Bachelor Thesis Topic

A Study on Fault Localization Capabilities of different Language Models

Motivation and Background

Localizing bugs in a software system is generally a very time-consuming task. Recently, fault localization techniques have been developed that utilize statistical language models, which were originally intended to be used with natural languages instead of programming languages. The exact approaches to building these language models differ, though, which brings up the question of what techniques to employ best when building language models that are intended to be used by fault localization methods.

Goals

The goal of this thesis is to identify important properties that language models should possess when used in the context of fault localization. This will be done by examining the influence of changing different parameters and by identifying aspects that are important, but also by identifying aspects that can safely be neglected without affecting fault localization results.

Description of the Task

The specific tasks are:

- Develop an understanding of the principles behind statistical language models and their application in software fault localization.
- Conduct a range of experiments, executing fault localization on an existing benchmark of buggy Java projects. Thereby, examining language models with different sizes, orders and generation methods, including the use/invention of different tokenization strategies as a preprocessing step.
- Perform an experimental evaluation and comparison of the experiments' results.

Research Type

Theoretical Aspects:	★★☆☆☆
Industrial Relevance:	★☆☆☆☆
Implementation	★★★★☆

Prerequisite

The student should be enrolled in the bachelor/master of software engineering/informatics program, and has completed the required course modules to start a bachelor/master thesis.

Skills required

Programming skills in Java. Understanding of, or willingness to learn, the software engineering methods and the statistical techniques needed for the project.

References

[1] Ray, B. et al. (2016). On the "Naturalness" of Buggy Code. Proceedings of the 38th International Conference on Software Engineering, 428–439.

Contacts

Simon Heiden, Humboldt-Universität zu Berlin, Institut für Informatik, Lehrstuhl Software Engineering, Unter den Linden 6, 10099 Berlin, Germany

Application

Please contact me during my office hours or send me an email with the title: "[ThesisProject]-FLCLM" to se-career@informatik.hu-berlin.de