Developing Interactive Jupyter Notebooks for SDSC HPC Systems

Dhruv Kumar, Leo Gu, Sarah Xie
Mentored by Mary Thomas

SDSC SAN DIEGO SUPERCOMPUTER CENTER

UCSD

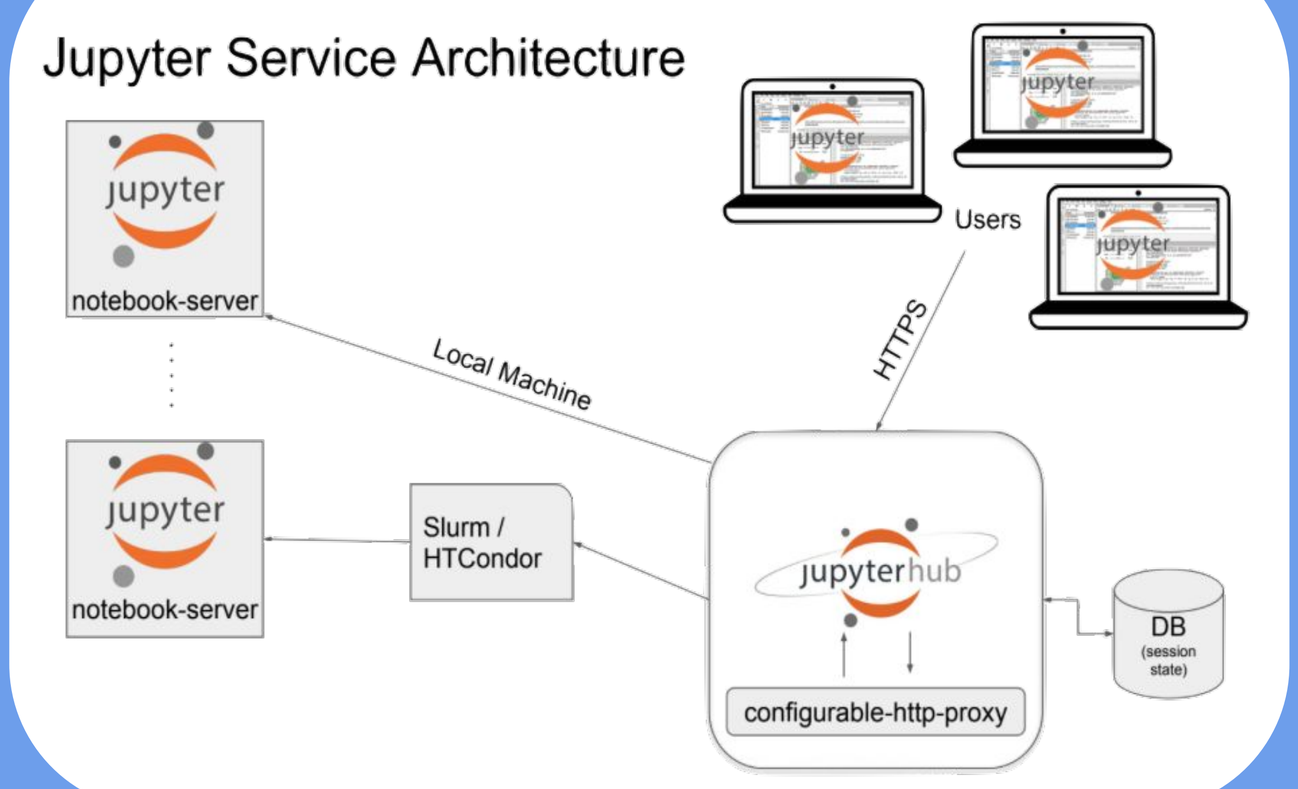
Abstract

The goal of this research project is to develop a set of Jupyter Notebooks that can be used to train users on the Comet supercomputer. These notebooks will be hosted on the SDSC GitHub.io training pages. The purpose of this project is to understand the basics of High Performance Computing, Jupyter notebooks and GitHub. For this, we worked with existing projects to develop new notebooks. The research components are to contribute to the body of knowledge needed for hosting live, dynamic, interactive services that interface to HPC systems, and to learn how to develop interactive notebooks to be used for education and training of the users of these systems. As result of our efforts, we produced several new Jupyter notebooks that will be used for Comet/SDSC training.

