Ben Rombaut
CISC873
Milestone 2 - Summary
Project Title: An Empirical Study on the Evolution of Deep Learning Library Uses
GitHub Repo: https://github.com/brombaut/CISC873_TermProject

My goals for this milestone were to begin analyzing how the popularity of deep learning libraries over time, as well as to have a proof-of-concept for extracting the API calls that clients use from these deep learning libraries.

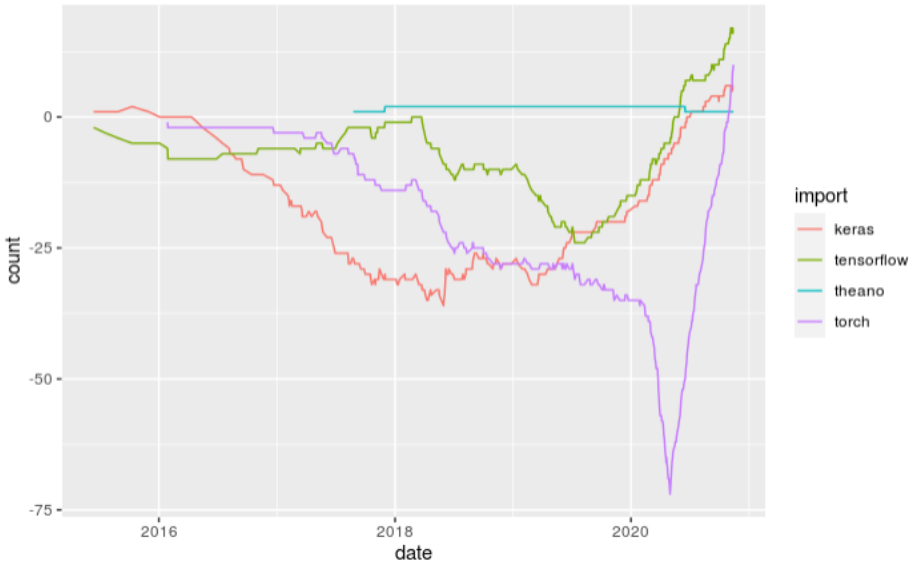
I had previously written a python script [1] that are able to extract which deep learning libraries are imported in files (if any) [2] [3], and writes this information to a CSV [4]. I extended this script to also parse out the call statements from these files [5] [6], and write this data to a CSV[7].

With these scripts, I made a pipeline script [8] that reads in a list of repositories, checks out each release of each of the repositories, and runs the script mentioned above [1] on the codebase, extracting all the import and call statement data for each release for each project and adding it to their respective CSVs. In addition, information about each release of the repository is written out [10]. At the moment, the list consists of only 20 repositories from “The Scent of Deep Learning: An Empirical Study” [11]. When running the full analyses, I will have to ensure that we are looking at a large number of projects from a wide range of domains, in an attempt to minimize any bias.

Once I had these CSV’s generated, I wrote a script to analyze the import differences for each release on each project. For example, for project P, if at release R1, they have a file F that imports the library L, but at release R2, F no longer imports L, then we know that the import was removed from the file from R1 to R2. This process can be generalized up to the project level across releases. So for project P, if at release R1, they don’t have any files that imports the library L, but at release R2, P has at least one file that imports the library L, then we know that the project P added the library L from R1 to R2. The logic for this is in [9].

All work mentioned above is what was done for the data collection and transformation phase of the project. For the data analyse phase, I spent quite a bit of time working with R to try and get some results I wanted. This term has been my first experience with using R, and the language was relatively foreign to be. Towards the end of the previous week, I decided to switch to using Jupyter Notebooks, as I have more experience with Python.

As you can see from the graph below, there is clearly an issue in my data transformation scripts for extracting the import diffs for projects over time. I expected that it would not be possible for the number of projects the library is imported in to be negative, as you can only remove an import after it has been added, but the graph says otherwise.



For the next milestone, I plan on fixing the issue with the import counts not being tallied correctly. Once I have fixed this issue, I am confident I will be able to augment this approach to be able to analyze the client's usages of each library's API over time. Additionally, I am going to organize the project repo a bit better, because it has become quite cluttered.

[1] https://github.com/brombaut/CISC873_TermProject/blob/main/repo_analyzer.py
[2]https://github.com/brombaut/CISC873_TermProject/blob/main/data_transform_scripts/library_imports_finder.py
[3]https://github.com/brombaut/CISC873_TermProject/blob/main/data_transform_scripts/source_imports_parser.py
[4]https://github.com/brombaut/CISC873_TermProject/blob/main/data_transform_scripts/parse_import_jsons_to_csv.py
[5]https://github.com/brombaut/CISC873_TermProject/blob/main/data_transform_scripts/function_calls_collector.py
[6]https://github.com/brombaut/CISC873_TermProject/blob/main/data_transform_scripts/function_calls_finder.py
[7]https://github.com/brombaut/CISC873_TermProject/blob/main/data_transform_scripts/parse_call_jsons_to_csv.py
[8] https://github.com/brombaut/CISC873_TermProject/blob/main/extract_imports_and_calls.sh
[9]https://github.com/brombaut/CISC873_TermProject/blob/main/data_transform_scripts/parse_import_diffs.py
[10]https://github.com/brombaut/CISC873_TermProject/blob/main/data_transform_scripts/write_repo_release_to_csv.py
[11] http://swat.polymtl.ca/~foutsekh/docs/hadhemi-MSR2020.pdf