(title = "Linguistic variables for the top 6
threads",

                                                      y = "Average", x = "Threads") + theme_bw(base_size = 15) +

  theme(axis.text.x = element_blank(), axis.ticks.x=element_blank()) +

  geom_text(aes(label=labels), angle =90, hjust=1) +

  scale_fill_discrete(name = "Legend") +

  xlab("Thread") +

  ylab("Average")





#take the variables/column that we want to analyse or visualise with to plot the graph out.

top_author_thread_averages2 = top_author_thread_averages[ ,
(names(top_author_thread_averages) %in% c("Group.1", "ppron",

                                                                              "i", "we", "you", "shehe","they",

                                                                              "affect", "posemo", "negemo"))]


top_author_thread_averages2 = melt(top_author_thread_averages2, id="Group.1")

top_author_thread_averages2


labels = c("Thread: 144080", "Thread: 145223", "Thread: 231526", "Thread: 297041", "Thread:
466035", "Thread: 472752",

          "Thread: 144080", "Thread: 145223", "Thread: 231526", "Thread: 297041", "Thread: 466035",
"Thread: 472752",
     "Thread: 144080", "Thread: 145223", "Thread: 231526", "Thread: 297041", "Thread: 466035", "Thread: 472752",
     "Thread: 144080", "Thread: 145223", "Thread: 231526", "Thread: 297041", "Thread: 466035", "Thread: 472752",
     "Thread: 144080", "Thread: 145223", "Thread: 231526", "Thread: 297041", "Thread: 466035", "Thread: 472752",
     "Thread: 144080", "Thread: 145223", "Thread: 231526", "Thread: 297041", "Thread: 466035", "Thread: 472752",
     "Thread: 144080", "Thread: 145223", "Thread: 231526", "Thread: 297041", "Thread: 466035", "Thread: 472752",
     "Thread: 144080", "Thread: 145223", "Thread: 231526", "Thread: 297041", "Thread: 466035", "Thread: 472752",
     "Thread: 144080", "Thread: 145223", "Thread: 231526", "Thread: 297041", "Thread: 466035", "Thread: 472752")
```

```
#plot a bar chart graph for the top 6 threads and analyse the trend. It compares the top 4 linguistic variables
#against each other to analyse if there is a difference in the language.
ggplot(data = top_author_thread_averages2, aes(x = seq(1:length(value)), y = value, fill = variable)) +
  geom_bar(stat = 'identity', position = 'dodge') + labs(title = "Linguistic variables for the top 6 threads",
                                               y = "Average", x = "Threads") + theme_bw(base_size = 15) +
  theme(axis.text.x = element_blank(), axis.ticks.x=element_blank()) +
  geom_text(aes(label=labels), angle =90, hjust=1) +
  scale_fill_discrete(name = "Legend") +
  xlab("Thread") +
  ylab("Average")
```

```
#create a dataset that sort the top 6 threads with the posts in the forum in ascending order.
data = top_6_thread_posts[order(top_6_thread_posts$ThreadID),]
#Filter out the thread with the thread id of 231526 and store it in thread_231526
thread_231526 = filter(data, data$ThreadID == "231526")
```

```
#Filter out the thread with the thread id of 466035 and store it in 466035

thread_446035 = filter(data, data$ThreadID == "466035")



##Use hypothesis testing to test if the thread 297041 and the thread 145223 are the same or different

t1 = t.test(thread_231526$Analytic,thread_446035$Analytic, conf.level = 0.99)

t1$p.value


t2 = t.test(thread_231526$Clout,thread_446035$Clout, conf.level = 0.99)

t2$p.value


t3 = t.test(thread_231526$Authentic,thread_446035$Authentic, conf.level = 0.99)

t3$p.value


t4 = t.test(thread_231526$Tone,thread_446035$Tone, conf.level = 0.99)

t4$p.value


t5 = t.test(thread_231526$ppron,thread_446035$ppron, conf.level = 0.99)

