Sentiment Analysis of Software Development Teams

Mamtaj Akter
makter2@uoregon.edu
November 2019

Large software development often requires an analysis of the developers' productivity and how the developer/user emails can be manipulated to analyze that. Also, to track the sentiment of a developer team or of an individual developer with respect to time is crucial to analyze developers' emotions. These issues have always been a great challenge for the software engineering researchers who analyze to quantify developers' effort towards software developments.

In this project, the programming language that I will use is Python 3. The main goal of this project is to analyze the developer/user emails by identifying an email's topic. In other words, answering the question: Is the email about a bug or a software feature?. Therefore, to perform the topic labeling, I will use Latent Dirichlet Allocation (LDA) model. Also, I will study the emails and the Github issues of different software to perform the sentiment analysis - how developers' sentiment varies over time or how one developer's sentiment affects the whole team. TextBlob is a python library that has Naïve Bayes as a classification feature that can be used to process this sort of textual data to perform sentiment analysis.

The data that is required to study the topic labeling and sentiment analysis of software has already been collected. In [1], Shweta Gupta has mentioned how difficult the data collection process was. She collected the publicly available mailing list of PETSc and Linux using web scraping. The average number of emails in PETSc project was 20,000. The major challenge was collecting Linux email data which had a size around 6.5 GB. For the Github issues she cloned SPACK repositories' issue trackers and the size of this data was around 20 MB.

Reference:

[1]: Gupta S., "Software Development Productivity Metrics, Measurements and Implications", master's thesis, University of Oregon, June 2018.

*https://www.cs.uoregon.edu/Reports/MS-201806-Gupta.pdf