Day 1: Introduction to Working with Open Data

Raymond Yee

January 22, 2013 (`http://bit.ly/wwod1301`)

Course Overview

INFO 290T- Working with Open Data
http://www.ischool.berkeley.edu/courses/290t-wod Spring 2013 /
CCN: 41640
T,Th 2:00-3:30pm 210 South Hall
Office Hours: TBD (tentatively: T, Th 3:30-4:30pm, 303A South Hall)
Instructor: Raymond Yee, Ph.D.
Contact info: `yee@berkeley.edu` (@WorkingOpenData / @rdhyee)
Tutor: Jacob Portnoff (`jacob.portnoff@ischool.berkeley.edu`)

bspace: `https://bspace.berkeley.edu/portal/site/226977ea-5d47-4d0e-9119-616d78c8641a`

Goals Today

- Introduce prospective students as to the purpose, structure, content of the course
- Begin to think together about open data and using Python to analyze open data
- Start building a learning community that will work together
- Build basic communication structures

Course Description

Open data – data that is free for use, reuse, and redistribution – is an intellectual treasure-trove that has given rise to many unexpected and often fruitful applications. In this course, students will

1. learn how to access, visualize, clean, interpret, and share data, especially open data, using Python, Python-based libraries, and supplementary computational frameworks
2. understand the theoretical underpinnings of open data and their connections to implementations in the physical and life sciences, government, social sciences, and journalism.

Working with Open Data (WwOD) is a *technical* course with a strong focus on the social-political context and domains of application of open data.

Prerequisite

Info 90 (Programming for Computing Applications) or equivalent background with Python.

Expectations

- Participants will work on tangible projects related to the overall theme. They can select from a list of projects I design or they can propose other projects of comparable scope and intent. However, we will find ways for the projects to combine together into a larger super-project.
- Participants will be heavily involved in learning from and teaching each other, depending on each other for the course's collective success. The class will provide support for students to work together, not only in their own project group, but also course-wide and even with people working with us outside of the class.
- To ensure that our projects remain grounded in the "real world," we'll be working to engage outside users for our projects from the outset.
- The course will be designed to enable the larger community to participate.

Main Textbook

Wes McKinney. `Python for Data Analysis`. (O'Reilly Media, 2012). I strongly recommend getting a paper copy as well as accessing any electronic versions

- oreilly.com
- Proquest.safaribooksonline at UCB

Supplementary Materials

I plan to supplement the book with materials covering the following topics:

- open data, open content in various fields
- using JavaScript, HTML5, CSS together with Python for data presentation, analysis, and visualization, (e.g., d3.js)

In addition to survey materials on the public domain, creative commons, and open data movements, I'll focus us on

- Wikipedia, dbpedia, Freebase data
- census data

and other data sets still to be determined, probably large open scientific data sets

Course Outline

Tentative

1. Course Intro Lecture
2. Review of Python & iPython (Chap 3 and Chap 13) + setup EPD
3. NumPy and Getting started with pandas I (Chap 4, 5)
4. Open Government Data I: US Census data: introduction how to work with
5. Plotting and Visualization (Chap 8)
6. Data Loading, Storage, and File Formats (Chap 6)
7. Wikipedia: API, data structure
8. NumPy and Getting started with pandas II (Chap 4, 5)
9. Freebase, dbpedia, wikidata
10. Data Wrangling: Clean, Transform, Merge, Reshape (Chap 7)
11. working with geodata I
12. JavaScript-based visualization I
13. Publishing open data including LOD
14. Google Refine
15. working with geodata II

Grading Scheme

Grading Scheme:

Tentative

1. problem sets (30%)
2. mid-term exam (20%)
3. project proposal (5%)
4. final project (25%)
5. participation (20%)

Projects

In the projects, students will synthesize and demonstrate what they have learned throughout the course. The projects need to be thorough analyses of some open dataset or datasets.