t5$p.value


t6 = t.test(thread_231526$i,thread_446035$i, conf.level = 0.99)

t6$p.value


t7 = t.test(thread_231526$we,thread_446035$we, conf.level = 0.99)

t7$p.value


t8 = t.test(thread_231526$you,thread_446035$you, conf.level = 0.99)

t8$p.value


t9 = t.test(thread_231526$shehe,thread_446035$shehe, conf.level = 0.99)
```

```
t9$p.value

t10 = t.test(thread_231526$they,thread_446035$they, conf.level = 0.99)
t10$p.value

t11 = t.test(thread_231526$affect,thread_446035$affect, conf.level = 0.99)
t11$p.value

t12 = t.test(thread_231526$posemo,thread_446035$posemo, conf.level = 0.99)
t12$p.value

t13 = t.test(thread_231526$negemo,thread_446035$negemo, conf.level = 0.99)
t13$p.value


##Analyse the top 20 threads and bottom 20 threads in the dataset and compares the mean of all the linguistic variables with each other to see if there's a
##difference in these different groups.

#Find the top threads in the forum which is the thread that appeared in the forum for the most number of times
total_threads = as.data.frame(table(webforum$ThreadID))
total_threads = total_threads[order(total_threads$Freq),]


#create a dataset and group all the data with the top 20 threads that have the most number of posts in the forum.
#All the data in the dataset will be the same thread ID which is top 20 threads posted at different times and different forums.
top_threads = as.numeric(as.character(tail(total_threads$Var1, n=20)))
top_threads = webforum[webforum$ThreadID %in% top_threads,]
```

```
#create a dataset and group all the data with the 20 thread that posted the least number of times in
the forum.

#All the data in the dataset will be the same thread ID which is author posted the least amount of
times

#at different times and different forums.

bott_threads = as.numeric(as.character(head(total_threads$Var1, n=20)))

bott_threads = webforum[webforum$ThreadID %in% bott_threads,]
```

```
top_threads_averages3 = aggregate(top_threads, list(top_threads$ThreadID), mean)

top_threads_averages3 = top_threads_averages3[, (names(top_threads_averages3) %in%
c("Group.1", "Analytic", "Tone", "Clout",

                                    "Authentic", "ppron","anger","anx","social",

                                    "affect", "posemo", "negemo" ))]
```

```
#Find the mean of the selected linguistic variables for the top 20 threads

top_mean_thread = summarise(top_threads_averages3, Analytic_mean = mean(Analytic),
Clout_mean = mean(Clout),

                             Authentic_mean = mean(Authentic), Tone_value = mean(Tone),
ppron_mean = mean(ppron),

                             affect_mean = mean(affect), posemo_mean = mean(posemo),
negemo_value = mean(negemo),

                             anx_mean = mean(anx), anger_mean = mean(anger),social_mean =
mean(social),

                             )
```

```
melted_top_mean_thread_averages = melt(top_mean_thread)

melted_top_mean_thread_averages
```

```
labels=c("Analytic", "Tone", "Clout", "Authentic", "ppron","anger","anx","social","affect", "posemo",
"negemo")
```

```
#Plot the bar chart for the mean of selected linguistic variables for the top 20 threads

ggplot(data = melted_top_mean_thread_averages, aes(x = seq(1:length(value)), y = value, fill = variable)) +

  geom_bar(stat = 'identity', position = 'dodge') +

  theme(axis.text.x = element_blank(), axis.ticks.x=element_blank()) +

  geom_text(aes(label=labels), angle =90, hjust=2) +

  scale_fill_discrete(name = "Legend") +

  xlab("Thread") +

  ylab("Thread Average")
```

```
bott_threads_averages3 = aggregate(bott_threads, list(bott_threads$ThreadID), mean)

bott_threads_averages3 = bott_threads_averages3[, (names(bott_threads_averages3) %in% c("Group.1", "Analytic", "Tone", "Clout",