Research Process

In order to create a persistent JupyterHub web service, we learned about batch scripts and launching Jupyter Notebooks through conda and linux. Using the interactive software that Jupyter supports, we designed a series of tutorials that guide users through Python basics, running jobs using the Comet Slurm manager, and real-world applications using Python. We then studied High Performance Computing (HPC) and used our experience of the bash environment to launch the Jupyter notebooks on both laptops and Comet. Additionally, we learned about JupyterHub and set up a demo server. Subsequently we did testing for the reverse proxy service which allows users to securely launch a Jupyter Notebook on comet. Lastly, we researched different applications of Python and individually created Jupyter Notebooks on various packages and applications that branched into different fields of science such as machine learning and bioinformatics. As part of our learning curve, we learned to use the Atom IDE, and GitHub repositories to store our work and to clone onto different machines.

Jupyter Service Architecture

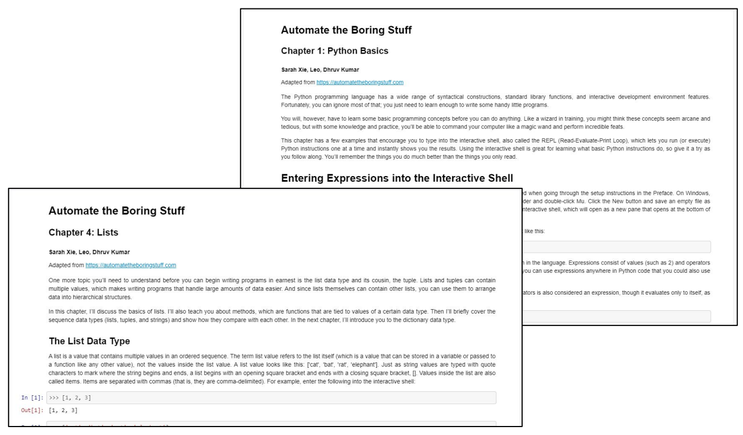


Technologies Learned

- Web and High Performance Computing Architectures
  - Accessing and using the Comet cluster
  - Unix, Running jobs
- Code development:
  - Python
  - JSON
  - GitHub repositories
- Jupyter Services:
  - Jupyter Notebooks
  - Using Kernels (Python, Java, Bash)
  - JupyterLab
  - Plugins/packages: vim, keras, numpy, Tensorflow
  - Reverse Proxy Service
- Connecting to Comet Supercomputer using SSH Tunneling

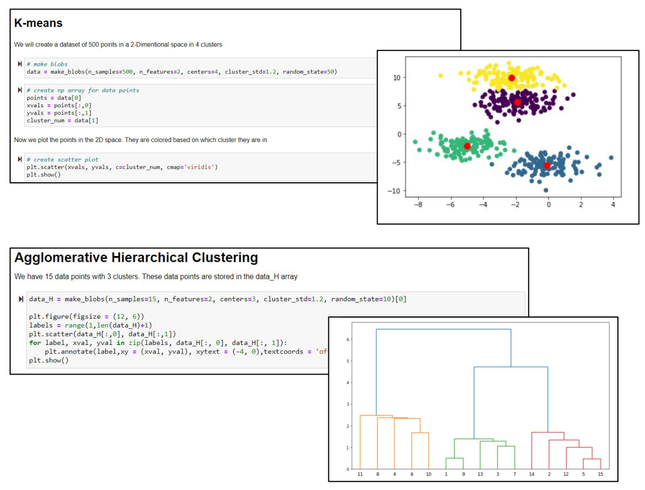
Automate the Boring Stuff with Python https://automatetheboringstuff.com/

