World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:15, No:02, 2021

Open Science Index, Computer and Information Engineering Vol:15, No:02, 2021 publications.waset.org/abstracts/133701.pdf

Semantic Differences between Bug Labeling of Different Repositories via Machine Learning

Authors : Pooja Khanal, Huaming Zhang

Abstract : Labeling of issues/bugs, also known as bug classification, plays a vital role in software engineering. Some known labels/classes of bugs are 'User Interface', 'Security', and 'API'. Most of the time, when a reporter reports a bug, they try to assign some predefined label to it. Those issues are reported for a project, and each project is a repository in GitHub/GitLab, which contains multiple issues. There are many software project repositories -ranging from individual projects to commercial projects. The labels assigned for different repositories may be dependent on various factors like human instinct, generalization of labels, label assignment policy followed by the reporter, etc. While the reporter of the issue may instinctively give that issue a label, another person reporting the same issue may label it differently. This way, it is not known mathematically if a label in one repository is similar or different to the label in another repository. Hence, the primary goal of this research is to find the semantic differences between bug labeling of different repositories via machine learning. Independent optimal classifiers for individual repositories are built first using the text features from the reported issues. The optimal classifiers may include a combination of multiple classifiers stacked together. Then, those classifiers are used to cross-test other repositories which leads the result to be deduced mathematically. The produce of this ongoing research includes a formalized open-source GitHub issues database that is used to deduce the similarity of the labels pertaining to the different repositories.

Keywords : bug classification, bug labels, GitHub issues, semantic differences

Conference Title : ICDAMLKD 2021 : International Conference on Data Analytics, Machine Learning and Knowledge Discovery

Conference Location : London, United Kingdom

Conference Dates : February 15-16, 2021