# Project report

For this project, I plan to do a research on whether there is a relation between Reddit comments and the hour when those comments are created. To be more specific, I analysis comments under the ‘xkcd’ subreddit. The reason why I made this filter is because, from the given set of data, this one is a significant label that can distinguish different comments. Also, the ‘xkcd’ is an interesting web comic mixed with romance, sarcasm, math, and language that people who are interested in it have some similarity in personality and there should be some sentiment kinds in comments regarding these comics. Thus, I made a filter to reduce other types of subreddit comments. In terms of the comment part, what I thought previously is that I can do a sentiment analysis on the comments to figure out the polarity of them. Base on the sentiment analysis result, an idea hits me that there might be a relation between the polarity of comments and the created date(hour) of comments since I had recently guessed that people tend to be in the positive emotion status in the day and negative emotion after work, which is usually after work. To make it clear, this is the origin of my project.

To refine data, the first step is to extract the data from the Reddit comments corpus by filtering subreddit as ‘xkcd’. Then I removed the instance whose author is empty. (However, when I was writing reports, I realized that it should not remove whatever the author is empty or not because the record existed and we didn’t care much about who created) Then, I removed the instance, which has no comment body part, because this is highly related to my result. A convenient case for me to refine data is that this data set was already gathered by other people online, so the content of data and score didn’t need to be filtered since they are already cleaned as they were extracted from Reddit.

At this stage, I selected features that related to my topic, author, comment_body, and created_at time. The next major step is to clean the comment_body to prepare for the further sentiment analysis. Firstly, I scanned the data and found some kinds of low_valued data. The URL, numbers, and punctuation should be removed. Secondly, I made all letters be lowercased so that could avoid sentiment analysis on same meaning words repeatedly. There are other refined parts like converting the timestamp to the real datetime so that I can extract the hour.

Now, the comment_body is cleaned. I used sentimentIntensityAnalyzer (from NLTK) to do sentiment analysis. It outputted a dictionary with scores of the compound, positive, negative and neutral. I filtered data by ignoring the compound and neutral because my topic only focuses on the pos/neg comment. Then, I labeled a polarity for each comment by taking the max score between pos and neg and choose the polarity of the max score as the polarity of the comment. Later, I grouped data by hour and polarity and counted the number of comments. Finally, I filter the hour by treating 7:00 to 18:00 as day time and 19:00 to 6:00 as night time.

Chi_square is the analysis method I used, I grouped the data by polarity and took the sum of the number of comment. Then, apply the test method on the data.

The p-value is far from the 0.05, which means there are no relationships between the polarity of comments and whether the comments are created in the morning or night.

While there are many things can be refined later. The reason of setting morning and night label is not firm, especially when I paid attention to the number of comments. For example, people tend to be optimistic when they wake up, which influences what they will comment in Reddit. Moreover, when 16:00, people might be tired of work or study, they might have the bad attitude to Reddit and the content in Reddit might be a source to pull all their bad emotion status out. So, I shouldn’t label morning and night and the 24 hours should be a better choice to find inner relations. Even if I used morning and night, a reasonable explanation should be mentioned. In addition, there should be more refined operations in data cleaning since I still noticed some low_valued word, like “awwwwwww”.

[I didn’t do much in analysis part and I planned to draw the trend of the number of the positive /negative comments and a stacked bar chart to saw the distinguish of comments, as well as the polarity proportion of them. I will analyze the relation between the hour and the strength of the polarity of comments to move deep in his topic.