vPython: Multi-system Analysis (May 2022)

1. Findings

When running Linear Regression programs (pandas and sklearn are used) on both machines:

- The traces in all parts (before/in/after A area) are different between my laptop and RPi
- RPi's trace has higher numbers than my laptop's (~100 000 pops/pushes)
- RPi's traces also have a normal distribution.

Things that have been checked are identical in both machines:

- vPythons Pandas and Sklearn versions (used in the program) but others haven't been checked (e.g. standard libraries)
- vPython versions appeared as 3.9.7+
- ceval.c
- The code used to train linear regression models

With the above findings, more simple programs are run **on my laptop alone** to collect traces and identify. Most significantly, I found that:

- Traces from the Programs importing both *pandas* and *sklearn* once are idempotent
- Traces from the Programs importing sklearn 8 times are idempotent
- Traces from the Programs importing *pandas* 8 times are **NOT** idempotent (Figure 1)
- Traces from the Programs importing *pandas* and reading a CSV (8 times) are **NOT** idempotent (Figure 2)

There are 25 runs and 25 according traces for each mentioned program and it’s a decent number to identify variance in the traces.

simple_one > importing4_manypd.py

```
1   import sys, pandas
2   sys.modules.pop('pandas')
3   import pandas
4   sys.modules.pop('pandas')
5   import pandas
6   sys.modules.pop('pandas')
7   import pandas
8   sys.modules.pop('pandas')
9   import pandas
10  sys.modules.pop('pandas')
11  import pandas
12  sys.modules.pop('pandas')
13  import pandas
14  sys.modules.pop('pandas')
15  import pandas
```

Figure 1

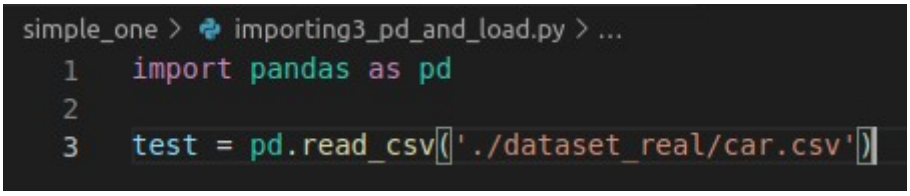

Figure 2

This shows that the importing process of some libraries are not stable. And we should be mindful about the arbitrage nature of working with some Machine Learning model (Random Forest, random train test dataset split for instance). Random seed can be used to produced the same result but their traces also needs to be verified.

2. To-dos

- Run the same simple programs on RPi to
  - Identify idempotence in the mentioned libraries
  - The number of memory operations in different libraries compared to my laptop's. With this information, we can decide what to do next.