COMP0104 Coursework Ethics Review Request for Group A
ANONYMOUS AUTHORS

In Test-Driven-Development, tests are written before the tested code. If a project adopted TDD, the git repository should reflect this. If a new class (file) is created, then the same or an earlier commit should also create a new test class (file).

1 ANALYSIS DESCRIPTION

This proposal aims to explore the default coursework question provided by the brief as shown above. The analysis be breaking down this question into the following set of research questions:

RQ:

1. How often is a test class (file) created (a) before or (b) in the same commit as a tested class (file)?
2. Are there test cases created after committing the code, if so, how often did this occur?
3. How does the size of a commit impact the results?
4. How can you link a test class (file) to a tested class (file)?
5. When a new class is created, do earlier commits always create a new test class?

The research questions will be answered using data from the Apache project code repository [1]. Firstly, the data will be filtered to only include repositories relevant to our analysis, i.e. those explicitly stated to have used TDD during development. PyDriller will then be used to mine these chosen repositories for the required datasets (described below). Finally, key statistics will be drawn in order to answer the aforementioned chosen research questions.

2 ETHICS IMPLICATIONS

The Apache Project is published under Apache License 2.0, which permits modification and distribution of the project. There is no objection to large-scale repository mining in terms and conditions.

Although the proposed analysis does not include the contributor behind a commit as a data point, the resultant dataset is likely to contain unique identifiers of relevant commits, possibly allowing the contributors to be traced by extraneous individuals through the commit metadata (can be obtained via `git log` or `git show <commit hash>`).

To protect the contributors in the repository, the committers' ids will not be included in the research dataset which will be published. Our program will not try to access the committers' ids while mining repositories as they are not helpful to our research questions.

3 DATA ELEMENTS

1. A data set containing the time of each commit that initially creates test files. It may also include the commit id, the content of the commit e.g. the files' name and the size of the commit.

This data set is used to track when test files are created and what are the contents of the test file commit.

2. A data set containing the time of each commit containing the tested code. It may also include the commit id, the content of the commit e.g. The files' name and the size of the commit.

   This data set is used to track when the tested code is added to the repository and.

3. A data set containing the build logs, including the result of the testings and the corresponding commit (including commit id).

4 REFERENCES

[1] "Source Code Repositories at Apache," *infra.apache.org*. https://infra.apache.org/version-control.html (accessed Dec. 05, 2021).