                                                  "Authentic", "ppron","anger","anx","social",

                                                  "affect", "posemo", "negemo" ))]
```

```
#Find the mean of the selected linguistic variables for the bottom 20 threads

bott_mean_thread = summarise(bott_threads_averages3, Analytic_mean = mean(Analytic), Clout_mean = mean(Clout),

                         Authentic_mean = mean(Authentic), Tone_value = mean(Tone), ppron_mean = mean(ppron),

                         affect_mean = mean(affect), posemo_mean = mean(posemo), negemo_value = mean(negemo),

                         anx_mean = mean(anx), anger_mean = mean(anger),social_mean = mean(social),

)


melted_bott_mean_thread_averages = melt(bott_mean_thread)
```

```
labels=c("Analytic", "Tone", "Clout", "Authentic", "ppron","anger","anx","social","affect", "posemo",
"negemo")

#Plot the bar chart for the mean of selected linguistic variables for the bottom 20 threads

ggplot(data = melted_bott_mean_thread_averages, aes(x = seq(1:length(value)), y = value, fill =
variable)) +
  geom_bar(stat = 'identity', position = 'dodge') +
  theme(axis.text.x = element_blank(), axis.ticks.x=element_blank()) +
  geom_text(aes(label=labels), angle =90, hjust=2) +
  scale_fill_discrete(name = "Legend") +
  xlab("Thread") +
  ylab("Thread Average")





##Use hypothesis testing to test if the variables are similar

t1 = t.test(top_threads$Analytic,bott_threads$Analytic, conf.level = 0.99)

t1$p.value


t2 = t.test(top_threads$Clout,bott_threads$Clout, conf.level = 0.99)

t2$p.value


t3 = t.test(top_threads$Authentic,bott_threads$Authentic, conf.level = 0.99)

t3$p.value
```

```
t4 = t.test(top_threads$Tone,bott_threads$Tone, conf.level = 0.99)
t4$p.value

t5 = t.test(top_threads$ppron,bott_threads$ppron, conf.level = 0.99)
t5$p.value

t6 = t.test(top_threads$affect,bott_threads$affect, conf.level = 0.99)
t6$p.value

t7 = t.test(top_threads$posemo,bott_threads$posemo, conf.level = 0.99)
t7$p.value

t8 = t.test(top_threads$negemo,bott_threads$negemo, conf.level = 0.99)
t8$p.value

t9 = t.test(top_threads$anx,bott_threads$anx, conf.level = 0.99)
t9$p.value

t10 = t.test(top_threads$anger,bott_threads$anger, conf.level = 0.99)
t10$p.value

t11 = t.test(top_threads$social,bott_threads$social, conf.level = 0.99)
t11$p.value




####TaskB part3
library(igraph)
```

```
#We now take the top thread and analayse all the linguistic variables and see if it changes over time.


#Get the top thread in the total threads. The top thread will have the most number of posts compared to other threads.

top1_thread = as.numeric(as.character(tail(total_threads$Var1, n=1)))

top1_thread


#Group the thread into different dataset by months,days and hours

top1.1_thread = webforum[webforum$ThreadID %in% top1_thread,]

top1.1_my_thread = aggregate(top1.1_thread,by=list(top1.1_thread$my), mean)

top1.1_day_thread = aggregate(top1.1_thread,by=list(top1.1_thread$day), mean)

top1.1_day_thread$Group.1 = factor(top1.1_day_thread$Group.1, levels = c("Monday","Tuesday","Wednesday","Thursday","Friday","Saturday","Sunday"))


top1.1_hour_thread = aggregate(top1.1_thread,by=list(top1.1_thread$hour), mean)




#Plot the graph for Month,Year for the top 5 variables. The trend for the top 5 variables in 1 month for a year.

get_my_graph1(top1.1_my_thread)

