EAWOC: Exploratory Analysis of Commit Messages in World Of Code

Addi Malviya Thakur, Mahmoud Jahanshahi, Rekha Bhupatiraju, Minoo Oliaee, Prachi Patel

I. OBJECTIVE

World of code (WoC) has enabled research on the global properties of Free and open-source software (FOSS)[2]. The objective of this project would be to conduct an exploratory analysis on the $\sim$1.5 Billion commit messages present in WoC. The exploratory analysis will involve the following:

1) Perform *descriptive statistics* to understand the basic features of the commit messages in WoC[1]. In particular, we will gather fundamental insights of WoC by calculating the *measures of central tendencies* and the *dispersion* of the commit messages in the WoC.
2) *Term frequency extraction* and analysis to discover the most common words used in the commit messages
3) Plot the *temporal distribution* of first and last 10,000 commit messages of random 1000 projects with at least 50,000 commits. Also, perform *distribution fitting* to finding the best distribution that fits the underlying commit messages at each repository level.
4) Hypothesis testing for the presence or absence of *Pareto Principle* for the number of commits made by the software developers. Specifically, we will study if 80% of commit message come from 20% of committers for a random sample of 100,000 repositories (at each repository level).
5) Develop *static and interactive visualizations* to demonstrate the findings.

II. MOTIVATION

Since its inception Free/Libre and open-source software (FLOSS) has made tremendous impact in the software community and its ecosystem. FLOSS is software that is both free software and open-source software where anyone is freely licensed to use, copy, study, and change the software in any way, and the source code is openly shared so that people are encouraged to voluntarily improve the design of the software[3]. GitHub is a popular place to host open-source projects and is the largest source code repository in the world. It offers a socio-collaborative working environment for the open-source projects, where individuals can participate to develop and release software. Adding comments during pull and push updates to the software code is popular way to communicate the changes and interact with the community. Comments informs other developers of bug reports, planned improvements, changes, and feedback, among other activities. WoC allows an easy and programmatic access to such comments for research and analysis[2].

An exploratory analysis of comments would allow insight in the development and progress of software, its updates, and contemporary interest and usage. Furthermore, such an analysis would usher novel research methodologies to understand team dynamics, software evolution cycle, identification of common pitfalls - what works and what doesn't; eventually providing a holistic approach towards improving software development and engineering processes.

III. DATASET

The WoC collection of software repository contains cross-reference authors, projects, commits, blobs, dependencies, and history of the FLOSS ecosystems. This data is updated on a regular basis and contains billions of git objects. For the purpose of this study we will use WoC's large and frequently updated collection of version control commit data of the entire FLOSS ecosystems.

WoC API's will be used to fetch and process the data related to comments. We plan to extract comments from repositories with at least 100,000 commits made in a period of time (say, last 5 years). We will also be collecting meta data related to these commit messages including:

1) commit info: timestamps, number of associated blobs, number of associated files
2) user info: number of projects, number of commits
3) repository info: number of authors, blobs, time since creation

The exact extent to which we plan to dive deep into each of these three areas will depend on the results obtained at each stage of our project.

IV. ROLES AND RESPONSIBILITIES

We take a team driven approach to address the objectives of this project. The roles and responsibilities allows for rapid turnout and will benefit from individuals specialized skill-sets and interest. To this end, following are initial roles and responsibilities.

- Addi Malviya Thakur: will lead the data analysis for the project and will assist in data preparation and visualization. She will primarily work on objectives #1, #2, #3, #4.
- Mahmoud Jahanshahi: will lead the data collection for exploratory analysis and will serve as an interface to access the WoC data through queries and scripting. He will also contribute to Objective #1, #3, and #4.
- Rekha Bhupatiraju: will lead the visualization and plotting of results and findings for the project. She will assist on objectives #1 - #5 and primarily on #5.
- Minoo Oliaee: will primarily contribute to the data preparation for exploratory analysi as well as the data collection and curation of the WoC data through queries

and scripting. She will assist in Objective #1, #2, and #4.

- Prachi Patel: will primarily contribute to the visualization and plotting of results and findings for the project. She will assist on objectives #1 - #5 and closely team up on #5.

V. Time-line of Milestones

In Table-I, we have shown the tentative timeline for our project leading up the submission of final report during the exam week. The listed task corresponds to the objectives proposed earlier in the project proposal and the timeline is tracked at a weekly scale until the end of November.

TABLE I
Project timeline

Task/Timeline	9/27/2021	10/4/2021	10/11/2021	10/18/2021	10/25/2021	11/1/2021	11/8/2021	11/15/2021	11/22/2021	11/29/2021	12/6/2021	Exam Week
Data Collection and Curation	🟩	🟩										
Descriptive Statistics	🟩	🟩	🟩	🟩								
Term Frequency Analysis			🟩	🟩	🟩	🟩						
Temporal Distribution Analysis					🟩	🟩	🟩	🟩				
Distribution Fitting						🟩	🟩	🟩				
Hypothesis Testing for Pareto Principle								🟩	🟩			
Interactive Visualizations		🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩		
Report Writing				🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩

We will begin with the data collection and curation from WoC repository. This process will involve extraction, preprocessing, cleaning, and converting in a format for exploratory data analysis. Several of tasking as shown will be executed in parallel for efficiency purposes. Efforts that includes visualization and reporting writing will be done on a regular basis to include the updates towards final deliverable.

VI. Expected Outcomes

This project will result in the first-ever large-scale analysis of the WoC code repository of more than ∼1.5 Billion commit messages. The temporal distribution analysis will be an essential metric for project managers and scrum masters to understand what efforts take time and how to optimize the project planning for the future. The distribution fitting of commit messages will result in the overall progression of software development. It will enable a longer-term decision on release management and deployment deadlines in a real-world setting. Finally, we plan to host and release the finding through an interactive visualization portal that would allow other researchers and the scientific community to benefit from their work.

References

[1] Zealure Holcomb. *Fundamentals of descriptive statistics*. Routledge, 2016.

[2] Yuxing Ma, Tapajit Dey, Chris Bogart, Sadika Amreen, Marat Valiev, Adam Tutko, David Kennard, Russell Zaretzki, and Audris Mockus. World of code: Enabling a research workflow for mining and analyzing the universe of open source vcs data, 2020.

[3] Walt Scacchi. Free/open source software development. In *Proceedings of the the 6th joint meeting of the European software engineering conference and the ACM SIGSOFT symposium on The foundations of software engineering*, pages 459–468, 2007.