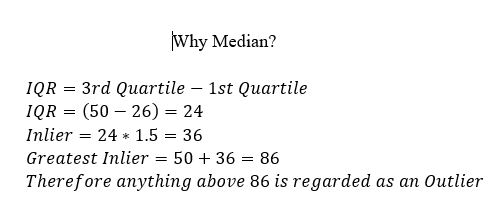
**FIT3152 Assignment 1**

The aim of this report is to provide an analysis of the webforum dataset which was given to us on moodle. The dataset contains x variables and is based on linguistic analysis of huge numbers of threads and posts between the year 2002 and 2011 which was conducted using Linguistic Inquiry and Word Count (LIWC). The main purpose of this report is to investigate whether participants in an online forum who are communicating directly via threads uses similar language and whether the language used by participants in the forum changes over time.

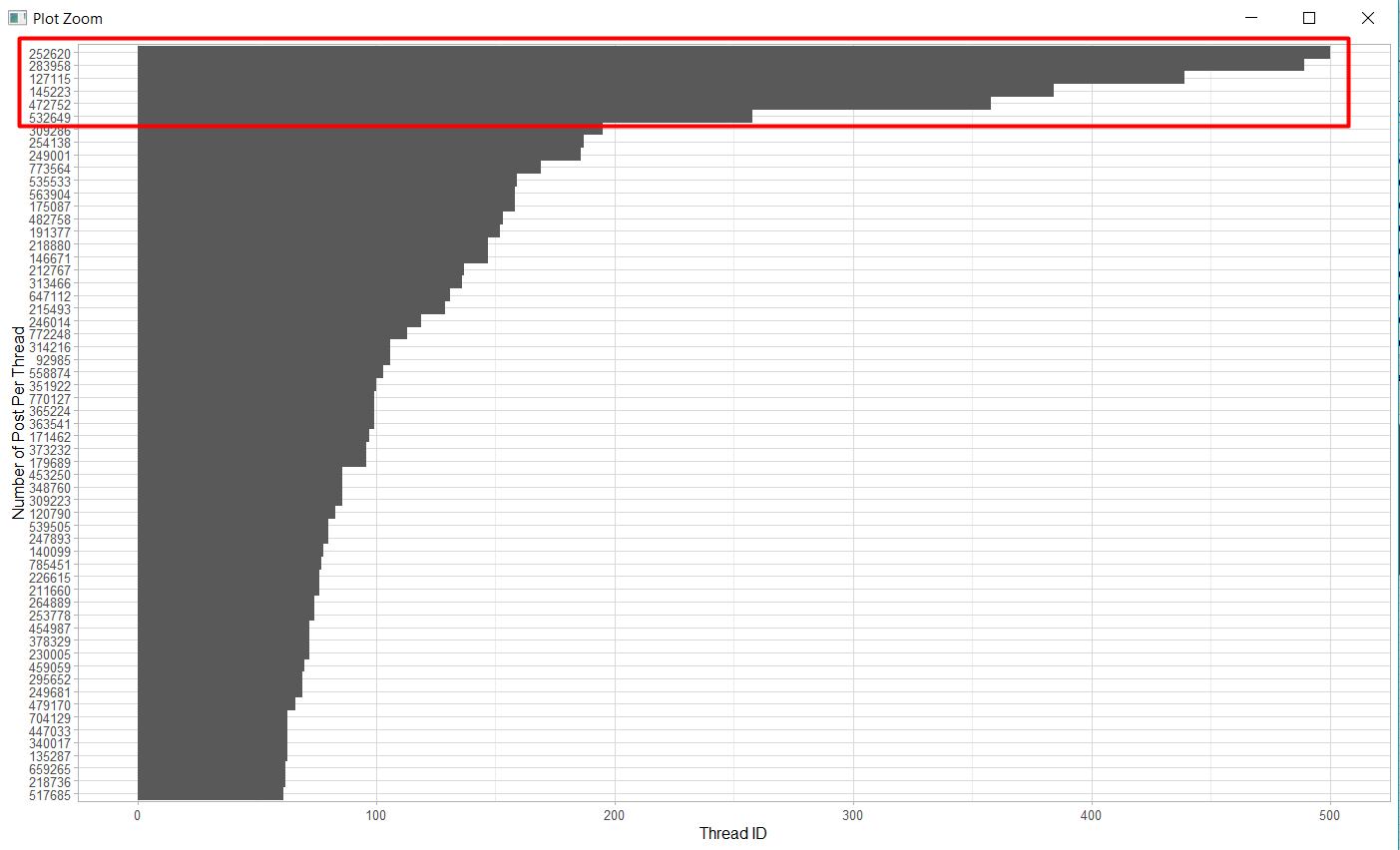
1. **Subset of the data to be analysed**



Posts that has 0 Word Count (WC) is excluded because it contains images or diagrams and will not contribute to the report which analyses about language (text). Posts are then grouped by their threadID.

In grouping the threadID, number of authors in each thread is also taken into consideration (count number of authors in each thread and take average of it. Median is chosen as average because there exist outliers [see figure 1]. Threads which has lower number of authors than the median are not included). This is done to ensure that there are several different authors communicating with each other.

*Figure 1: Why Median?*

Next, we analyze number of posts in the threads and take the top 6 threads base on the graph in figure 2 (Thread with more than 200 posts). We then take the 6 threads (252620, 283958, 127115, 145223, 472752, and 532649) mentioned above and use it for this assignment

*Figure 2: Number of Post in All Threads*

## Analysing a User’s Language in Different Threads

## Using the top 6 threads above, Analytic is chosen as the attribute to be analysed. Thread 127115 is chosen as the thread to be analysed as it has the top mean of analytic with 84.038 based on Figure 3 (shown by the color red).

## Going further into thread 127115, year 2009 is chosen because it has the most number of post compared to the other years in this thread (refer to Figure 4).

## Thread 127115 in 2009 was then analysed by AuthorID to get the author with the most number of posts. From Figure 5, author 47875 with the colour orange has most number of post. A conclusion can be derived that author 47875 uses analytic in thread 127115. Next step will be analysing author 47875’s language usage in other thread.

*Figure 3: Mean of Analytics Between Top 6 Threads*

## 

## 

*Figure 5: Posts in Thread 127115 by AuthorID*

*Figure 4: Number of Post by Year in Thread 127115*

## Author 47875 has posted in 3 different threads (472752, 145223, and 532649) which will then be analyzed except thread 532649 because author 47875 only posted once in that thread therefore it is assumed that the sample is not enough. After aggregating the top 6 threads, posemo (positive emotion) was found to be the significantly higher in thread 472752 compared to the other threads (refer to Figure 6) and anx (anxiety) was found to be significantly higher in thread 145223 compared to the other threads (refer to Figure 7).

## T-test was also conducted to further support the fact that thread 472752 has higher posemo than the other threads in the top 6 threads (refer to appendix t.test 1) and thread 472752 has higher anx than the other threads in the top 6 threads (refer to appendix t.test 2).

## 

*Figure 7: Summary of Top 6 Threads Showing Highest anx for Thread 145223*

*Figure 6: Summary of Top 6 Threads Showing Highest Posemo for Thread 472752*

## Another t.test with the confidence level of 0.95 is then conducted on both author 47875’s post in the thread 472752 and 145223 to see whether his/her post still uses Analytic or has change according to the thread’s theme/attribute. Referring to Figure 8 and Figure 9, p-value was less than the critical value (0.05) which as a result accepts the alternative hypothesis. The conclusion is author 47875’s post in thread 472752 has higher positive emotions compared to his/her posts in other threads and author 47875’s post in thread 145223 has higher anxiety compared to his/her posts in other threads which indicates that his/her language change according to the thread’s theme/attribute.

*Figure 8: t.test of Author 47875’s Post in Thread 472752 with His/Her Other Posts Regarding Posemo*

## 

*Figure 9: t.test of Author 47875’s Post in Thread 145223 with His/Her Other Posts Regarding anx*

## To further support the conclusion above, a random thread is taken from the top 6 threads with most posts using random sampling. Thread 472752 was chosen then as the random thread (Figure 10). The same approach as in the analytic thread was chosen which is taking the author with most post in thread 472752 (author 39170) and seeing his/her posts in another thread to see his/her behavior regarding language usage (Figure 11). Author 39170 has posted in thread 127115 which is the analytic thread therefore the method before can be reused as well. T.test with confidence level of 0.95 was conducted to test the significance of analytics on average and the result was that the p-value was less than the critical value (0.05) which as a result accepts the alternative hypothesis (Figure 12). This finding further supports the analysis that user’s language in the posts changes according to the language used by others in the thread.

*Figure 10: Random Sampling of Thread*

## 

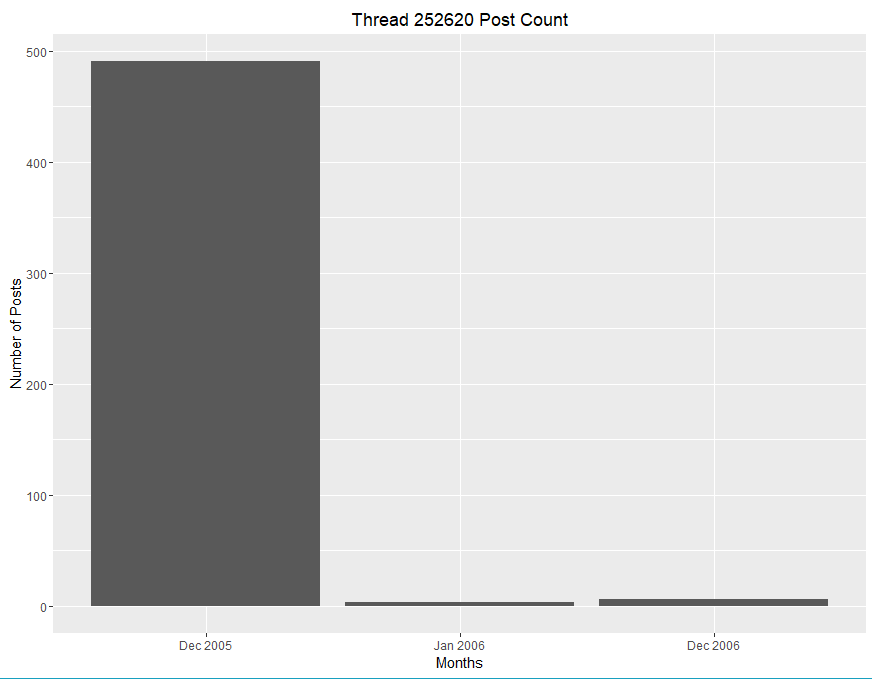
*Figure 11: Author with Most Post in Thread 472752 and His/Her Posts in Other Threads*

## 

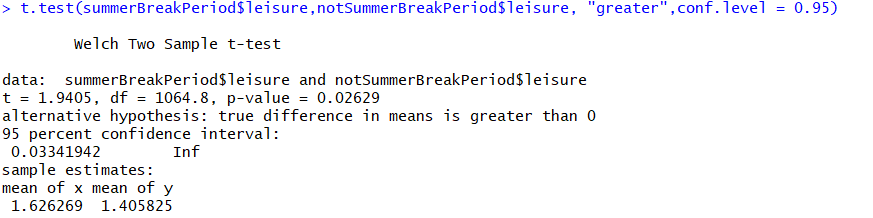
*Figure 12: t.test of Author 39170’s Post in Thread 127115 with His/Her Other Posts Regarding anx*

**----------------------------------------UNTIL HERE---------------------------------------**

**Analyzing Language Change Over the Time**

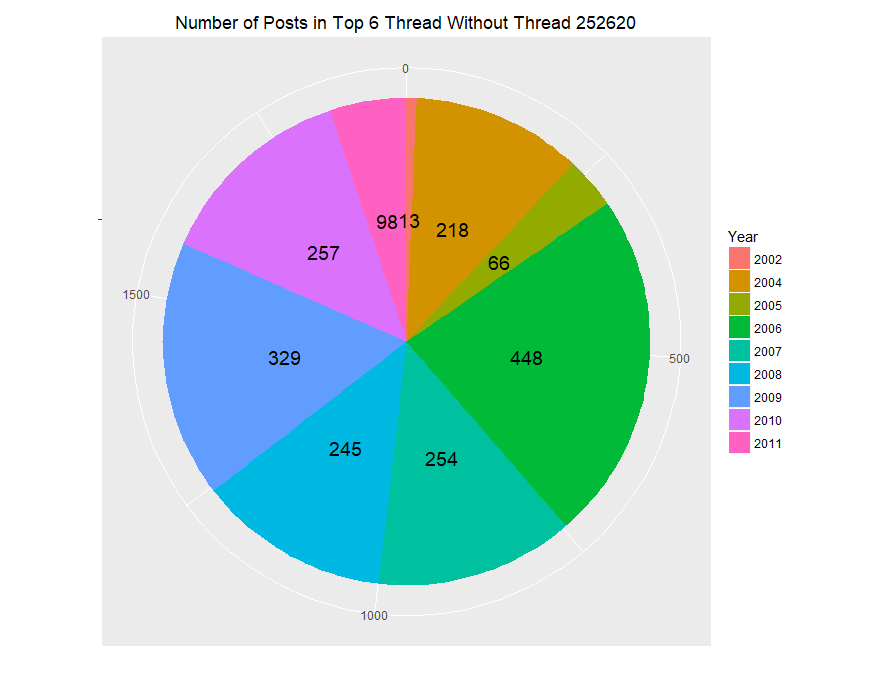
We took the top 6 threads and noticed that thread 252620 only has posts on Dec 2005, Jan 2006 and Dec 2006 where in Dec 2005 it has a significant number of posts as can be seen in the diagram. Therefore, we eliminate the thread 252620 to prevent bias and affect the results and findings from our analysis.