#Plot the graph for Month,Year for the selected variables. The trend for the selected linguistic variables in 1 month for a year.

get_my_graph2(top1.1_my_thread)



#Plot the graph for Day,Week for the top 5 variables. The trend for the top 5 variables in 1 day for a week.

get_day_graph1(top1.1_day_thread)

#Plot the graph for Day,Week for the selected variables. The trend for the selected linguistic variables in 1 day for a week.

get_day_graph2(top1.1_day_thread)
```

```
#Plot the graph for Hour,Day for the top 5 variables. The trend for the top 5 variables in 24 hours for a day.

get_hour_graph1(top1.1_hour_thread)

#Plot the graph for Hour,Day for the selected variables. The trend for the selected linguistic variables in 24 hours for a day.

get_hour_graph2(top1.1_hour_thread)




####Task C

## Now we look at the social network of the participants in the forum. We analyse the network by looking at the top author and group the thread id together.

## We only look at two months for the same threads.


#Find the most popolar author in the forum which is the author that appeared in the forum for the most number of times

popular_author = as.data.frame(table(webforum$AuthorID))

#Sort the dataset from lowest number of authors to the highest

popular_author = popular_author[order(popular_author$Freq), ]




#Get the highest number of author that appear in the forum by getting the last row data as the dataset is already sorted.

most_popular_author = as.numeric(as.character(tail(popular_author$Var1, n=1)))

#Group the data with just the most popolar author which is the author that posted most number of times in the forum.

top_author_threads = webforum[webforum$AuthorID %in% most_popular_author,]


#Group the Thread ID of the forum by grouping with the data that has all the most popular author.

top_thread = as.data.frame(table(top_author_threads$ThreadID))
```

```r
#Sort the dataset
top_thread = top_thread[order(top_thread$Freq), ]
top_thread
#Get the highest number of author that appear in the forum by getting the last row data as the dataset is already sorted.
top_thread = as.numeric(as.character(tail(top_thread$Var1, n=1)))

#Group the data with the top thread in the forum which means the thread have the most number of posts
top_thread_posts = webforum[webforum$ThreadID %in% top_thread,]
#Get the item that is needed by using subset function
top_thread_posts = subset(top_thread_posts, select = c(ThreadID,AuthorID,my))

#Group the author ID together by grouping the Author ID and month,year together
author_thread_average = top_thread_posts %>% group_by(AuthorID,my) %>% summarise(Total = length(AuthorID))
#Sort the dataset
author_thread_average = author_thread_average[order(author_thread_average$Total), ]
#Get the month and year that have the highest number of thread in the forum. The month is 2004-08.
author_thread_average = (as.character((tail(author_thread_average$my, n=1))))

#Create a dataset to store the top thread ID in the forum
network = webforum[webforum$my %in% author_thread_average,]
#Get the item that is needed by using subset function
network = subset(network, select = c(ThreadID,AuthorID,my))
#Group the network dataset by Thread ID and Author ID
network = network %>% group_by(ThreadID,AuthorID) %>% summarise(length(AuthorID))
#Sort the dataset
network = network[order(network$ThreadID), ]
```

```
#For first thread

first_thread = filter(network, ThreadID == '145223')

#Convert the Author ID to string type

string1 = toString(first_thread$AuthorID)

#split every AuthorID with a ,

final_first_thread = unlist(strsplit(string1, ", "))

#Adjacency matrix for the graph

adj_matrix1 = matrix(1, nrow = (length(final_first_thread)),  ncol = (length(final_first_thread)),
dimnames = list(final_first_thread,final_first_thread))

adj_matrix1


#Plot the social network for this matrix

g1 = graph.adjacency(adj_matrix1, mode = 'undirected', weighted = NULL, diag =FALSE)

plot(g1)



#For second thread

second_thread = filter(network, ThreadID == '104567')

#Convert the Author ID to string type

string2 = toString(second_thread$AuthorID)

