Sentiment analysis using R

Author: Kuah Jia Chen

Table of Contents

Question a.1

How active are participants over the longer term (that is, over months and/or years)? Are there periods where activity increases or decreases?

To analyze the participants' activity over the long term (i.e., months and years), a preprocessing step is required to generate a suitable graph. Firstly, I will group the data by months and years, combining columns with the same month and year. Then, I will calculate the frequency of occurrences for each group. This process will result in a dataset that contains the number of posts for each month and year, allowing us to analyze the participants' activeness over time.

```
> head(date_table)
# A tibble: 6 x 2
  Date       No_of_post
  <date>          <int>
1 2002-01-01        118
2 2002-02-01         55
3 2002-03-01         66
4 2002-04-01         86
5 2002-05-01        125
6 2002-06-01         34
> mean(date_table$No_of_post)
[1] 166.6667
```

Figure 1. Sample of the Dataset and Average Post Count

Note: Please disregard the day (represented as "xxxx-xx-01") in each row, as it holds no significance. The day will not be taken into consideration when plotting the graph.

This figure showcases the first six rows of the dataset, which have been grouped based on months and years for the purpose of generating graphs. The second column represents the number of posts recorded in each month. Additionally, the figure displays the average post count across the entire dataset. The dataset is now prepared for graph generation, allowing for further analysis and visualization of the participants' activity over time.

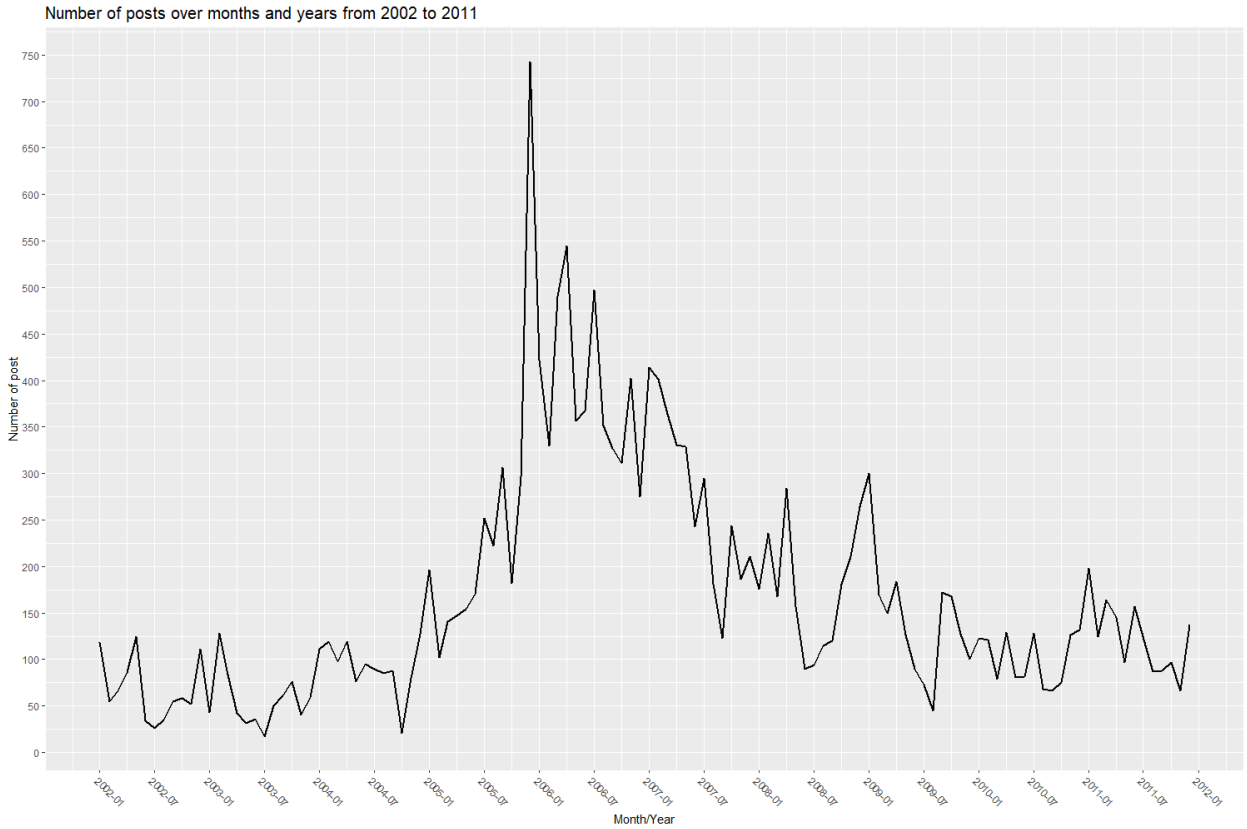

Figure 2. Time Series of Post Count from 2002 to 2011

This time series graph depicts the number of posts recorded over a span of ten years, from 2002 to 2011. The graph provides insights into the fluctuation and trend of post activity over different months and years. By analyzing this visualization, we can gain a better understanding of the participants' engagement and the overall dynamics of the forum during this time period.

From the analysis of Figure 1, we can observe that the average number of posts per month/year is approximately 167, indicating a relatively active participation level among the users.

Figure 2 provides further insights into the participants' activity patterns over the entire time period. Initially, from January 2002 to July 2004, the participant activity remained relatively stable with some minor variations. However, the time series graph highlights a significant surge in the number of posts starting from July 2004 to January 2006. During this period, the number of posts continuously increased, reaching a peak of around 750 posts in January 2006. This suggests a notable rise in participant engagement.

Subsequently, as depicted in Figure 2, there was a noticeable decline in post activity after January 2006. Particularly between January 2006 and July 2009, although experiencing occasional fluctuations, the overall trend demonstrates a consistent decrease in the number of posts. However, starting from July 2009, the post count tends to stabilize, indicating a relatively steady level of participant activity during this period.

In order to support the observation of a continuous increase in the number of posts from July 2004 to January 2006 compared to the period prior to July 2004, a hypothesis test will be conducted. The null hypothesis states that the mean number of posts in the period from July 2004 to January 2006 is greater than the mean number of posts before July 2004. The default confidence level for this test is set at 95%.

By performing the hypothesis test, we can assess the statistical significance of the observed increase in post activity during the specified time period and determine if it is unlikely to have occurred by chance.

```
> t.test(date_table_between_July_2004_Jan_2006$No_of_post,date_table_before_July_2004$No_of_post,alternative = "less")

        Welch Two Sample t-test

data:  date_table_between_July_2004_Jan_2006$No_of_post and date_table_before_July_2004$No_of_post
t = 3.1464, df = 17.923, p-value = 0.9972
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     -Inf 184.4275
sample estimates:
mean of x mean of y
189.05556  70.16667
```

Figure 3. Results of Hypothesis Testing using t-test in R

This figure presents the outcomes of the hypothesis testing conducted using the t-test in R. The aim of the test was to assess the statistical significance of the observed increase in post activity from July 2004 to January 2006, as compared to the period preceding July 2004. The default confidence level of 95% was employed for this analysis. The results of the t-test provide valuable insights into the significance of the observed differences in post counts, aiding in the evaluation of the hypothesis.

According to the results presented in Figure 3, the p-value obtained from the hypothesis testing is 0.9972. This indicates that if the null hypothesis were true, there would be approximately a 99.7% chance of observing a sample with a difference from the null as extreme or more extreme than the one we observed. Since the p-value is greater than the conventional significance level of 0.05, there is insufficient/weak evidence to reject the null hypothesis. Therefore, we cannot conclude that the mean number of posts in the period from July 2004 to January 2006 is greater than the mean number of posts before July 2004. This finding supports my earlier statement that the post activity increased during the period from July 2004 to January 2006.

