Day 10: Pip, Fixed Column Data, Freebase

Raymond Yee

February 21, 2013 (`http://bit.ly/wwod1310`)

Agenda

- Let's step back and explicitly fill in some Python development context
- Wikipedia
- Freebase
- Homework

Announcements

- Keeping Course outline up to date: `http://bit.ly/wwod13outline`
- In-class Midterm rescheduled to Day 17: Tuesday, March 19, 2013

Campus subscription to Safari books

UC Berkeley library proxy

Python distributions

Though the future of Python is Python 3.x, we are using Python 2.7.x in this course.

Setting up Python and handling *dependencies* is both eseential and often painful. I've tried to hide as many of the complexities as possible but recommending we all use the same distribution, namely Enthought Python Distribution (Academic).

Environments

- Piazza: Which operating system(s) are you using?
- Piazza: Which version of Enthought Python Distribution are you using?

Python modules

In addition to the awesome “batteries included” nature of Python because of the Python standard libraries, there’s a huge world of modules.

Many are available in PyPi.

Using pip instead of easy_install

I really recommend:

- Virtualenv and pip Basics
- Python Dev Environment Screencast (by Apreche) on YouTube

Facets for further interaction

Results from Piazza Polls

- How difficult did you find the homework from Day 4?
- Are you looking for more challenge / discussion / links to resources than currently presented?

Twitter

(@WorkingOpenData / @rdhyee)

Stackoverflow

Go, AJ, for posing this question and getting good answers on stackoverflow

Evernote

`https://www.evernote.com/pub/rdhyee/workingwithopendata2013`

Zotero

Using requests, lxml, and some simple geocoding example

notebook on requests, lxml, geocoding

Wikipedia

- user pages: e.g., `http://en.wikipedia.org/wiki/User:RaymondYee`
- example geographic pages:
  - `http://en.wikipedia.org/wiki/California`
  - `http://en.wikipedia.org/wiki/Alameda_county`
- Infobox

Wikidata

- `http://www.wikidata.org/wiki/Wikidata:Main_Page`
- `http://www.wikidata.org/wiki/Q99` for California
- `http://www.wikidata.org/wiki/Q107146` for Alameda County

Freebase

Use `https://dev.freebase.com/` instead of `http://freebase.com`

Planet example

- `https://dev.freebase.com/astronomy/planet?schema`
- `https://dev.freebase.com/astronomy/planet?instances`
- queryeditor: `http://tinyurl.com/an85xhs`

Freebase: some geographic examples

California: `https://dev.freebase.com/m/01n7q`

`http://www.freebase.com/view/en/california`

`http://dev.freebase.com/en/california` ->
`https://dev.freebase.com/m/01n7q`

California Freebase types: `http://tinyurl.com/af3g7ua`

Freebase: governors + party affiliation example

`http://tinyurl.com/a4f3r4s`

Freebase: centroids of states

`http://tinyurl.com/cjuy6k3`

dbpedia

We'll come back to `http://dbpedia.org/About`.

Homework

- Parse `DataDict.txt` (`https://raw.github.com/rdhyee/working-open-data/5ef3932b4ff7cadf1f06ca01eb852ad71361894a/data/census/DataDict.txt`) into a DataFrame – send me your notebook by Friday, March 1, 2013 at noon. Read the file from github. Hint: if you use requests to read the file, you may need to turn verify off for requests.get: `http://stackoverflow.com/questions/10667960/python-requests-throwing-up-sslerror`
- Generalize Freebase map for states into one that works for counties. Hints: `http://wiki.freebase.com/wiki/MQL_Read_Service#cursor` and `https://developers.google.com/freebase/v1/mql-overview#looping-through-cursor-results`.
- Explore what I've shown you in this class.
- Keep working on your projects.