Project Title – democratizing data science by analyzing Jupyter notebooks

Project Description:

Computational notebooks such as Jupyter Notebooks (https://jupyter.org/) allows data scientists to combine code, visualizations, and text in a single document. Many data scientists with even little technical knowledge are using them do a variety of data analytics tasks. Previous work has collected a huge amount of such notebooks (https://blog.jupyter.org/we-analyzed-1-million-jupyter-notebooks-now-you-can-too-guest-post-8116a964b536). The goal of the project is to get a better understanding on how data preparation, machine learning, and visualization tasks are done in these notebooks.

Duties/Activities:

1. Download the Data from https://library.ucsd.edu/dc/object/bb2733859v
2. Profile and collect statistics about the data.
3. Identify data preparation, visualization, and machine learning operations.
4. Classify these operations into semantic classes.
5. Create generic pipelines that could be used to characterize most of the notebooks.
6. Given a snippet of operations, is the next operation predictable?

Required Skills: Good analytical skills. Good command of Python, Pandas, and Matplotlib**.**

Preferred Intern Academic Level: At least juniors (3rd year undergrad students). The project can be modified to fit graduate students as well.

Learning Opportunities: Developing research and coding skills in data analytics and data curation. Intensive coding.

Expected Team Size: *1-2*

Mentors

Mourad Ouzzani mouzzani@hbku.edu.qa
Nan Tang ntang@hbku.edu.qa

References and Resources

Exploration and Explanation in Computational Notebooks

- Paper - https://adamrule.com/files/papers/chi_2018.pdf

- Post - https://blog.jupyter.org/we-analyzed-1-million-jupyter-notebooks-now-you-can-too-guest-post-8116a964b536
- Data - https://library.ucsd.edu/dc/collection/bb6931851t https://library.ucsd.edu/dc/object/bb2733859v
- Code - https://github.com/activityhistory/jupyter_on_github
- Talk - https://www.youtube.com/watch?v=trlfzLyDI6U

Towards Understanding Data Analysis Workflows using a Large Notebook Corpus

Paper - https://dl.acm.org/doi/pdf/10.1145/3299869.3300107?download=true https://dl.acm.org/doi/abs/10.1145/3299869.3300107