Similarly, to investigate the decrease in activity during the period from January 2006 to July 2009, a hypothesis test will be performed. However, in this case, the mean number of posts in the period from January 2006 to July 2009 will be compared with the mean number of posts in the period from October 2005 to January 2006. This comparison is more appropriate as the latter period represents the approximate peak of the graph, allowing for a fair assessment of the subsequent decrease. The null hypothesis states that the mean number of posts in the period from October 2005 to January 2006 is greater than the mean number of posts in the period from January 2006 to July 2009. The default confidence level of 95% will be utilized for this hypothesis test.

```
> t.test(date_table_between_Oct_2005_Jan_2006$No_of_post,date_table_between_Jan_2006_July_2009$No_of_post,alternative = "less")

        Welch Two Sample t-test

data:  date_table_between_Oct_2005_Jan_2006$No_of_post and date_table_between_Jan_2006_July_2009$No_of_post
t = 0.8298, df = 2.0466, p-value = 0.7539
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
    -Inf 637.524
sample estimates:
mean of x mean of y
 407.3333  264.5476
```

Figure 4. Results of Hypothesis Testing using t-test in R

According to the results depicted in Figure 4, the obtained p-value from the hypothesis testing is 0.7539. This indicates that if the null hypothesis were true, there would be approximately a 75.4% chance of observing a sample with a difference from the null as extreme or more extreme than the one we observed. Since the p-value exceeds the conventional significance level of 0.05, there is insufficient/weak evidence to reject the null hypothesis. Consequently, we cannot conclude that the mean number of posts in the period from October 2005 to January 2006 is greater than the mean number of posts in the period from January 2006 to July 2009. This finding supports my earlier statement that the post activity decreases during the period from January 2006 to July 2009.

Is there a trend over time?

By examining Figure 2, a clear trend in the number of posts over time becomes apparent. The time series graph exhibits a distinctive parabolic shape. From January 2002 to July 2004, the number of posts fluctuated without any clear pattern. However, starting from July 2004, there was a substantial and continuous increase in the number of posts until January 2006. Following this peak, a decline in the number of posts occurred from January 2006 to July 2009. Subsequently, after July 2009, the number of posts stabilized, displaying relatively consistent levels with minor fluctuations until the end of 2011.

Question a.2

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

To visualize the changes in all levels of linguistic variables throughout the duration of the forum, I will employ line graphs as a visual tool. Instead of overcrowding a single graph with all 18 lines, I have opted to create three separate graphs for better clarity. The first graph will focus on the four summary variables, the second graph will display variables related to pronouns, and the third graph will encompass the remaining linguistic variables. It is worth noting that word count was excluded from the analysis as it does not provide functional insights into the relationship between linguistic variables.

To generate the line graphs, data preprocessing is required. Initially, I will group the linguistic variables for each post based on the month and year. Subsequently, I will calculate the mean value for each linguistic variable within each month and year. Using the mean of the percentile or proportion of posts within a specific month and year for each linguistic variable enables better visualization of how the variable's level changes over time, accentuating fluctuations in the graphs.

```
> head(webforum_four)
        Date Average_of_Analytic Average_of_Clout Average_of_Authentic Average_of_Tone
1 2002-01-01              57.89390         68.62008              32.45754         36.70695
2 2002-02-01              61.02491         69.83873              33.41600         42.19982
3 2002-03-01              61.59879         67.16364              32.04273         37.12485
4 2002-04-01              58.47233         59.73709              39.54360         40.99802
5 2002-05-01              59.69344         60.69296              33.36536         46.97160
6 2002-06-01              65.14029         65.73765              23.58382         51.67588
```

Figure 6. First 6 Rows of the Pre-processed Data Frame

This figure displays the resulting data frame obtained after performing the preprocessing operations, allowing for further analysis and visualization. Each linguistic variable will undergo the same group by operation to create separate line graphs for each variable, as shown in Figure 7.

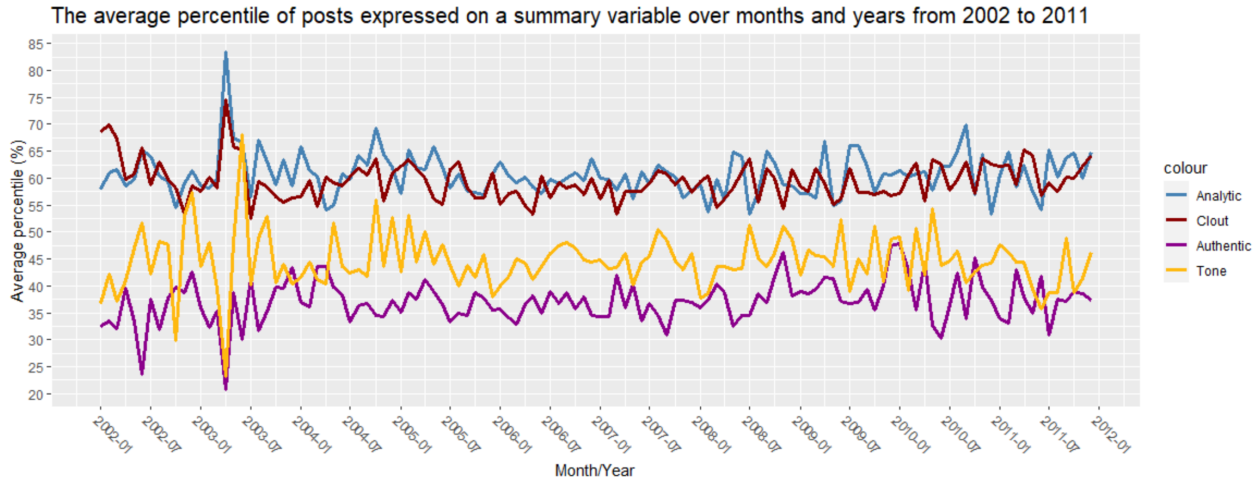

Figure 7. Average Percentile of Posts for Summary Variables over Time

In Figure 7, it is observed that the mean percentiles for Analytic and Clout show a consistent agreement over time, as their corresponding lines closely follow each other. Furthermore, the average percentiles for both Analytic and Clout tend to fall within the range of 55% to 70% for most of the duration. On the other hand, the mean percentile for Authentic consistently remains lower than Analytic and Clout over time, with values frequently ranging between 35% and 50%. However, a notable exception is observed around July 2003, where the mean percentile for Authentic experiences a significant rise, reaching approximately 70%. As for Tone, it consistently exhibits the lowest mean percentile among the summary variables throughout the entire period, typically ranging from 30% to 45%. In summary, the levels of the four summary variables do show variation over time, although they mostly remain within specific ranges.

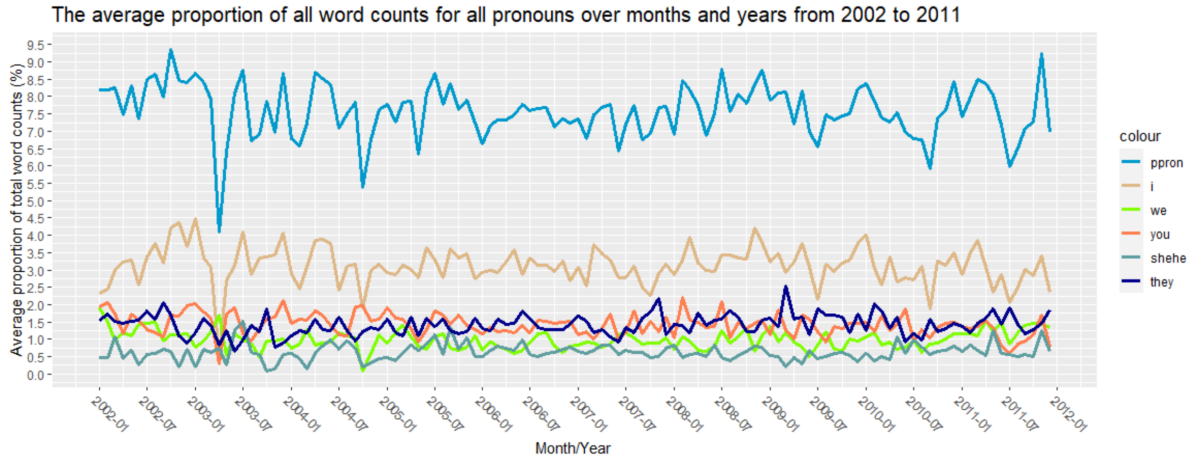

Figure 8. Average Proportion of Word Counts for Pronouns over Months and Years (2002-2011)

The figure depicts the average proportion of total word counts for all pronouns in the posts over the duration of the forum, spanning from 2002 to 2011. The graph reveals the fluctuation and relative distribution of different pronouns throughout this period.

Observing the graph, we can discern that the average proportion for the pronoun "ppron" consistently ranges between 6.0% to 9.0%, exhibiting the highest value among all other pronouns. This aligns with expectations as "ppron" encompasses the sum of all other pronouns displayed. The second-highest proportion belongs to the pronoun "i," which predominantly falls between 2.0% to 4.0% throughout the entire period, with a notable drop to approximately 1.0% in April 2003.

Of particular interest is the close alignment of the average proportion lines for the pronouns "we," "you," "shehe," and "they." These pronouns consistently maintain a percentage in the range of 0% to 2.0% over the long term, indicating their relatively balanced usage.

In conclusion, the levels of pronoun variables exhibit fluctuations over the duration of the forum. However, their values tend to remain within specific ranges, with "ppron" displaying the highest average proportion consistently.

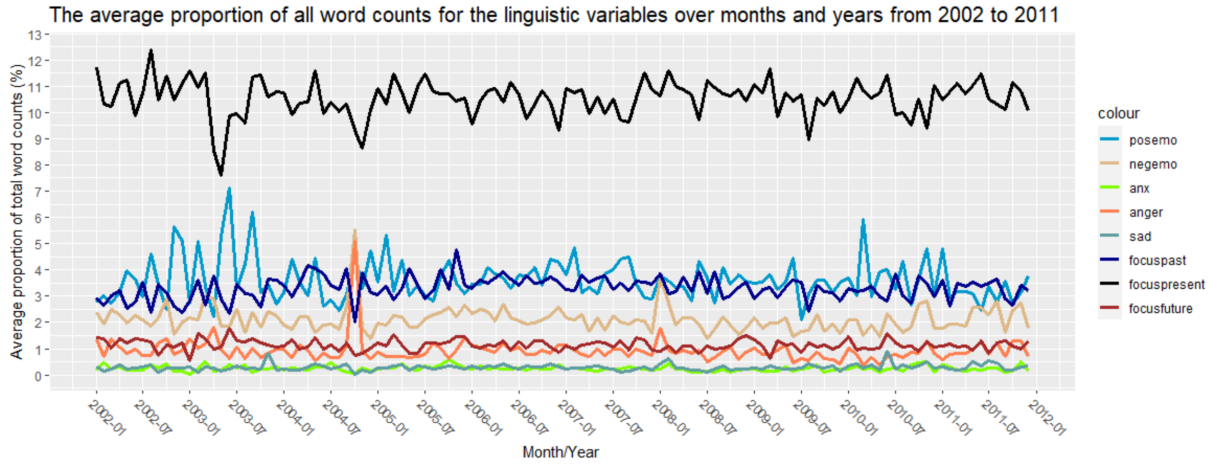

Figure 9. Average Proportion of Linguistic Variables in Posts over Months and Years (2002-2011)

The figure depicts the average proportion of total word counts for each linguistic variable in the posts over the span of months and years from 2002 to 2011.

Examining the graph, it becomes evident that the posts predominantly focus on the present, as indicated by the consistently higher average proportion for the "focuspresent" variable, ranging from 9% to 12%. This proportion surpasses that of the other variables displayed in the graph. Following closely are "posemo" and "focuspast," with their average proportions remaining within the range of 2.5% to 5% throughout the entire period.

Remarkably, the lines representing "negemo," "anger," "focusfuture," "anx," and "sad" appear to be closely aligned, signifying that these variables exhibit comparable average proportions, which generally range from 0% to 2.5% over the long term.

In summary, the observed linguistic variables in Figure 9 showcase variations in their average proportions over the duration of the forum. Despite the fluctuating trends depicted by the line graphs, the values consistently fall within specific ranges, indicating relative stability in their usage throughout the examined period.

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

To examine the long-term relationship between linguistic variables, a heatmap was generated to visualize the correlations between each pair of variables (excluding word count, as previously explained). To prepare the data for the heatmap, a series of pre-processing steps were performed.

First, the columns from the three data frames used to create Figure 7, Figure 8, and Figure 9 were combined into a single data frame using the cbind function. This consolidation allowed for a comprehensive analysis of the linguistic variables.

Next, the data frame was reshaped using the melt function. This transformation ensured that the data was in a suitable format for generating the heatmap.

Finally, the reshaped data frame was utilized to generate the heatmap, which provides a visual representation of the correlations between different linguistic variables.

By following this approach, we can effectively explore the relationships and potential patterns among the linguistic variables over the extended duration of the forum.

