Captain: Aishani Pal - aishani2
Alyxandra Merritt - merritt9
October 24, 2021

CS 410 Course Project Proposal

Our main goal is to analyze the emotions of GitHub commit comments associated with a person over time for a single project.

We will do so by analyzing multiple people across multiple projects. Our idea is inspired by the paper *Sentiment analysis of commit comments in GitHub: An empirical study* written by Emitza Guzman, David Azócar, and Yang Li. The study looks at sentiment for commit comments for the overall projects and analyzes several metrics including time and project approval. One Github commit comment feature the paper does not analyze, but we believe could be interesting, is to track comment sentiment for a singular user. Looking at this trend can tell us how a developer's mood changes over time as a deadline approaches. For example, we can answer questions like: *does a developer become more negative closer to a deadline due to frustration? Or is the developer likely to become more positive as a result of better progress over time?*

We plan to take a very systematic approach to complete this project. First, we will gather our data using the GHTorrent dataset mentioned in the paper. Then, we will begin developing our Sentiment Analysis tool. Finally, we will create graphs and metrics displaying our sentiment of users change over time. This will allow us to draw conclusions in a manner similar to the paper.

We will use the same GHTorrent dataset as mentioned in the paper (https://ghtorrent.org/msr14.html) and import it into MySql. Sentiment analysis and opinion mining will be discussed in the course in Week 11. We plan to leverage this lecture material with the approaches and tools mentioned to create our own sentiment analysis. Some packages we have previous knowledge of that might be helpful for this project are pandas to help work with the data, nltk for sentiment analysis, matplotlib for plotting data, and scipy for analysis.

With the completion of the project, we expect to see either a positive, negative, or neutral trend in emotions across time. We will conduct both an observational and mathematical analysis to draw conclusions based on the sentiment analysis data we derive.

We will evaluate our work by ensuring that our sentiment analysis tool produces similar results as the paper. That is, we will match our results of emotions with the programming language, time of day, and day of the week with those of the paper. Once we can reproduce the paper's results, we will continue with our research question regarding emotions over time for a single person.

We plan to use Python to complete this project. Also, we will use MySql Database queries on the dataset.

For our project team of two members the workload needs to be at least 40 hours and we have broken it down in the following table:

Task	**Estimated Time**
Data collection	2 hours
Build sentiment analysis tool	15 hours
Test our sentiment analysis with similar metrics as the paper	8 hours
Run our sentiment analysis on our metric	12 hours
Evaluate trend (observational and statistical analyses)	3 hours