Lab Assignment 3

Course: CS202 Software Tools and Techniques for CSE
Lab Topic: Multi-Metric Bug Context Analysis and Agreement Detection in Bug-Fix Commits
Date: 18th August 2025

Objective

This lab builds upon your Lab 2 file-level bug-fix commit dataset by incorporating additional code quality metrics into the analysis. The focus is on structural metrics for measuring Maintainability Index, Cyclomatic Complexity, and Lines of Code for each bug-fix before and after the change using **radon**. The metrics are then compared with Semantic and Token-Based Similarity measures to investigate the relationship between the code quality and the magnitude of changes.

Learning Outcomes

By the end of this lab, students will be able to:

- ✓ Use **radon** to calculate Maintainability Index, Cyclomatic Complexity, and Lines of Code.
- ✓ Compare structural metrics with change magnitude metrics (semantic & token similarity).
- ✓ Classify bug fixes as *major* or *minor* using thresholds.
- ✓ Identify and analyze cases where metrics give conflicting assessments.

Pre-Lab Requirements

- Lab 2 CSV dataset (rectified commit message dataset)
- Any Operating System (Windows, Linux, MacOS, etc.)
- Python 3.10 or later
- **radon** – https://pypi.org/project/radon
- CodeBERT (microsoft/codebert-base) https://huggingface.co/microsoft/codebert-base
- SacreBLEU https://github.com/mjpost/sacrebleu (Chosen for standardized, reproducible BLEU scoring widely used in research).

Lab Activities

(a) Start from lab 2 dataset

- ❑ Use your Lab 2 file-level dataset (produced in the "**Diff Extraction and Analyses**" step of Lab 2) as the starting point for this lab. While CSV format is recommended for consistency, a slightly different storage format can also be used, provided it contains at least the following information as the bare minimum:

Hash	Message	Filename	Source Code (before)	Source Code (current)	Diff	LLM Inference (fix type)	Rectified Message
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...

(b) Compute and report baseline descriptive statistics:

- ❑ Total number of commits and files.
- ❑ Average number of modified files per commit.
- ❑ Distribution of fix types from LLM Inference (fix type).
- ❑ Most frequently modified filenames/extensions (e.g., `.py`, `.java`).

(c) Structural Metrics with radon:

For each file in your dataset:

- ❑ Run **radon** on `Source Code (before)` and `Source Code (current)`.
- ❑ Extract:
  Maintainability Index (`MI`), Cyclomatic Complexity (`CC`), Lines of Code (`LOC`)
- ❑ Compute:
  `MI_Before, MI_After, CC_Before, CC_After, LOC_Before, LOC_After`
- ❑ Add columns:
  `MI_Change = MI_After – MI_Before`
  `CC_Change = CC_After – CC_Before`
  `LOC_Change = LOC_After - LOC_Before`

(d) Change Magnitude Metrics:

- ❑ Compute Semantic Similarity between `Source Code (before)` and `Source Code (current)` using CodeBERT.
- ❑ Compute Token Similarity using BLEU.
- ❑ Add `Semantic_Similarity` and `Token_Similarity` columns.

(e) Classification & Agreement:

- ❑ Define thresholds to classify a commit as *Major* or *Minor* based on: Semantic Similarity, Token Similarity, individually.
  Example: (These are just a suggested threshold, they can be adjusted)
  - Semantic Similarity >= 0.80 -> Minor Fix
  - Semantic Similarity < 0.80 -> Major Fix
  - Token Similarity >= 0.75 -> Minor Fix
  - Token Similarity < 0.75 -> Major Fix

Applying these thresholds to each commit, will get two separate classifications (one from Semantic Similarity, one from Token Similarity).

Note: Please reach out to the TAs for any queries/issues.

- ❑ Add columns `Semantic_class` and `Token_class`, which will comprise of the "***Major***" and "***Minor***" labels with respect to each commit.
- ❑ Create `Classes_Agree` column with values: "***YES***" if both `Semantic_class` and `Token_class` have the same value, otherwise "***NO***".

(f) Final Table Structure *(in continuation with the previous one)*:

..	MI_Change	CC_Change	LOC_Change	Semantic_Similarity	Token_Similarity	Semantic_Class	Token_Class	Classes_Agree
..	…	…	…	…	…	…	…	…
..	…	…	…	…	…	…	…	…

Resources

- Lecture 3 slides

Note: Please reach out to the TAs for any queries/issues.