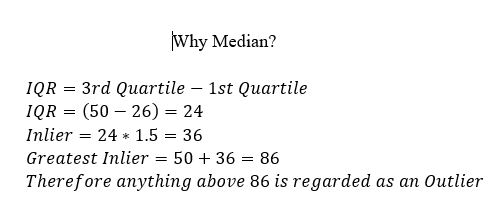
**FIT3152 Assignment 1**

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.

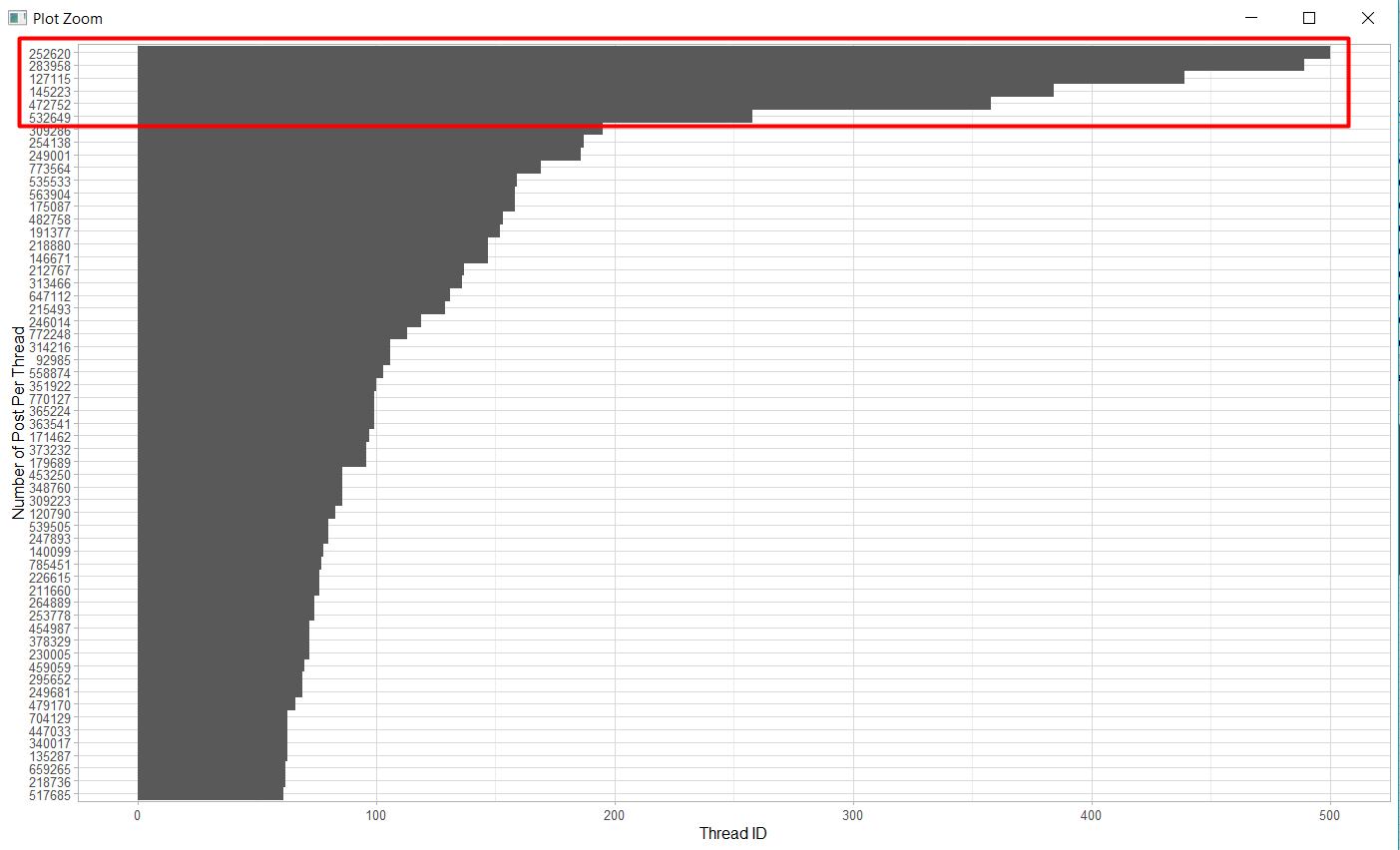
1. **Subset of the data to be analysed**

The first thing that we did is to first analyse the data as a whole and try to find out where can we make the data smaller. We then noticed the column in the data called ‘WC’ is the word count of each post. Some of the post has 0 word count, meaning the post probably contains images or diagrams. So we eliminates these posts that contains 0 word count.

Next we group the posts according to their threadID and then count the number of posts in each thread. In doing this we also take into account the number of authors in each thread (We count the number of authors in each thread and take the average of it. Here we choose median as average because there exist outliers. Then we eliminate the threads with number of authors lower than the median number of authors).

