Group 4, Presentation 1

By: Bill Yu, Kevin Amado, Leo Wei

Selected Paper

Information and Software Technology

Volume 139, November 2021, 106665

ELSEVIER

"*Won't We Fix this Issue?*" Qualitative characterization and automated identification of wontfix issues on GitHub

Sebastiano Panichella [a], Gerardo Canfora [b], Andrea Di Sorbo [b]

Show more

+ Add to Mendeley Share Cite

https://doi.org/10.1016/j.infsof.2021.106665

https://doi.org/10.1016/j.infsof.2021.106665

Problem Overview

We want to **predict whether an issue on a specific GitHub repository will be fixed** by analysing the issue title, description and discussion comments.

This is important because according to the paper, fixed and not fixed issues receive the same amount of attention and effort from the project developers and managers, and therefore an approach for timely identifying the issues that won't be fixed reduces the unproductive effort required to triage and resolve such issues. Further, the longer a "won't-fix" issue remains open, the more chances it will catch other people attention, creating a spiral of wasted time.

Data Collection

We will collect the data by using the GitHub REST API:

- https://docs.github.com/en/rest/issues/issues#list-repository-issues
- https://docs.github.com/en/rest/issues/comments#list-issue-comments

And a Python script that produces JSON documents from that information so that it can be loaded from-disk later and used in the analysis.

We'll pick only a project using the methodology mentioned in the paper:

- Take the top 1000 repositories ordered by number of stars
- Pick a C# project from that list that has a high number of issues, good issue labels, and a healthy community

Research Questions

By analyzing the data we want to visualize the statistics of the factors that relate with the resolution time of a won't fix issue, vs a will-fix issue:

- **Description Length**: Issue description length (number of characters);
- **Max Author Percentage**: The proportion of messages posted by the author who posted the majority of messages in the issue discussion;
- **Major Authors**: Number of unique authors who have posted more than one-third of the overall messages present in the issue discussion;
- **Mean Comment Size**: Average length of comments (number of characters) in the issue discussion;
- **Minor Authors**: Number of unique authors who have posted less than one-third of the overall messages present in the issue discussion;
- **Number of Actors**: Number of distinct authors participating in the issue discussion;
- **Number of Comments:** Number of total comments in the issue discussion;
- **Time To Close Issue**: Time lapse (in days) between issue opening and closing (with the wontfix label);
- **Time To Discuss Issue**: Time lapse (in days) between issue opening and last comment posted in the issue discussion.

Research Questions

By using ML we want to predict whether an issue is a “won’t fix”, or a “will fix”

Project Plan - Labelling

Each member will label 1000 issues with either “will-fix” or “won’t fix”.

For each record, the resulting label will be the one that is most agreed on, or, if no consensus is reached, it will be discussed together.

It is worth highlighting that for predicting whether an issue will be labeled as “won’t fix”, the machine learning models will be trained by exclusively using information that is immediately available at the issue opening (i.e., issue title and description, without considering the other features). However, when labelling, all the information available will be considered.

Project Plan - Preprocessing

We'll take the data collected and stored as JSON documents and perform:

- Markdown rendering and extraction of the text
- Stop-word removal using the English Standard Stop-word list
- Stemming (English Snowball Stemmer)
- Weight the stems using tf-idf

Producing a matrix with the issues as rows, and the all the keywords as columns, where the value of a column and row is the weight of the keyword if it is present in the title or description, or 0 if it is not present.