We want to compare the number of posts that contains leisure between summer break vacation period from within the rest since according to *http://www.pewinternet.org/2015/08/19/mobile-messaging-and-social-media-2015/2015-08-19\_social-media-update\_04/* (2015), the demographics of people using online forums are 23% of which they are between the age of 18-29 which showed that there are a majority of college or university students using online forums and summer vacation for most countries (except Australia and New Zealand) are between June to September (*https://www.summerdiscovery.com/blog/2013-11-15/summer-vacation-around-the-world*), hence we used this assumption for the basis of our analysis. We performed a t-test to prove that the number of posts that contains leisure increases during the summer break period compared to the rest of the months. The result can be found below:

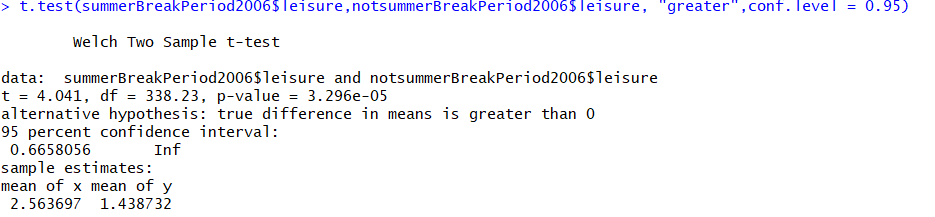


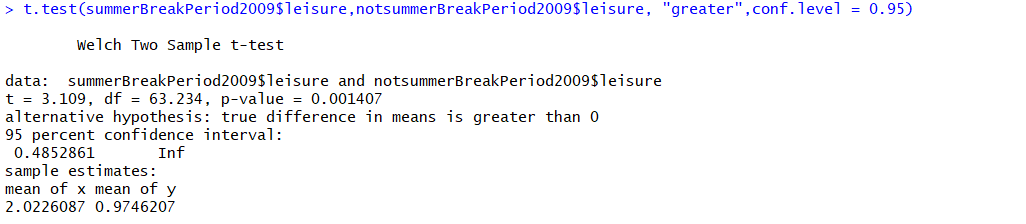
We performed a t-test on 95 percent confidence level and found out that the p-value is lower than the critical value which is 0.05. Consequently, we can reject the null hypothesis and therefore accept the alternative hypothesis where the number of posts about leisure on summer break period is more than the rest of the months in average.

We wanted to further prove the analysis by concentrating on the top 2 years where the number of posts is the highest in the top 5 threads. We found the year 2006 and 2009 to be the years that have the most posts as can be seen from the pie chart



We took the top 2 years because we chose the threshold to be 300. We performed another t-tests on 2006 as well as 2009 and found out the results support our initial analysis which can be seen below:





From above, the p-values are smaller than the critical value (0.05) which means we can accept the alternative hypothesis and therefore that the number of posts in 2006 and 2009 ‘s summer break period has a higher leisure compared with the rest of the months on that particular year on average. This as a result, further supported our initial findings.

## Finding significance of posts at night and posts at day

## We would also like to know whether posts made at night compared to posts made at day by authors have any significance in affecting the attributes. We defined posts that is between 6 a.m. to 6 p.m. to be categorized as day and 6.01 p.m. to 5.59 a.m. to be night. We subset the data to be within 2006 because that is the period where the top 5 threads (excluding the thread 252620 since on that thread, most posts are made on December 2005 which may affect the analysis performed) have the most posts.