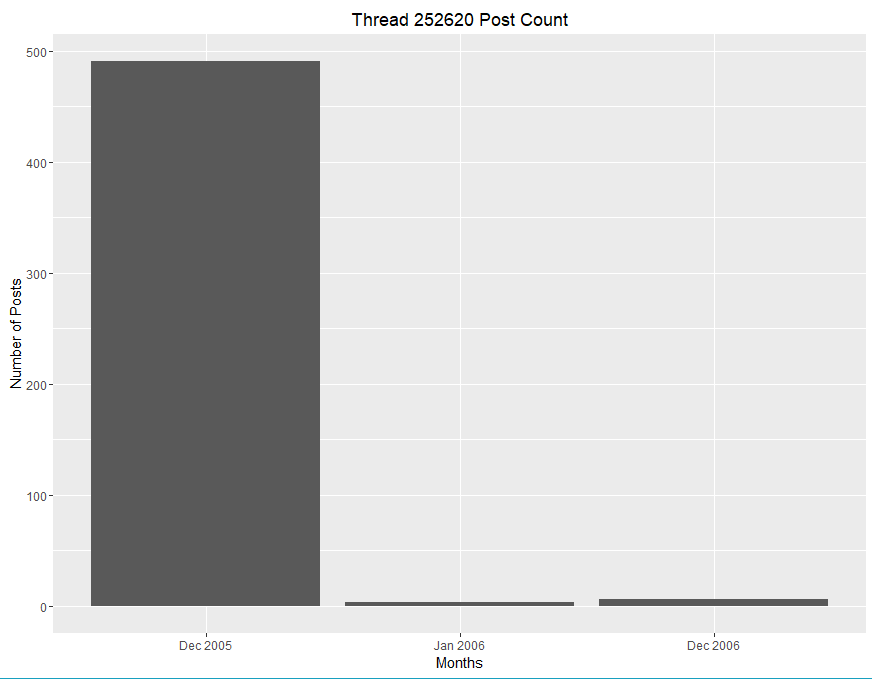


\*Statistics revision notes FIT1006 slide 83/410

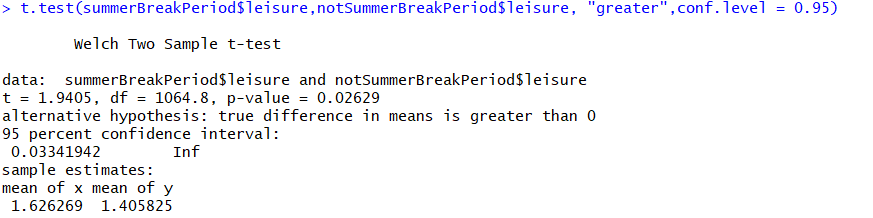
This is to make sure that there are several different authors in each thread and not only 2 people that keeps replying to each other. Next, we plot the number of post for every thread in the forum and we got the following graph:

Which gives us thread 252620, 283958, 127115, 145223, 472752, and 532649 as the top 6 threads with the most post (after thread 532649, the number of post is below 200 and we chose 200 as the threshold). We then take the 6 threads mentioned above and use it for this assignment.

**Analysing Language Change Over the Time**

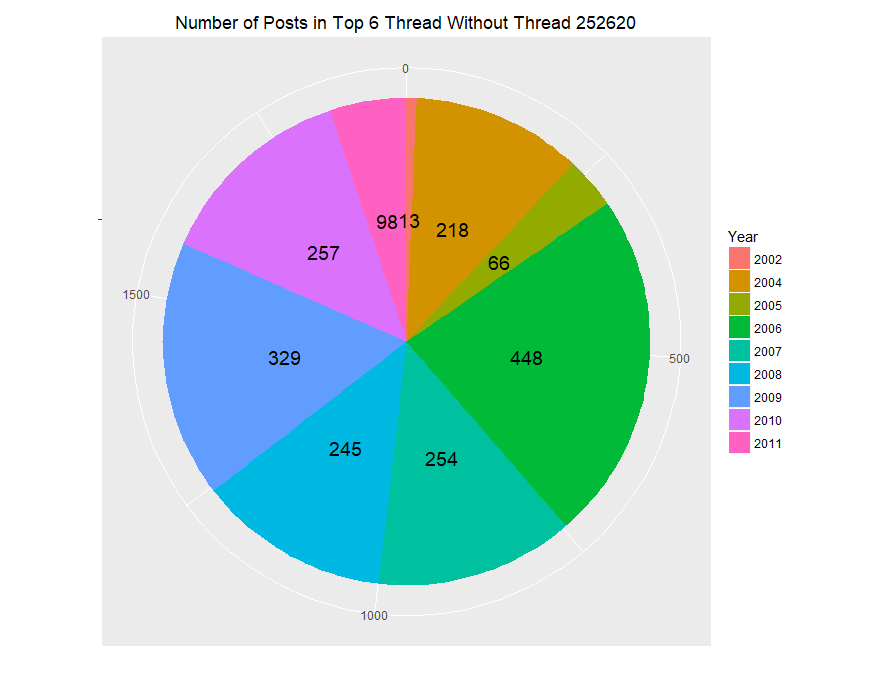
We took the top 6 threads and noticed that thread 252620 only has posts on Dec 2005, Jan 2006 and Dec 2006 where in Dec 2005 it has a significant number of posts as can be seen in the diagram. Therefore, we eliminate the thread 252620 to prevent bias and affect the results and findings from our analysis.