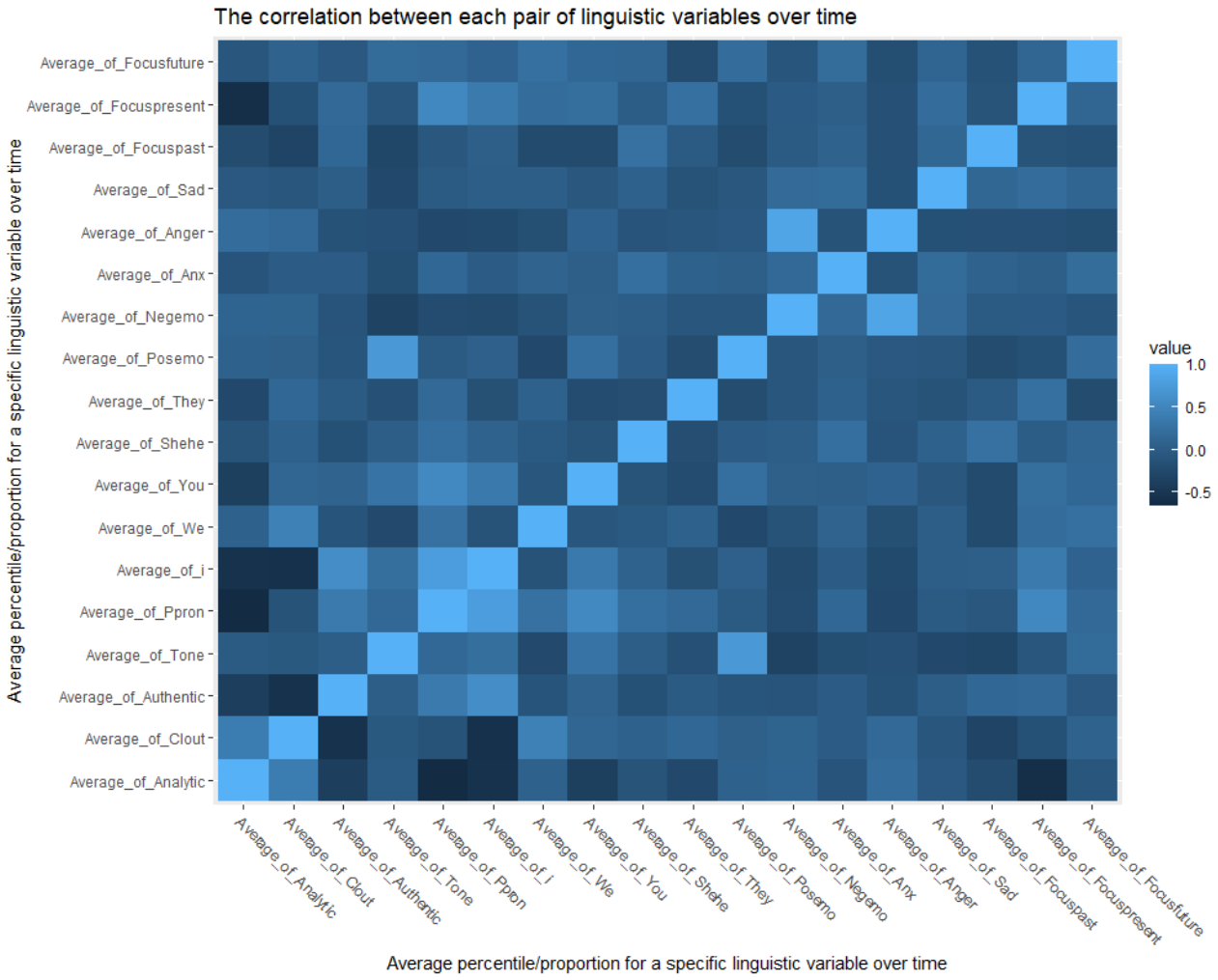

Figure 10. shows the correlation between each pair of linguistic variables over time

Figure 10 displays the heatmap generated from the pre-processed data frame, "correlation_all." This heatmap provides insights into the correlations between pairs of linguistic variables. While most pairs exhibit no significant correlation, there are a few exceptions worth noting.

For example, the correlation between Average_of_Anger and Average_of_Negemo is approximately 0.86, indicating a relatively strong positive correlation. Similarly, the correlation between Average_of_Ppron and Average_of_i is around 0.78, suggesting a moderate positive correlation. On the other hand, the correlation between Average_of_Analytic and Average_of_Ppron is approximately -0.65, indicating a moderate negative correlation.

Please note that these correlation values were obtained by examining the "correlation_all" data frame. Although there may be other pairs of variables with high positive or negative correlations, limitations regarding space prevent me from plotting scatter plots and conducting regression models for each pair.

To further explore the relationship between Average_of_Negemo and Average_of_Anger, I have created a scatter plot in Figure 11. In this plot, Average_of_Negemo is plotted on the x-axis, Average_of_Anger is plotted on the y-axis, and Average_of_Tone is included to

provide additional context. Additionally, a regression line is included (depicted in red) to demonstrate the positive correlation between the two linguistic variables.

By presenting this scatter plot and regression model, we can observe and analyze the relationship between Average_of_Negemo and Average_of_Anger, shedding light on potential associations between these variables.

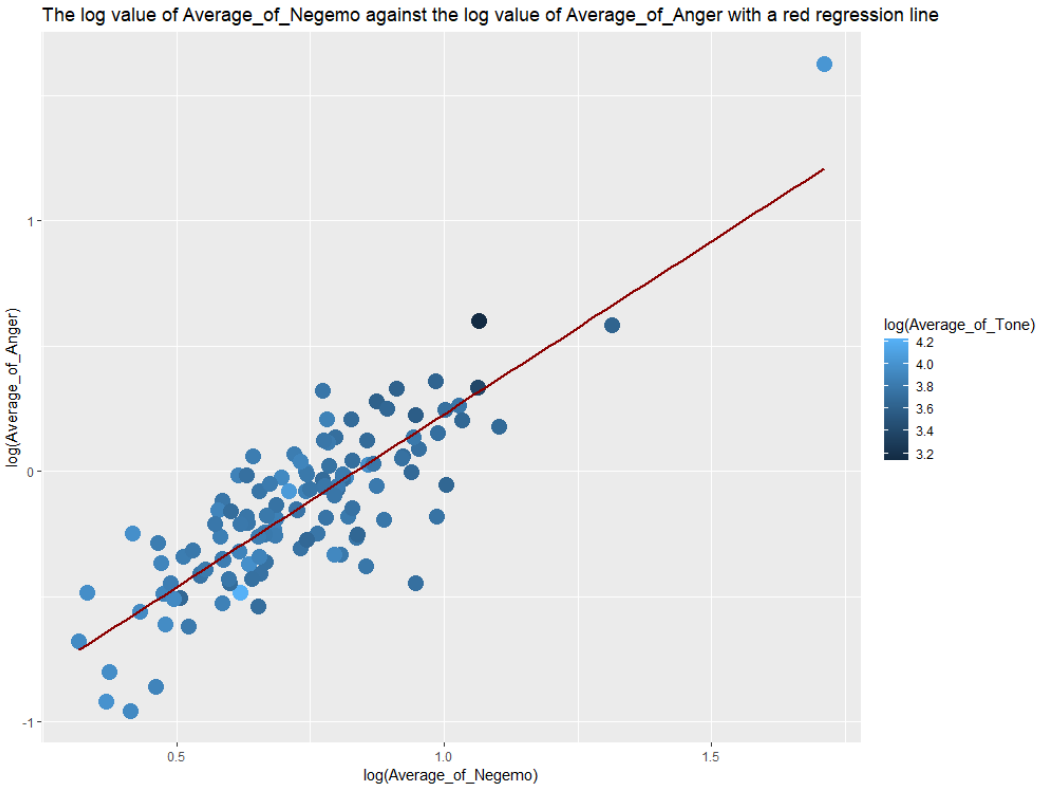

Figure 11. shows the log value of Average_of_Negemo against the log value of Average_of_Anger with a red regression line

Figure 11 illustrates the relationship between the logarithm of the Average_of_Negemo variable and the logarithm of the Average_of_Anger variable, which exhibits a positive correlation. By logarithmizing the variables, we achieve a more linear model, allowing us to better understand their relationship. The scatter plot and regression line highlight this positive correlation, indicating that as log(Average_of_Negemo) increases, log(Average_of_Anger) tends to increase as well. The red regression line provides an estimate of this relationship, with a slope of 1.37883 and an intercept of -1.15106.

In addition, the Average_of_Tone variable, represented by the color of the points, also exhibits a general increase as log(Average_of_Anger) and log(Average_of_Negemo) increase, as indicated by the darker color of the points (excluding the outlier). This suggests a potential relationship between the three variables.

Further insights into the regression model can be found in Figure 12, providing more details on the relationship between log(Average_of_Anger) and log(Average_of_Negemo) through the regression equation.

```
> fitted = lm(log(webforum_all_linguistic_variables$Average_of_Anger)~log(webforum_all_linguistic_variables$Average_of_Negemo))
> summary(fitted)

Call:
lm(formula = log(webforum_all_linguistic_variables$Average_of_Anger) ~
    log(webforum_all_linguistic_variables$Average_of_Negemo))

Residuals:
     Min       1Q   Median       3Q      Max
-0.60151 -0.10749 -0.00296  0.10728  0.41739

Coefficients:
                                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                                       -1.15106    0.06190  -18.59   <2e-16 ***
log(webforum_all_linguistic_variables$Average_of_Negemo)  1.37883    0.08176   16.86   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1789 on 118 degrees of freedom
Multiple R-squared:  0.7068,    Adjusted R-squared:  0.7043
F-statistic: 284.4 on 1 and 118 DF,  p-value: < 2.2e-16
```

Figure 12. shows the details of the regression model between log (Average_of_Anger) and log (Average_of_Negemo)

In conclusion, our analysis reveals a positive correlation between Average_of_Negemo and Average_of_Anger over the long term, as well as several other pairs of linguistic variables indicated by the heatmap. This suggests that the levels of these variables tend to vary in a similar direction. However, it's important to note that correlations do not imply causation, and further investigation is required to understand the underlying factors driving these relationships.

Question b.1

Using the relevant linguistic variables, is it possible to see whether threads are happier or more optimistic than other threads, or the forum in general, at different periods in time.

To narrow down the analysis, I focused on the **top ten** threads with the highest number of posts. To identify these threads, I grouped the data by ThreadID and calculated the frequency of each thread using the summarize function. By sorting the resulting data frame and using the head() function, I obtained the ThreadID for the top ten threads with the most posts.

Next, I needed to determine which linguistic variables to consider for this question. Based on the assignment specification, it is mentioned that "sad" is included as part of the "negemo" variable. Although it does not explicitly state whether "anx" and "anger" are included in "negemo," I assumed that these variables are already accounted for within "negemo." Therefore, for this question, I decided to focus on the average values of the "posemo" and "negemo" variables over time for each thread. These variables indicate the level of positive and negative emotions in each post within a thread, which allows me to assess the overall happiness of a thread.

Considering the number of threads and the amount of data, including the average values of "posemo" and "negemo" for all top ten threads in a single image would be overwhelming. Hence, I opted to divide the graphs into two images: the first image displaying the top four threads and the second image featuring threads ranked five to ten. To accomplish this, I filtered the data in the "webforum" data frame accordingly and used the group by function to calculate the average values of "posemo" and "negemo" for each thread within each month and year (by grouping based on ThreadID and Date).

Figure 13 illustrates the resulting data frame that will be used to plot the graphs. This data frame comprises the average values of "posemo" and "negemo" for the four threads over time. It is important to note that I will perform the same procedure for threads ranked five to ten to obtain the data frame for the second image.

```
> top_4_thread
# A tibble: 87 x 4
# Groups:   ThreadID [4]
   ThreadID Date       Average_Posemo Average_Negemo
      <int> <date>              <dbl>          <dbl>
 1   127115 2004-04-01           0              5.98
 2   127115 2004-05-01           2.54           2.54
 3   127115 2004-06-01           1.54           3.08
 4   127115 2004-08-01           3.23           0
 5   127115 2004-09-01           0.21           1.68
 6   127115 2005-01-01           0.57           1.14
 7   127115 2006-05-01           1.78           2.78
 8   127115 2006-08-01           2.31           1.91
 9   127115 2006-09-01           2.24           2.62
10   127115 2007-01-01           2.42           0.82
# ... with 77 more rows
```

Figure 13. shows the summarized average posemo and negemo scores per month for each thread, obtained by grouping the data using ThreadID and Date

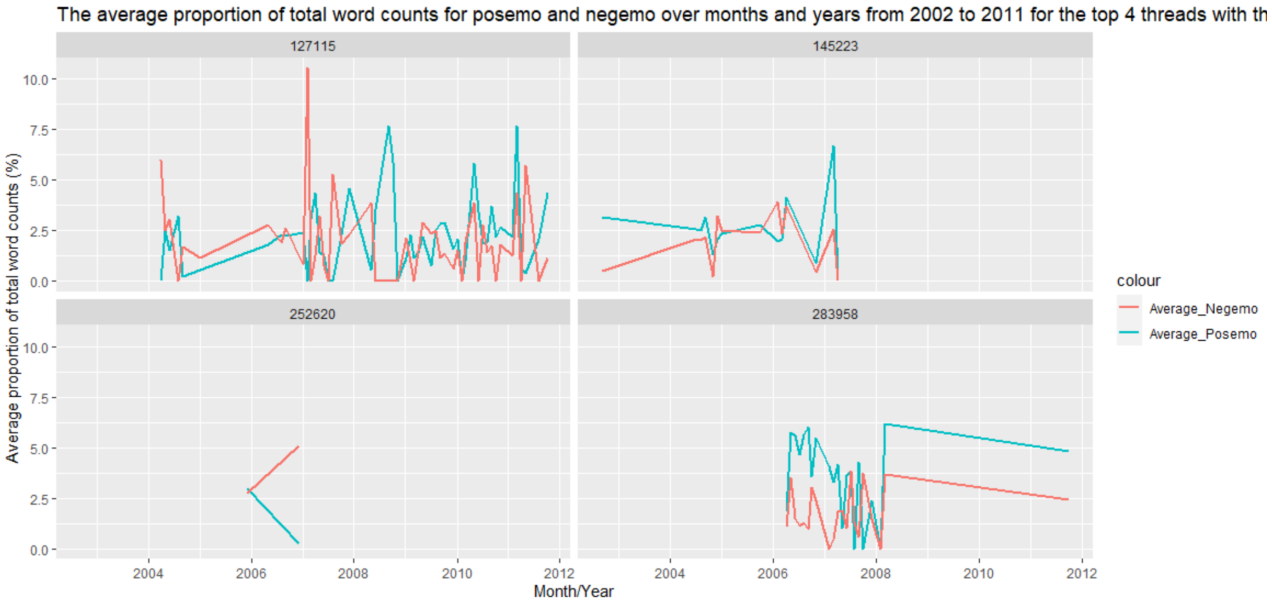

Figure 14. shows the evolving word count distribution of positive and negative emotions across the top four threads from 2002 to 2011

In Figure 14, we can observe the average proportion of positive emotions (posemo) and negative emotions (negemo) over a period of time for the top four threads with the highest number of posts. Let's analyze each thread individually.

Starting with the thread identified by ID 127115, we notice that both posemo and negemo become active around 2004 and remain active until 2012. However, it is challenging to distinguish between the values of these two variables for this thread, as they are quite close to each other. Throughout this period, the average proportion of negemo and posemo mostly ranges between 0% and 5.5%. Therefore, despite the relatively high average posemo, I would not consider this thread to be optimistic, as the average negemo is equally prominent.

Moving on to the second thread (ID 145223), it surprisingly exhibits similar patterns to the previous thread. The average posemo and average negemo are quite comparable, as shown in the image. However, before 2005, the average posemo was slightly higher than the average negemo. Although this thread has some positive sentiment, it is not significantly different from its negative sentiment.

Now let's examine the thread with ID 252620. Throughout the entire period, the average negemo consistently appears to be higher than the average posemo. Consequently, it is evident that this thread lacks optimism entirely.

Lastly, we analyze the last thread among the top four (ID 283958). Excluding the year 2007, the average posemo for this thread is higher than the average negemo by approximately 2.5%. Based on this information, it appears that the last thread (ID 283958) is the most optimistic and happiest among the top four threads.

I have identified these four threads to determine which one is the most optimistic. This analysis is crucial because I plan to compare the most optimistic thread among these four with the most optimistic thread among the top 5 to 10 threads. By conducting a hypothesis test, I aim to establish which thread stands out as the most optimistic and happiest among the top ten.

I want to highlight that I will only compare one pair of threads in this analysis. The constraint of a page limit prevents me from conducting a hypothesis test for every possible pair of threads to determine the happiest one. Therefore, I have chosen to focus on comparing the thread that emerges as the most optimistic and happiest among these top four threads with another thread from the top 5 to 10 threads.

With these improvements, the text provides a clear explanation of the purpose behind identifying the most optimistic threads and highlights the limitations of conducting a hypothesis test for every pair of threads.

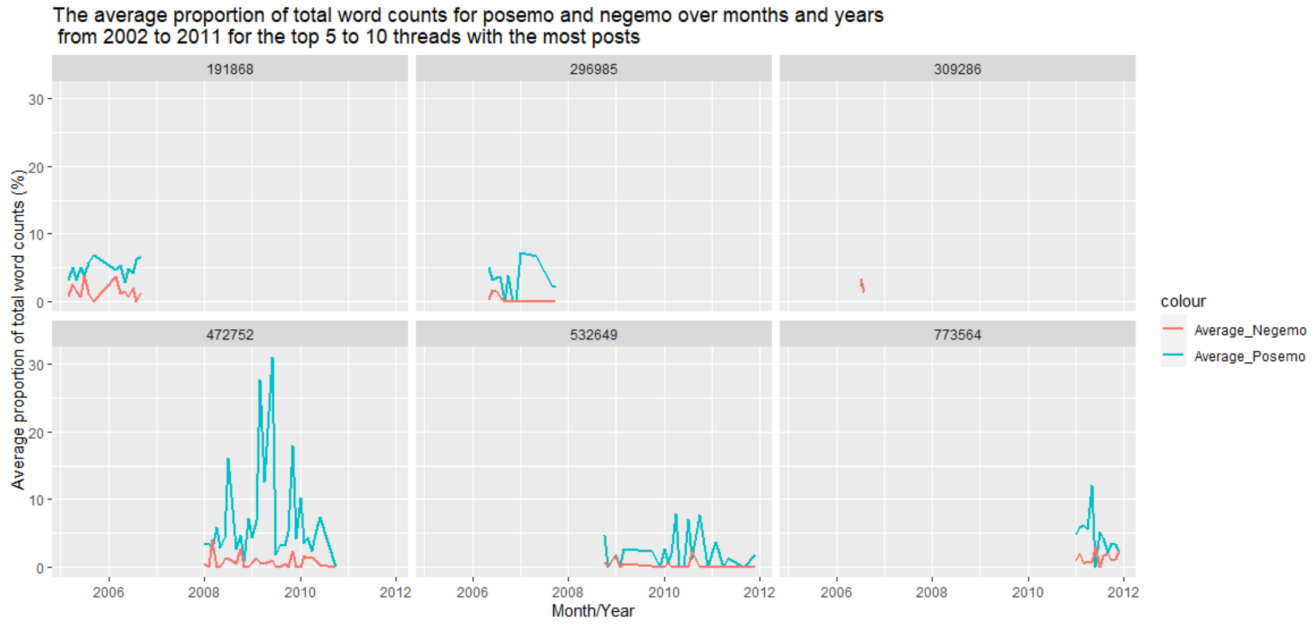

Figure 15. shows the average proportion of total word counts for posemo and negemo over months and years from 2002 to 2011 for the top 5 to 10 threads with the most posts

Figure 15 displays the average proportion of total word counts for positive emotions (posemo) and negative emotions (negemo) from 2002 to 2011 for the top 5 to 10 threads with the most posts. By examining this figure, we can determine the most optimistic thread among these six threads.

Notably, with the exception of the thread with ID 309286, the remaining five threads consistently show a higher average posemo compared to negemo throughout the entire period. Threads with ID 191868, 296985, 532649, and 773564 exhibit slightly higher average posemo values within the range of 0% to 10% for the majority of the time. However, this pattern is not observed in the thread with ID 472752. From the figure, it is evident that this thread has significantly higher average posemo values compared to the other threads, even surpassing 30% in 2009. In summary, based on the details presented in the figure, we can conclude that the thread with ID 472752 is the most optimistic and happiest among these six threads.

Now that we have identified the most optimistic and happiest thread among the top four threads (ID 283958) and among the top 5 to 10 threads (ID 472752), the next step is to compare these two threads to determine which one is more optimistic. While my current observation suggests that the thread with ID 472752 is more optimistic and happier than the thread with ID 283958, it is important to provide evidence to support this finding.

To gather evidence, I will perform two hypothesis tests: one comparing the average posemo over time and another comparing the average negemo over time for both threads. For the first hypothesis test, I will filter the data frame used to plot the graph and calculate the average posemo for each month of a year for both threads. I will then use the t.test function to perform the test. In this case, the null hypothesis is that the average posemo for the thread with ID 472752 is greater than the average posemo for the thread with ID 283958 throughout the entire period, with a 95% confidence level.

The results of this hypothesis test are presented in Figure 16. From the results, we can observe that the p-value is 0.9889. This indicates that if the null hypothesis were true, there would be approximately a 98.9% chance of observing a sample with a difference from the null as extreme or more extreme than the one we observed. Therefore, there is insufficient/weak evidence to reject the null hypothesis, suggesting that the average posemo for the thread with ID 472752 is greater than the average posemo for the thread with ID 283958 throughout the entire period.

Based on this analysis, we can conclude that the thread with ID 472752 is indeed happier than the thread with ID 283958.

```
> t.test(average_posemo_negemo_for_283958$Average_Posemo,average_posemo_negemo_for_472752$Average_Posemo,alternative = "greater")

        Welch Two Sample t-test

data:  average_posemo_negemo_for_283958$Average_Posemo and average_posemo_negemo_for_472752$Average_Posemo
t = -2.3975, df = 33.613, p-value = 0.9889
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -5.91264      Inf
sample estimates:
mean of x mean of y
 3.629541  7.096304
```

Figure 16. shows that result of the hypothesis testing using t.test in R

Similarly, for the second hypothesis test, I will calculate the average negemo for each month of a year for both threads by filtering the data frame used to plot the graph. Then, I will employ the t.test function to perform the test. The null hypothesis states that the average negemo for the thread with ID 283958 is greater than the average posemo for the thread with ID 472752 throughout the entire period, with a 95% confidence level.

The results of this hypothesis test are presented in Figure 17. The p-value obtained is 0.9972. This implies that if the null hypothesis were true, there would be approximately a 99.7% chance of observing a sample with a difference from the null as extreme or more extreme than the one we observed. Consequently, there is insufficient/weak evidence to reject the null hypothesis, suggesting that the average negemo for the thread with ID 283958 is greater than the average posemo for the thread with ID 472752 throughout the entire period.

Based on this analysis, we can conclude that the thread with ID 283958 is indeed more pessimistic than the thread with ID 472752.

```
> t.test(average_posemo_negemo_for_283958$Average_Negemo,average_posemo_negemo_for_472752$Average_Negemo,alternative = "less")

        Welch Two Sample t-test

data:  average_posemo_negemo_for_283958$Average_Negemo and average_posemo_negemo_for_472752$Average_Negemo
t = 2.9365, df = 37.753, p-value = 0.9972
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     -Inf 1.489001
sample estimates:
mean of x mean of y
1.7673513 0.8214885
```

Figure 17. shows that result of the hypothesis testing using t.test in R

By considering the results of both hypothesis tests conducted above, we can confidently state that the thread with ID 472752 is not only happier but also more optimistic and less pessimistic than the thread with ID 283958. Therefore, based on these findings, we can conclude that the thread with ID 472752 is the most optimistic thread among the top ten threads.

Question c.1

Create a non-trivial social network of all authors who are posting over a particular time. For example, over one month. To create this, your social network should include at least 30 authors, some of whom will have posted to multiple (2 or more) threads during this period. Your social network should be connected, although some authors may be disconnected from the main group. Present your result as a network graph.

To begin constructing the social network, it is essential to examine the number of posts for each month of a year. This analysis will help in selecting an optimal month that strikes a balance between having a sufficient number of posts (to avoid an overcrowded network graph) and ensuring a minimum of 30 authors who have made posts during that month. To achieve this, I will group the webforum dataset by the Date column and utilize the summarize function to count the number of rows within each group.

After careful consideration, I have chosen to build the social network using the data from July 2009. This specific timeframe satisfies the requirements of having at least 30 authors and allows for a manageable number of vertices (representing unique AuthorIDs) and edges in the resulting network graph. Once the timeframe is established, I will create a new data frame by filtering the webforum dataset to include only the rows posted during July 2009. Subsequently, I will plot the social network graph based on this refined data frame.

By following this approach, I can ensure that the resulting network graph will provide a clear and meaningful representation of the relationships among authors while avoiding issues of overcrowding.

