Captain: Aishani Pal - aishani2
Alyxandra Merritt - merritt9
November 15, 2021

Progress Report

Progress Made

We have completed the data collection portion of our project. Using the data from GHTorrent, the SQL database is saved locally. We created a script to query the data and save the information we need in organized .csv files. The relevant data we have saved are the project IDs, user IDs, and comment text. If we need any additional pieces of data later on, updating the script should not take long.

We have also begun creating our own sentiment analysis tool. We are using the nltk package. So far, we have done tokenization and POS tagging on our comments.

Remaining Tasks

First, we need to finish creating our own sentiment analysis tool. Then, we will use the tool to replicate the paper's findings. Once we have reasonable results, we will continue with our own research goal: *to analyze the emotions of GitHub commit comments associated with a person over time for a single project.* Lastly, we will chart all of our findings.

Challenges

The dataset was quite large and initially created problems when reading the data we needed. However, after studying the SQL schema, we were able to extract the specific information we needed for our project.

At first, we wanted to use SentiStrength, the tool used by the paper we are following, to accurately compare our findings with those of the paper. However, we found out that we have to pay 1000 euros in order to use the software. Thus, we have decided to not use SentiStrength to compare. Instead, we will use our own sentiment analysis tool to attempt to replicate the findings of the paper before proceeding with our own research question.

Since we shifted the project slightly to focusing on our own sentiment analysis tool, we both planned to work on this task and agreed to update each other on progress that was made. Aishani was able to complete tokenization and POS tagging on the comments on her machine successfully. When Alyxandra attempted to build upon this work, she experienced challenges running this script on her own machine. Despite spending several hours attempting to resolve the error, running the code on multiple different environments and machines, and seeking help from Aishani, it is still unclear why Alyx is unable to run the existing code. This will be one of the first things that we plan to resolve as it is necessary that both of us are able to run and contribute code from our own machines.