#split every AuthorID with a ,

final_second_thread = unlist(strsplit(string2, ", "))

#Adjacency matrix for the graph

adj_matrix2 = matrix(1, nrow = (length(final_second_thread)),  ncol = (length(final_second_thread)),
dimnames = list(final_second_thread,final_second_thread))

adj_matrix2

#Plot the social network for this matrix

g2 = graph.adjacency(adj_matrix2, mode = 'undirected', weighted = NULL, diag =FALSE)

plot(g2)


#merge the first graph and the second graph together by using union function %u%

merge_graph = (g1 %u% g2)
```

```
plot(merge_graph)




#For third thread

third_thread = filter(network, ThreadID == '150445')

#Convert the Author ID to string type

string3 = toString(third_thread$AuthorID)

#split every AuthorID with a ,

final_third_thread = unlist(strsplit(string3, ", "))

#Adjacency matrix for the graph

adj_matrix3 = matrix(1, nrow = (length(final_third_thread)),  ncol = (length(final_third_thread)), dimnames = list(final_third_thread,final_third_thread))

adj_matrix3

#Plot the social network for this matrix

g3 = graph.adjacency(adj_matrix3, mode = 'undirected', weighted = NULL, diag =FALSE)

plot(g3)


#Merge the graph 1, graph 2 and graph 3 together using the union function %u%

merge_graph2 = (g1 %u% g2 %u% g3)

plot(merge_graph2)




#plot a histogram to visualise the degree distribution of the merge graph

hist(degree(merge_graph2),      breaks  =          5,         col        =          "grey")


#Show a vertex summary of the merge graph such as calculating all the degree, betweenness, closeness centrality

#Eigenvector centrality and the average path of the merge graph.

degree = as.table(degree(merge_graph2))

betweenness = as.table(betweenness(merge_graph2))

closeness = as.table(closeness(merge_graph2),   digits = 2)
```

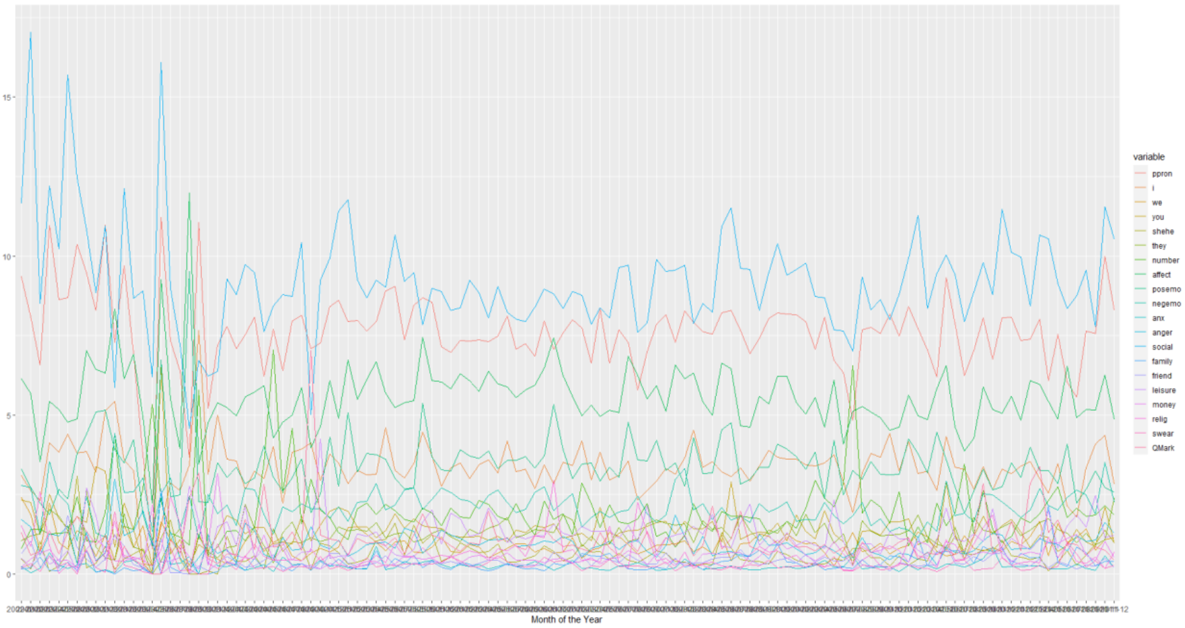
```
eig        =          as.table(evcent(merge_graph2)$vector)
averagepath = average.path.length(merge_graph2)
#find the dimameter of the merge graph
diameter(merge_graph2)

