Reflection Report

Huaifeng Zhang

1. What did you learn, in your individual session, about static analysis for ML and the pynblint tool?

During my individual session with Luigi, we used pynblint to analyze some notebooks I have. Static analysis is helpful to improve the maintainability and understandability of code. Maintainability and understandability are vital for large projects which need to be developed collaboratively. pynblint detected some issues existing in my notebooks by static analysis. The detected results fall into 2 categories: repository-level results and notebook-level results. The detected repository-level results include repository-not-versioned and dependencies-unmanaged; and the detected notebook-level results include too-few-MD-cells, cell-too-long and so on.

2. Will pynblint be useful to you in your WASP PhD project? Why or why not?

pynblint would be useful to my WASP PhD project. Because in my project, I need to use notebooks to analyze relevant data. pynblint could detect issues and helps me improve maintainability and understandability of my notebook.

3. Ideas for how the tool could be improved?

It would be great if pynblint could report code coverages of notebooks.

4. What do you see as the limits for static analysis tools in ML? For code, models, and for data?

Most codes in ML projects are related to data manipulation tasks. Data manipulation tasks usually can be performed parallelly. But limited by the programming skills of developers, the code implementations to do these tasks are usually in a sequential way, which is less efficient. As far as I know, there are no static analysis tools detecting these codes.

ML models, especially deep learning models, consist of many matrix operations. This is quite different from traditional programs, which consist of many if-else statements. Because of this underlying difference, static analysis for traditional programs might need to do some modification.

Data is the foundation of ML. Metadata, which provides information of this data is very important for developers. But usually, codes of ML projects include little information about metadata. Developers have to check metadata in separate files. Because static analysis aims to analyze code, it can obtain little information of data.