CS 5306: Project 1 Proposal

Natasha Armbrust, nka8
Fatima AlGhamdi, faa52

Analysis of Github Contribution in Connection to Project Sentiment

Crowdsourcing Platform: Github

We plan to analyze the crowdsourcing platform Github, a web-based version control repository that allows easy project collaboration and is widely used for open-source software projects. As computer scientists, software collaboration is an important and crucial part of software development. We are particularly curious in determining factors behind the amount of contributions to a project and how these factors can influence the progress of contributions over time. The factor we are targeting for our CS 5306 project is project sentiment. From class and our personal experiences, we have learned that successful groups have high turn-taking and social intelligence [1]. We believe tone and sentiment is very important for group collaboration. Thus, we wanted to turn to Github, an online group collaboration setting, to analyze sentiment in a virtual crowdsourcing platform.

We plan to answer this question by doing sentiment analysis on pull request comments. We will be using the dataset MSR 2014 Mining Challenge Dataset. This dataset provides data from the top-10 starred software projects for the top programming languages on Github, which gives 90 projects total. The dataset includes projects, users, pull requests, pull request comments, and pull request history and can be downloaded and accessed through MySQL. We will use Natural Language ToolKit for sentiment analysis on pull request comments to get the percentage of positive, negative, and neutrality in the comment. Project contribution amount will be assessed over time via pull request creations. Pull request comment sentiment will be assessed over time via the time the comment was created. Our final report will include our analysis driven by data visualizations.