#create atable to find out the most important person in the network
table = as.data.frame(rbind(degree, betweenness, closeness, eig))
table = t(table)
table



##Repeat the process but do it for the next month. 2004-09
#Create a dataset to store the top thread ID in the forum
network2 = webforum[webforum$my %in% '2004-09',]
network2
#Get the item that is needed by using subset function
network2 = subset(network2, select = c(ThreadID,AuthorID,my))
#Group the network dataset by Thread ID and Author ID
network2 = network2 %>% group_by(ThreadID,AuthorID) %>% summarise(length(AuthorID))
#Sort the dataset
network2 = network2[order(network2$ThreadID), ]
#For first thread


first_thread = filter(network2, ThreadID == '145223')
#Convert the Author ID to string type
string1 = toString(first_thread$AuthorID)
#split every AuthorID with a ,
final_first_thread = unlist(strsplit(string1, ", "))
#Adjacency matrix for the graph
```

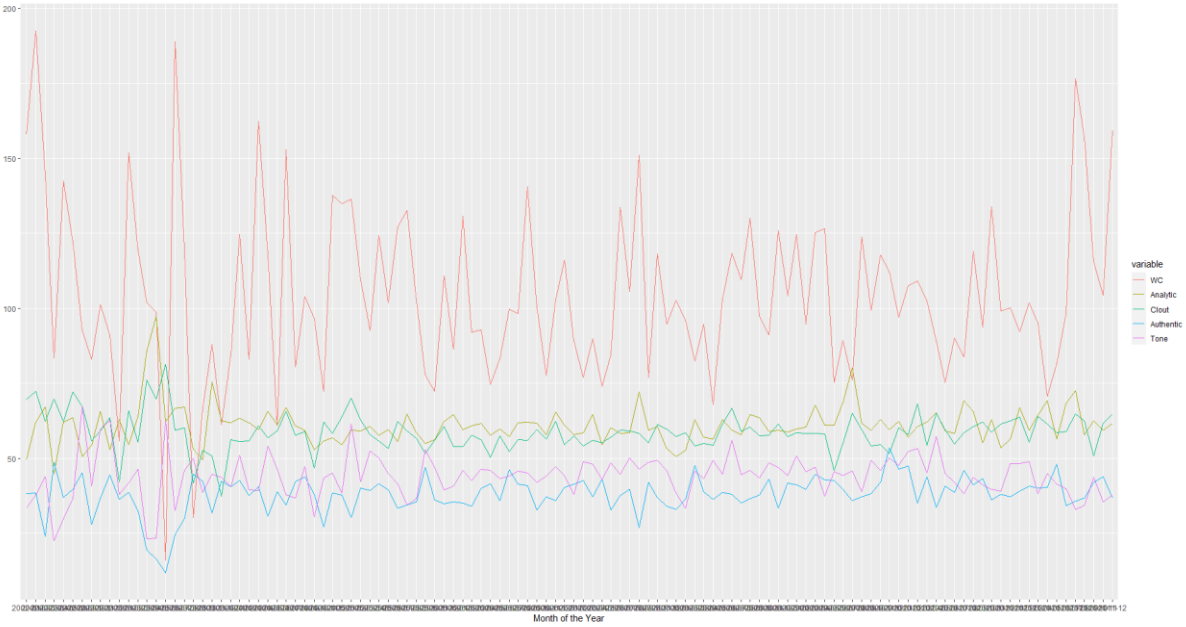
```
adj_matrix1 = matrix(1, nrow = (length(final_first_thread)),  ncol = (length(final_first_thread)),
dimnames = list(final_first_thread,final_first_thread))
adj_matrix1
#Plot the social network for this matrix
g1 = graph.adjacency(adj_matrix1, mode = 'undirected', weighted = NULL, diag =FALSE)
plot(g1)