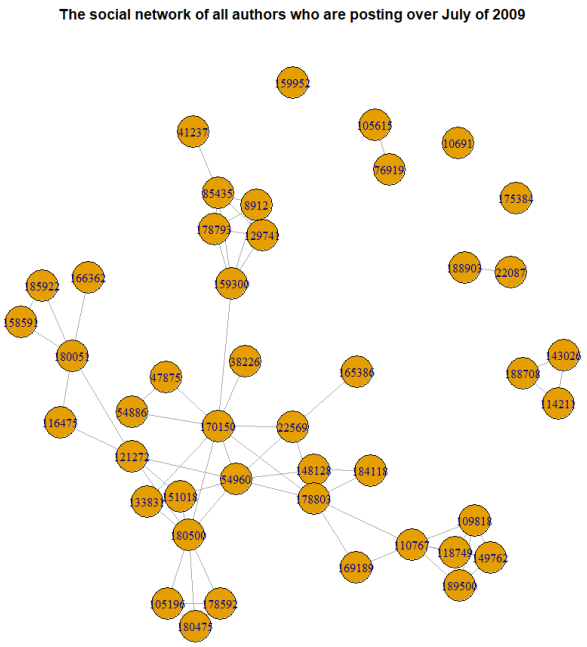

Figure 18. shows the social network of all authors who are posting over July of 2009

The resulting network graph is displayed in Figure 18. In this graph, each vertex represents an AuthorID, while the edges connecting the vertices indicate that these authors posted to the same thread during the selected time frame (July 2009). The graph comprises a total of 67 edges and 44 vertices, indicating the inclusion of 44 unique authors.

It's important to note that this is an unweighted graph, meaning that the number of times two vertices posted to the same thread (with different ThreadIDs) within the specified time frame is not considered. Each edge represents the occurrence of the two vertices posting to the same thread, regardless of the number of times they did so.

From Figure 18, we can observe that the majority of vertices are interconnected, forming a main group. However, approximately 10 vertices appear disconnected from this main group. After obtaining the network graph, we can analyze it from an overall perspective by examining its diameter, average path length, graph density, transitivity, and degree distribution. These metrics are presented in Figure 19.

The diameter of the network graph is found to be 7, indicating the longest shortest path between any two vertices. The average path length, which represents the average number of steps required to travel between any two vertices, is approximately 3.28. The graph density, which measures the proportion of possible edges present in the graph, is approximately 0.07. Furthermore, the transitivity, which measures the likelihood of interconnectedness among a vertex's neighbors, is approximately 0.52.

To visualize the distribution of degrees for the vertices in the network graph, we can plot a histogram, as shown in Figure 20. The histogram reveals that the majority of vertices (around 22) have a degree of 0 to 2, while approximately 12 vertices have a degree of 2 to 4. Only a minority of vertices exhibit a degree between 4 to 10.

These analyses provide valuable insights into the structure and characteristics of the network graph, enabling a deeper understanding of the relationships and connectivity among authors during the specified time frame.

```
> diameter(g)
[1] 7
> average.path.length(g)
[1] 3.275618
> graph.density(g)
[1] 0.07082452
> transitivity(g)
[1] 0.5234043
> # get.adjacency(g) # the adjacency matrix is too big to show in the console
> hist(degree(g),breaks=5,col = "grey")
>
```

Figure 19. shows all the factors regarding the network graph shown in figure 18

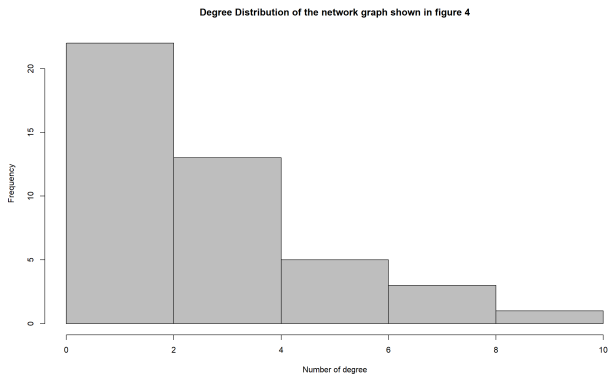

Figure 20. shows the degree distribution of the network graph shown in figure 18

Question c.2

Identify the most important author in the social network you created.

In order to determine the most important author within the social network constructed in the previous question, I will begin by calculating various network centrality measures for each vertex. These measures include degree centrality, betweenness centrality, closeness centrality, and eigenvector centrality. The results of these calculations are stored in a data frame, as depicted in Figure 21, which showcases the top few rows of the data frame containing the network centrality measures for each vertex.

```
> head(stats)
       degree betweenness   closeness eigenvector
149762      4    0.000000 0.007812500   0.2150063
129741      4    0.000000 0.008333333   0.2115239
148128      3    5.333333 0.010526316   0.4267857
170150      9  266.500000 0.014285714   0.9989390
175384      0    0.000000         NaN   0.0000000
105615      1    0.000000 1.000000000   0.0000000
```

Figure 21. shows the top six lines of the data frame that stores the network centrality measures for each vertex

Once the data frame is generated, I can proceed to sort it based on each network centrality measure. Upon examining the sorted data frame, it becomes apparent that the author with ID 170150 holds the highest level of importance within the social network. This conclusion is drawn from the fact that author 170150 consistently ranks among the top performers across multiple centrality measures. Specifically, they achieve the top rank in terms of betweenness and degree centrality, second place in eigenvector centrality, and eighth place in closeness centrality. This overall performance surpasses that of all other authors in the network. Therefore, it can be concluded that the author with ID 170150 is the most important individual within the social network established in the previous question.

Looking at the language they use, can you observe any difference between them and other members of their social network?

After identifying the most important author, the next step is to compare the language used by this author with other authors who posted during the same time frame. To begin, I will calculate the average values of linguistic variables for the posts made by the author with ID 170150 specifically during the month of July 2009.

Linguistic variable	Analytic	Clout	Authentic	Tone	ppron	i	we	you	shehe
Average value (%)	70.75	65.9	27.3	37.7	5.00	0.99	0.53	1.34	0.00
Linguistic variable	they	posemo	negemo	anx	anger	sad	focus past	focus present	focus future
Average value (%)	2.13	2.14	1.45	0.17	0.44	0.17	1.83	11.51	0.78

The table above displays the average percentage/percentile of linguistic variables for the posts made by author 170150 during July 2009. These values were generated using R code. It's important to note that the "WC" (word count) variable is not included in this analysis. This decision was made because the word count alone may not provide meaningful insights into the language differences between authors, as it simply represents the number of words. Additionally, I plan to plot histograms to visualize the distribution of values for all linguistic variables among the authors in the social network, excluding author 170150. Consequently, excluding the "WC" column helps maintain a manageable x-axis range, ensuring clarity in the subsequent graphs (i.e., Figure 22 on the following page) by avoiding extremely small bars. Therefore, for the purpose of analyzing the language differences between author 170150 and other authors, the "WC" column has been omitted.

To compare the language used by author 170150 with that of other authors who posted during the same time frame, histograms will be plotted for each linguistic variable (except WC) to depict the distribution of values among all authors, excluding author 170150.