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:

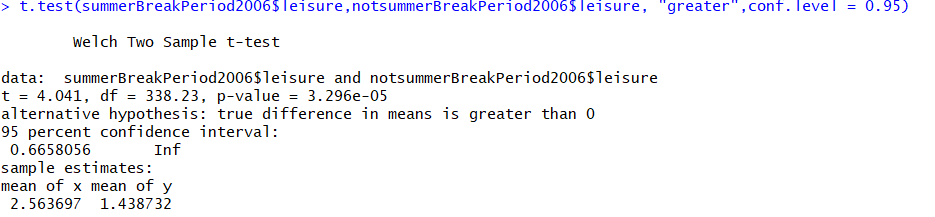


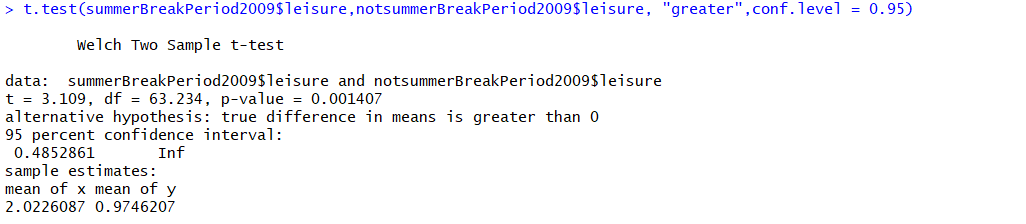
We performed a t-test on 95 percent confidence level and found out that the p-value is lower than the critical value which is 0.05. Consequently, we can reject the null hypothesis and therefore accept the alternative hypothesis where the number of posts about leisure on summer break period is more than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

## Analysing usage of similar language

## Using the top 6 threads that we have subset earlier, we want to analyse the usage of similar language in a thread. We picked analytic attribute of a thread and there is no significance in picking this so it is randomly selecting an attribute. We calculated the mean of the top 6 threads based on the analytics and discovered that the mean of thread 127115 is the highest of the top 6 threads. The graph of the mean of analytics of the top 6 threads can be seen on the side.

## We, then assumed this thread to be the most analytical thread in the top 6 threads that we have subset and picked the year with the most posts which can be seen in the bar graph that in 2009, thread 127115 has the most posts.

## 

## We then further subset the data based on the number of posts of an author and get the author that has the most posts in thread 127115 in 2009. As can be seen in the graph below, author 47875 has the most posts in 2009 and in fact he/she has the most posts in thread 127115 of all time.

## We discovered author 47875 has posted not only in thread 127115 but also in other threads as well. The number of posts of author 47875 in other threads can be seen from the graph below

## 

## After we have aggregated all the data, we found out that thread 472752 has the most positive emotion which can be seen in the table and also we did the t-test to compare the positive emotion in thread 472752 with the rest of the threads in the top 6 threads.

## 

## Then, we did a t-test as well on the posts on thread 472752 of author 47875 and the p-value as can be seen from the t-test below, is smaller than 0.05 which tells us to accept the alternative hypothesis and therefore positive emotion in posts that the author 47875 posted in thread 472752 is higher on average than positive emotions on other threads that the author 47875 posted.

## We also took the thread 145223 as a sample to analyse as well since the author 47875 also posts in this thread. We found out the level of anxiety in thread 145223 is the highest among the top 6 threads which can be seen from the table and t-test is also performed which also gives a small p-value which is less than 0.05.

## 

## Hence, we finally performed a t-test as well for the posts that the author 47875 posted in thread 145223 which compared to the other threads that the author posted, is higher in terms of level of anxiety on average based on the p-value less than 0.05

## We decided not to analyse the thread 532649 which the author 47875 has also posted because he/she had only posted once and we believed that the sample size is too small to deduce an analysis.

## Furthermore, in thread 127115, since it is the most analytical thread among the top 6 threads, we would like to test authors that have posted in this thread, are they following the language (which is analytical) used in this thread. Since there would be a lot of t-tests if we include all the 5 other threads to be compared, we just use a random sampling to give a random thread. Hence, it gives the thread 472752.

## 

## Then, we picked the author that has the most number of posts and he/she has also posted in thread 127115 which is important in proceeding further for the analysis.

## 

## We then compare the analytics of the thread 127115 that the author post with the other remaining threads where the author has posted so it gives the overall behavior/characteristic of the author in posting in a thread. We use t-test again to test the significance of the analytics on average and we found the p-value is less than the critical value (0.05) and hence we accept the alternative hypothesis and therefore supported our analysis that people in a certain thread tends to use similar language of that thread.

## 

## 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 of time 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:

