Ben Rombaut
CISC873
Milestone 1 - Summary
Project Title: An Empirical Study on the Evolution of Machine Learning Library Uses
GitHub Repo: https://github.com/brombaut/CISC873_TermProject

Since our last discussion on selecting a project topic, I have been talking with Professor Foutse Khohm about different research areas pertaining to Machine Learning and Software Engineering. We discussed examining how APIs in popular machine learning libraries, such as TensorFlow and PyTorch, evolve over time. We determined that this area could be the topic of two research papers: one from the provider's point of view, examining how providers change their libraries over time, and one from the client's point of view, examining how client's use these libraries and how they react to changes in these APIs. For this project, we will look at the latter point of view. Specifically, we look to answer the following research questions:

RQ1) What ML libraries do clients use?

RQ2) Do clients change the ML libraries they use over time?

RQ3) What phases of the ML pipeline do clients use these library APIs for?

- (Pre-processing, training, deployment)
- Are certain ML libraries used more for a specific phase?
- Do certain phases change more frequently than others?

RQ4) How do clients react to deprecated APIs?

RQ5) What is the lag for clients updating to new releases of the ML APIs?

- What is the cost of these updates? One heuristic to look at is code churn.

We have begun building a list of repositories that use machine learning libraries to analyze. This includes repositories used in "The Scent of Deep Learning: An Empirical Study"[1], of which Prof. Khohm is one of the authors. Additionally, I have written a script to scrape repository information from any categories specified on "Awesome Open Source – Artificial Intelligence Categories" [2]. We will have to ensure we collect repositories from a variety of domains.

For analyzing the repositories, one of the first steps we had to take was ensure we could extract the imports from python files in these repositories. I wrote the *extract_imports.sh* script, which is located at the base directory of my GitHub repository, which does the following steps:

```
for a list of repositories:
    clone the repo;
    fetch the tags for the repo;
    for all tags:
        checkout tag;
        run ./repo-analyzer/main.py
write all generated JSON files to a single CSV
```

The *run ./repo-analyzer/main.py* command runs a program I wrote that reads in a Python files in a given directory, parses the source code, generates the AST, and then walks the AST to find Imports statements. Once it has collected all the Import statements, it writes them to a JSON file with the repo name, repo version, and file name.

For the next milestone, I plan on doing preliminary analysis of a few sample repositories, including modeling the imports of these projects over time. Additionally, I will extend my script to extract function calls from the python source code of these projects, so that I will be able to analyze the evolution of the client's uses of machine learning libraries. Once I have done a proof-of-concept with extracting this data from a sample of repositories, I will build a dataset with a much larger list of clients.

[1] http://swat.polymtl.ca/~foutsekh/docs/hadhemi-MSR2020.pdf
[2] https://awesomeopensource.com/categories/artificial-intelligence