Extracting Metadata From Jupyter Notebooks

Presented by Ben Galewsky (bengal1@Illinois.edu)

ILLINOIS
NCSA | National Center for Supercomputing Applications

Summer 2018 FUNding Friday Project

We were awarded a FUNding Friday grant for a project to automatically extract metadata from Jupyter Notebooks on GitHub

Project Team:

- Keith Maull - NCAR
- Sean Gordon: HDF Group
- Ben Galewsky: NCSA

Project "**NBMeta**"

- https://git.io/fAf5T (https://github.com/ESIPFed/NbMeta)
- contains project motivation, some data investigations and is active

How to Find and Analyze Notebooks

GitHub API is great - you can query the entire universe to find repos that contain Jupyter Notebooks:

```
'language:"Jupyter Notebook" is:public'
```

Then query the repo to find files with `.ipynb` extension

Jupyter Notebooks are represented as JSON Documents.

Drawbacks to GitHub API

They have very restrictive API rate limits

This means that the scripts that query notebooks must have extensive sleep periods to avoid triggering Abuse Limits.

What to do?

Let’s see, we need search a set of JSON documents...

What to do?

Let’s see, we need search a set of JSON documents...

Analysis Environment

- Created a three node Docker Swarm with 0.5Tb attached storage
- Deployed a sharded Mongo DB
- Collections:

repositories	Repositories along with GitHub metadata
notebooks	The notebooks' content as json
imports	Output from pipeline that extracts the libraries that the notebook imports
links	Output from a pipeline that extracts URLs from the notebook code.

The Data: Why is this interesting?

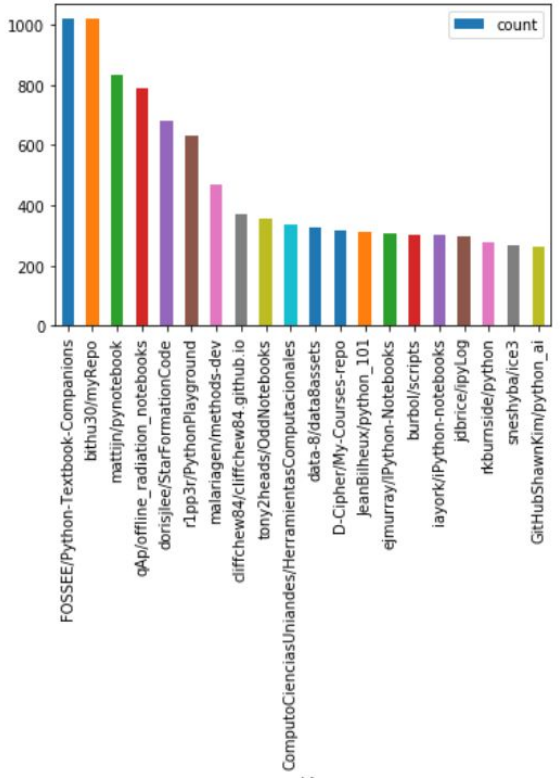
- Quantify and *Qualify* the number of notebooks in the Gitverse (extending the work of Rule, et al, 2018 doi:10.1145/3173574.3173606)
- Understanding these issues leads to metadata framework(s) for Notebooks
- Metadata for notebooks leads to well-indexed notebooks and hence findable notebooks
- Notebooks that cannot be found, cannot be used and valuable work is unnoticed and ignored
- Data nerds love to ask questions ...

Questions we have enabled ...

- What are the popular notebooks (by proxy through repo stargazers)?
- What are active notebooks (through commit changes, etc.)?
- Is there a taxonomy of notebooks that we can uncover? (e.g. tutorial/training notebooks, analysis notebooks, paper notebooks, demo, etc.)
- What are the relationships between links (urls) and the domain, discipline and purpose (type) of the notebooks?
  - Can domain or purpose (type) be inferred from links and other attributes? (e.g. training, analysis, demo, etc.)
  - What can DOIs allow us to infer?

… and for the data nerds

What are the top 20 repos by number of `.ipynb` files?



(and more) …

	ipynb	py	other
coells_100days	0.953271	0	0.046729
fchollet_deep-learning-with-python-notebooks	0.904762	0	0.0952381
lijin-THU_notes-python	0.827586	0.0229885	0.149425
fastai_courses	0.533333	0.186667	0.28
dennybritz_reinforcement-learning	0.492063	0.301587	0.206349
jakevdp_PythonDataScienceHandbook	0.444444	0.0588235	0.496732
aymericdamien_TensorFlow-Examples	0.44	0.44	0.12
nlintz_TensorFlow-Tutorials	0.428571	0.428571	0.142857
norvig_pytudes	0.396226	0.169811	0.433962
ageron_handson-ml	0.375	0.0416667	0.583333

What is the mix of `.ipynb` to `.py` to other file types sorted by `.ipynb` percent (descending)?

What’s next?

- Broad analysis of the data, disseminating the analysis and outcomes
- Developing a strategy for sharing the full dataset, metadata and data server (?)
- Development of recommendations, guidelines, best practices and tools to enable “metadata-first” notebooks

Credits

Thank you to National Data Service and San Diego Supercomputer Center for hosting our database!

SDSC
SAN DIEGO SUPERCOMPUTER CENTER

Double thank you to ESIP for the grant and opportunity to explore this.

bengal1@Illinois.edu

ILLINOIS
NCSA | National Center for
Supercomputing Applications