2k18 – csm -11 Ali ahsan

**Log Chunks: A Data Set for Build Log Analysis**

**AUTHORS:**

CarolinE . Brandt, Anni bale Panichella, Andy Zaidman, Moritz Beller

DATE: MSR ’20, October 5–6, 2020, Seoul, Republic of Korea

**ABSTRACT:**

Build logs are textual by-products that a software build process creates, often as part of its Continuous Integration (CI) pipeline. Build logs are paramount source of information for developers when debugging into and understanding a build developers who caused the original build failure. The width and depth of the Log Chunks data set are intended to make it the default benchmark for automated build log analysis techniques.failure. Recently, attempt stop partly automate this time-consuming, purely manual activity have come up, such as rule-or information-retrieval-based techniques. Log Chunks contain seminally labeled log part (chunk) describing why the build failed. We externally validated the data set with the chunk.

1 **INTRODUCTION :**

Continuous Integration (CI) has become a common practice in softwareengineering[10]. Manys of software projects use CI[2,10,17] to detect bug searly [8,18], improve developer productivity[10,13] and communication[7].CI builds produce logs which are port results of various sub-steps within the build. For Log Chunks, we queried GH Torrent from 2018-04-01 for the three most popular repositories of each of the 30 most popular languages to cover abroad range of development languages. Among there sulting repositorie sare, for example, Microsoft/Type Script, git/git and jwilm/alacrity

Log Sampling. Travis CI builds comprise number of jobs that actualize a build process in different environments. Hence, the out come from different jobs might be different. For each building Log Chunks , we download the logo fthe first job that has the same state as the overall build. We inspected the collected build log sanddis carded logs from three repositories. One had only one failed build, two others had empty build logon Travis CI. In total, we collected 797 logs from 80 repositories spanning 29 languages.

**importance**

In this paper, we introduce Log Chunks, across-validate dataset encompassing 797 build logs from 80 projects using Travis CI. For each log, we annotated the log chunk describing why the build fail eland provided keyword sad eveloper would use to search for the log chunks well as categorization of the log chunks according to the informant within the log. Log Chunks advances the research fie l do build log analysis by introducing benchmark rigorously examine research contributions[15]and opening various research possibilities that previously required tedious manual classification.