#For second thread
second_thread = filter(network2, ThreadID == '104567')
#Convert the Author ID to string type
string2 = toString(second_thread$AuthorID)
#split every AuthorID with a ,
final_second_thread = unlist(strsplit(string2, ", "))
#Adjacency matrix for the graph
adj_matrix2 = matrix(1, nrow = (length(final_second_thread)),  ncol = (length(final_second_thread)),
dimnames = list(final_second_thread,final_second_thread))
adj_matrix2
#Plot the social network for this matrix
g2 = graph.adjacency(adj_matrix2, mode = 'undirected', weighted = NULL, diag =FALSE)
plot(g2)



#For third thread
third_thread = filter(network2, ThreadID == '150445')
#Convert the Author ID to string type
string3 = toString(third_thread$AuthorID)
#split every AuthorID with a ,
final_third_thread = unlist(strsplit(string3, ", "))
#Adjacency matrix for the graph
adj_matrix3 = matrix(1, nrow = (length(final_third_thread)),  ncol = (length(final_third_thread)),
dimnames = list(final_third_thread,final_third_thread))
```

```
adj_matrix3

#Plot the social network for this matrix

g3 = graph.adjacency(adj_matrix3, mode = 'undirected', weighted = NULL, diag =FALSE)

plot(g3)

#Merge the graph 1, graph 2 and graph 3 together using the union function %u%

merge_graph2 = (g1 %u% g2 %u% g3)

plot(merge_graph2)
```

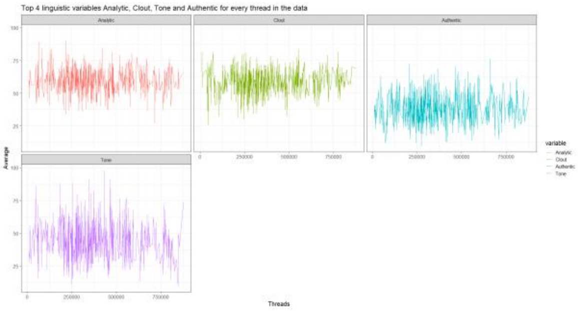
Graph Appendix

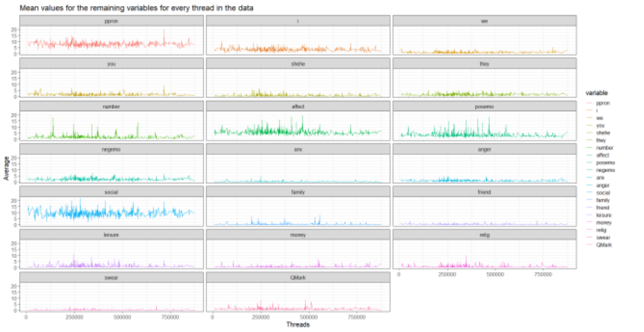
The graph below showed all the variables in the dataset for every month of the year from 2002 to 2011.





Look at all the linguistic variables and find the mean of it for every thread in the dataset:





From the graph above, we can see that the language used over different groups or threads are really different because we can see that for some linguistic variables it has a really high value in Analytic, Clout and Tone but it will be low in Authentic whereas some other threads have higher tone value and lower in Analytic, Clout and Authentic value. This proves that it is possible to see a change in different language in different groups/threads.