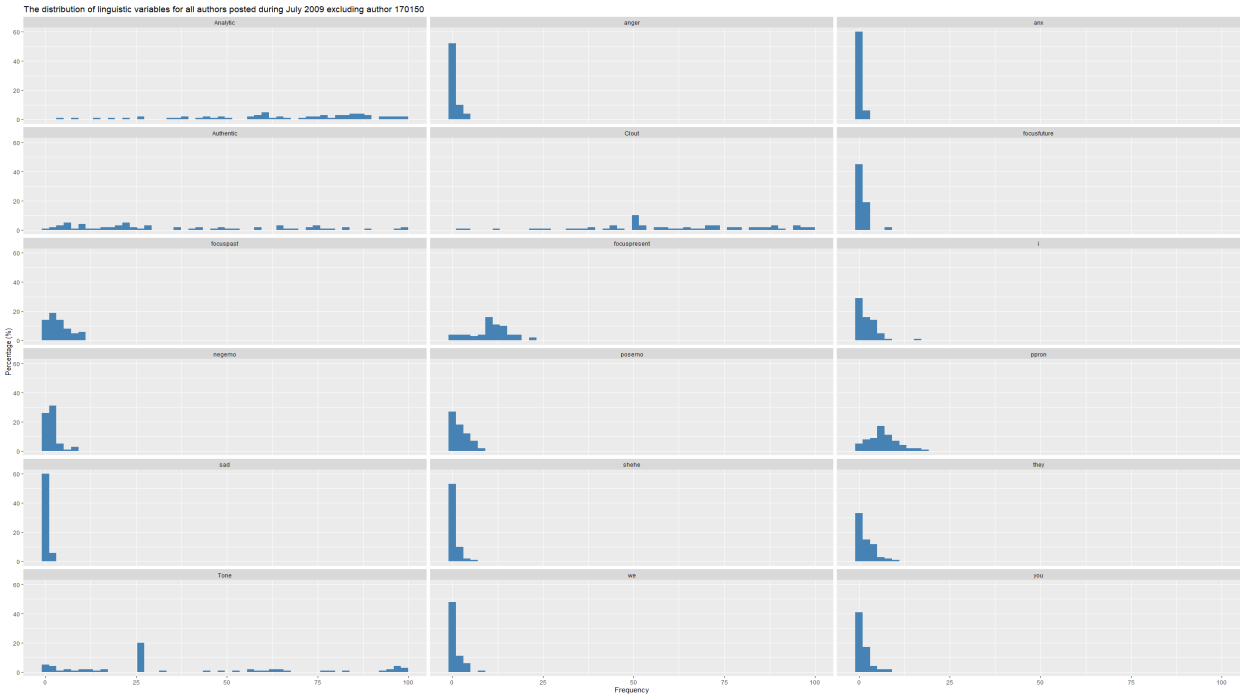

Figure 22. shows the distribution of linguistic variables for all authors who posted during July 2009, excluding author 170150

Figure 22 displays the distribution of linguistic variables for all authors who posted during July 2009, excluding author 170150. Upon analyzing the histograms in Figure 22, we observe that there are no distinct patterns in the distribution of analytic, authentic, tone, and clout variables. Therefore, it is challenging to identify clear differences between author 170150 and other authors in terms of these variables.

For the variables anger, anx, focusfuture, focuspast, i, negemo, posemo, sad, shehe, they, we, and you, the ranges of these variables for other authors fall within the 0% to 12.5% range. Interestingly, author 170150's values for these linguistic variables also fall within this range, making it difficult to discern differences between them.

Furthermore, the variables focuspresent and ppron exhibit a larger range among other authors, spanning from 0% to 25%. Surprisingly, the average values of focuspresent and ppron for author 170150 also fall within this range. This piques my interest in investigating whether there is a significant difference in the means of focuspresent and ppron between author 170150 and other authors. Consequently, I will conduct two hypothesis tests to support my findings. It's important to note that due to space limitations, I have selected only two variables, focuspresent and ppron, for hypothesis testing.

For the first hypothesis test, I will filter the "data_long" data frame (used to create Figure 22) and the "post_in_July_of_2009_all_variables" data frame to obtain the values of focuspresent for author 170150 and other authors. Then, I will utilize the t.test function to test the null hypothesis that the mean of focuspresent does not differ between author 170150 and other authors during July 2009. The confidence level is set at the default value of 95%. The result of this hypothesis test is provided in Figure 23.

Based on the result in Figure 23, the p-value is 0.7064. This indicates that if the null hypothesis were true, there would be a 70.6% chance of observing a sample with an extreme or more extreme difference from the null hypothesis. Insufficient/weak evidence is found against the null hypothesis that the mean of focuspresent does not differ between author 170150 and other authors during July 2009, as the p-value is greater than 0.05. Therefore, we cannot reject the null hypothesis.

```
> data_for_focus_present_other_author = data_long[(data_long$name=="focuspresent"),2]
> data_for_focus_present_other_author
 [1]  9.52 10.53 14.67  9.77  7.14 13.22 14.10  6.21 12.50 10.00 10.32  0.82  8.33 13.33 11.59  2.98  3.01 12.64 18.92 13.64 10.71 16.92
[23]  2.73  0.00 12.00 16.67 10.84  2.86 11.29  4.48  5.39  9.38 11.11  8.00  0.00 10.89 10.53 12.50 15.83 13.89  8.70 12.42  3.92 13.51
[45] 15.49 13.79  4.26 22.22  0.00  6.93 12.84 11.76 18.18 14.29  9.30 12.06 15.00 18.52 10.08 10.26 21.62 12.63  5.00 17.58 10.39  9.43
> data_for_focus_present_author_170150 = post_in_July_of_2009_all_variables[(post_in_July_of_2009_all_variables$AuthorID == 170150),22]
> data_for_focus_present_author_170150
[1]  4.76 14.29 12.68  9.26 21.74 13.11  4.76
> t.test(data_for_focus_present_other_author,data_for_focus_present_author_170150,alternative="two.sided")

        Welch Two Sample t-test

data:  data_for_focus_present_other_author and data_for_focus_present_author_170150
t = -0.39258, df = 6.9651, p-value = 0.7064
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.443977  4.610557
sample estimates:
mean of x mean of y
 10.59758  11.51429
```

Figure 23. shows that result of the hypothesis testing using t.test in R

For the second hypothesis test, I will filter the "data_long" data frame and the "post_in_July_of_2009_all_variables" data frame to obtain the values of ppron for author 170150 and other authors. Subsequently, I will employ the t.test function to test the null hypothesis that the mean of ppron does not differ between author 170150 and other authors during July 2009. The confidence level is set at the default value of 95%. The result of this hypothesis test is provided in Figure 24.

According to the results in Figure 24, the p-value is 0.2652. This implies that if the null hypothesis were true, there would be a 26.5% chance of observing a sample with an extreme or more extreme difference from the null hypothesis. Insufficient/weak evidence is found against the null hypothesis that the mean of ppron does not differ between author 170150 and other authors during July 2009, as the p-value is greater than 0.05. Therefore, we cannot reject the null hypothesis.

```
> data_for_ppron_other_author = data_long[(data_long$name=="ppron"),2]
> data_for_ppron_other_author
 [1]  7.94 13.45 10.00  1.50  7.14 12.40  3.85 10.34  6.25  5.56  5.56  1.64 16.67  0.00  8.70  5.36  1.50  6.59  5.41  9.09  0.00 10.77
[23]  2.73  0.00  6.00  8.33  4.82  0.71  6.45  7.46  7.78  4.69  9.26  4.00  1.37  4.95 15.79  7.50  5.00 10.19  5.80  7.45  1.96  5.41
[45]  7.04  8.05  4.96  0.00  5.88  2.97  6.42 11.76 18.18 11.69 13.95 11.28 10.00 11.11  3.36  5.98  2.70  5.26  5.00  6.59  5.19  7.55
> data_for_ppron_author_170150 = post_in_July_of_2009_all_variables[(post_in_July_of_2009_all_variables$AuthorID == 170150),10]
> data_for_ppron_author_170150
[1] 0.00 2.86 5.63 5.56 8.70 9.84 2.38
> t.test(data_for_ppron_other_author,data_for_ppron_author_170150,alternative="two.sided")

        Welch Two Sample t-test

data:  data_for_ppron_other_author and data_for_ppron_author_170150
t = 1.2002, df = 7.8118, p-value = 0.2652
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.585280  4.996579
sample estimates:
mean of x mean of y
 6.701364  4.995714
```

Figure 24. shows that result of the hypothesis testing using t.test in R

Furthermore, it is noteworthy that the 95% confidence intervals for both hypothesis tests include zero. This indicates that we cannot disregard the possibility of there being no significant difference at a population level between author 170150 and other authors.

In conclusion, considering the results of both hypothesis tests and the insights gained from Figure 22, it can be concluded that the language used by the most important author (author 170150) does not significantly differ from that of other authors in the social network.