## 

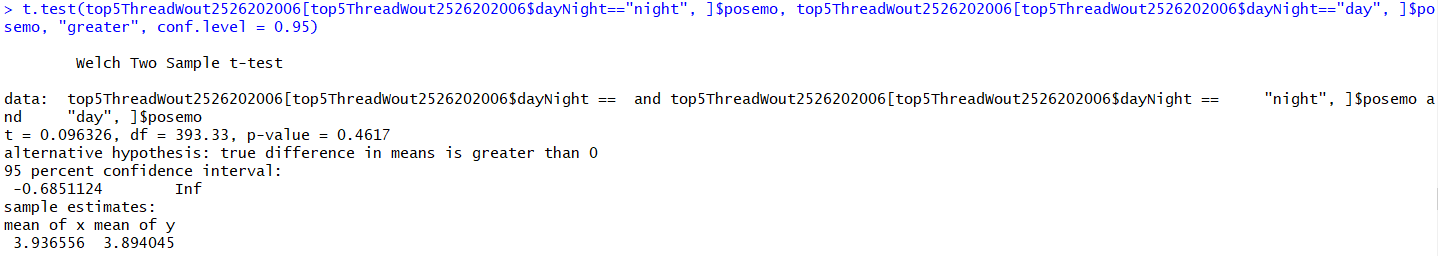
## It can be said that most of the attributes does not differ by much based on the mean of the attributes on day and night.

## As can be seen from the result of the aggregate above, the only significant difference of the attribute that is visible, is the word count. Hence, we performed a t-test to show this and the result can be seen below:

## 

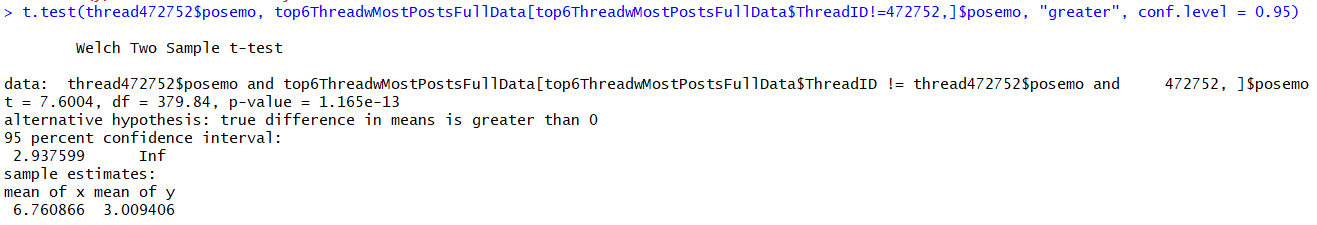
As can be expected from the low p-value and is less than 0.05, at day posts contain more words since people are active during the day and there is a drop in the word count of the posts at night where there is a higher chance that posts contain less words because most people wanted to relax and are exhausted by the activities performed during the day.

Although there is a slight difference between the other attributes such as positive emotion, negative emotion, anger etc. between day and night, the difference is not significant and can be seen from the high p-value from comparing the positive emotion (took as an example) of day and night below:

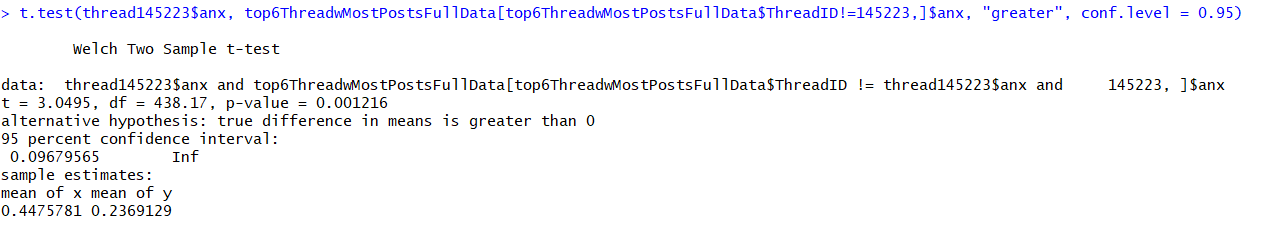


REFERENCES

\*Statistics revision notes FIT1006 slide 83/410

Appendix

*t.test 1: t.test for Posemo in thread 472752 against other threads in top 6 threads*



*t.test 2: t.test for anx in thread 145223 against other threads in top 6 threads*