By Dhruv Kumar, Leo Gu, Sarah Xie



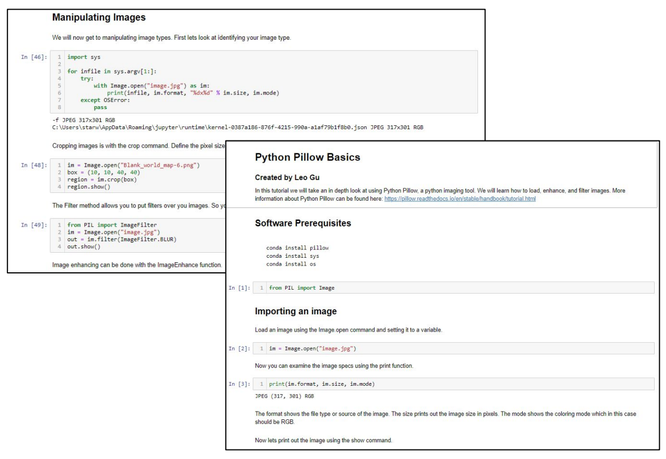
Machine Learning Clustering Algorithms

By Dhruv Kumar

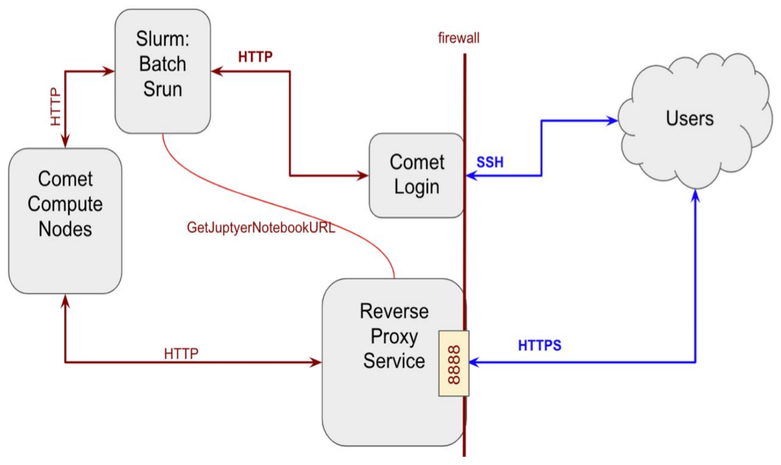


Python Imaging Library

By Leo Gu

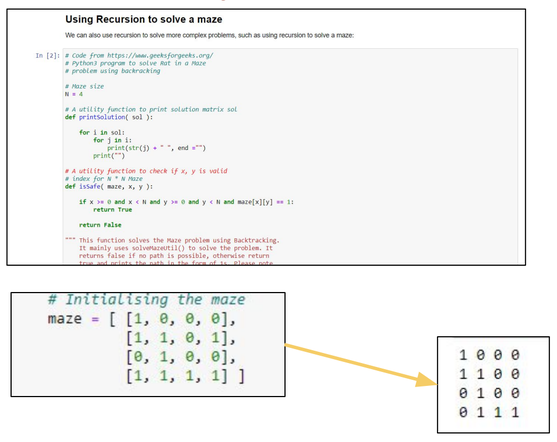


Reverse Proxy Service Architecture



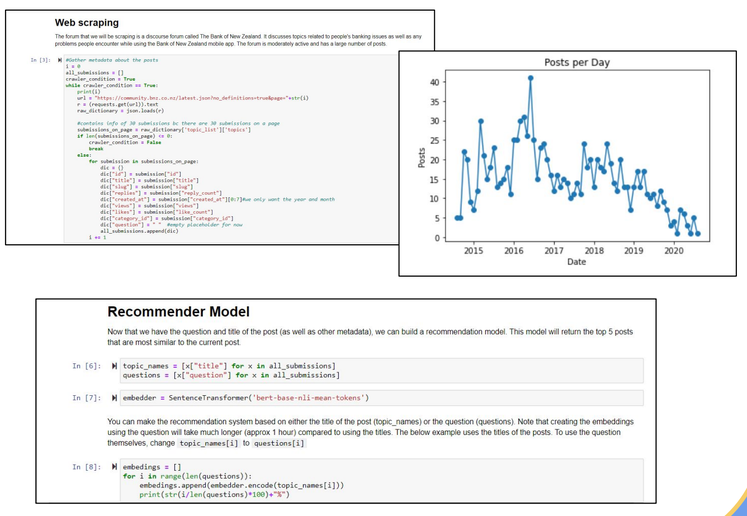
Maze Solving using Recursion in Python

By Sarah Xie



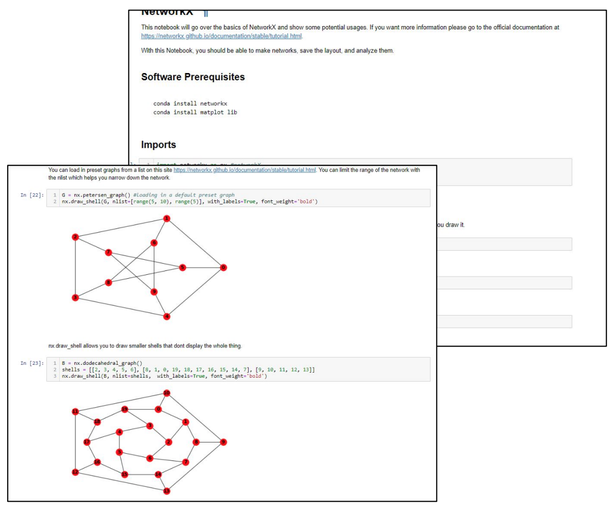
Web Scraping and Recommender System

By Dhruv Kumar



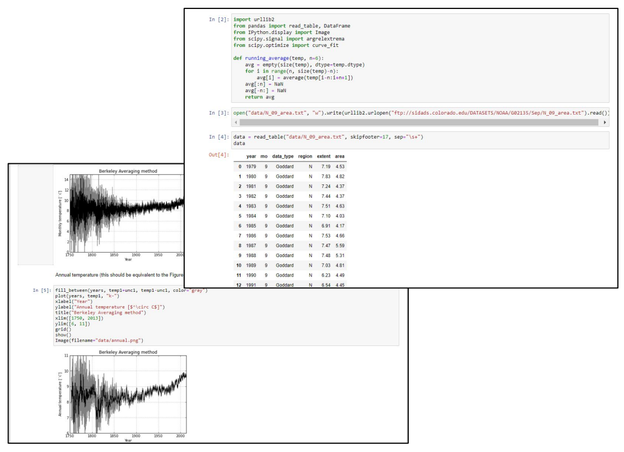
NetworkX and Applications

By Leo Gu



Climate Data and Analysis using Pylab

By Sarah Xie



Conclusions

Over the course of our internship at SDSC, we worked as a team to expand our understanding of High Performance Computing and the implementation of Jupyter Notebooks using Python. We also faced several challenges that we were able to overcome, including: installing miniconda, python package installation; launching notebooks on Comet; and spawning notebooks via a remote connection to Comet. Future work includes expanding our tutorial library and developing a JupyterHub authenticator that is compatible with UCSD accounts.

Acknowledgments

Ange Mason
Scott Sakai
James Mcdougal
Marty Kandes
Mahidhar Tatineni

References

1. Sinkovits, Robert: https://github.com/sinkovit/PythonSeries
2. Jupyter Project Repository, https://github.com/jupyter
3. Notebook references:
   https://networkx.github.io/documentation/stable/
   https://github.com/certik/climate
4. Automate the Boring Stuff with Python, https://automatetheboringstuff.com/