Students will have opportunities to brainstorm ideas, choose a specific focal point (drawing from structured feedback from other students and the instructor), craft a proposal for their projects, and then present their work at the end of the course.

Laptops in classroom

Ideally I would like everyone to bring a notebook computer to class so that we can work together in class on programming assignments. **If you are not able to do so, check in with me.**

Why Python?

see McKinney's narration: `http://proquest.safaribooksonline.com/book/programming/python/9781449323592/1dot-preliminaries/id2700570`

Working definition of open data

From
http://en.wikipedia.org/w/index.php?title=Special:Cite&page=Open_data&id=532390265:

> *Open data is the idea that certain data should be freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control.*

http://opendefinition.org/:

> *A piece of content or data is open if anyone is free to use, reuse, and redistribute it — subject only, at most, to the requirement to attribute and/or share-alike.*

data.gov as a good example

`http://www.data.gov/`

`http://www.data.gov/about`:

> *The purpose of Data.gov is to increase public access to high value, machine readable datasets generated by the Executive Branch of the Federal Government."*
>
> *A primary goal of Data.gov is to improve access to Federal data and expand creative use of those data beyond the walls of government by encouraging innovative ideas (e.g., web applications). Data.gov strives to make government more transparent and is committed to creating an unprecedented level of openness in Government. The openness derived from Data.gov will strengthen our Nation's democracy and promote efficiency and effectiveness in Government.*

datasets in data.gov

What is the population of California?

Some sources for data on the Population of California

`http://quickfacts.census.gov/qfd/states/06000.html`

`http://en.wikipedia.org/wiki/California#Population`

OKFestival as indicator of vibrancy of the international open data community

- `http://okfestival.org/`
- `http://okfestival.org/after/`
- video streams
- start with opening plenary
- final report

list of working groups

`http://okfn.org/wg/` includes:

- open science
- open bibliography
- open spending

other ties

- OKFN

Activity

- What you hope to learn and accomplish in the course?
- Name 2 to 4 types of data (or datasets) that interest or intrigue you. Bonus: explain why
- What's one of the more complicated example of Python programming you've done so far?
- What questions do you have for the instructor?

Examples of Open Data

- http://www.data.gov/
- https://data.sfgov.org/
- https://data.acgov.org/
- http://data.openoakland.org/
- http://www.socrata.com/discover/video-case-study-somerville-ma/
- http://sunlightfoundation.com/projects/
- http://oad.simmons.edu/oadwiki/Data_repositories#Astronomy
- http://opencontext.org/
- http://courtlistener.com/
- http://aws.amazon.com/publicdatasets/
- http://openmetadata.lib.harvard.edu/bibdata
- http://www.bart.gov/schedules/developers/api.aspx / http://www.bart.gov/schedules/developers/index.aspx
- http://www.ncdc.noaa.gov/
- http://www.propublica.org/tools/

Motivation: why I care about open data and why you might care

Traditional motivations given for open government data:

- transparency
- accountability
- efficiency
- innovation

My personal interests in the area:

- Open data useful testbed for working on data of all sorts, because of zero financial costs and minimal restrictions on use, reuse, redistribution
- Growing community around open data because of these low barriers. . . democratization of data. . . many more of us can participate in working with open data and attract a wide range of people I love to learn and to think and to understand, a big believer of computational and information systems as mind augmenters/extenders and open data (as

Quick Examples

money and politics

I read a story in Sunday's New York Times in connections made between political fund-raising and specific decisions being made in Congress. So how can we participate in this process we can certainly read stories which have made these connections. We first verify that yes, `The New York Times` is correct in making these connections. We can use this process to discover new connections – wan't obvious from reading article on paper but online article has links to opensecrets.org – e.g.,
`http://www.opensecrets.org/politicians/contrib.php?cycle=2012&cid=N00004643&type=C`

desire for transparency in China

Fascinating to read NYT saying that the Chinese middle class demanding government transparency:
`http://www.nytimes.com/2013/01/20/world/asia/in-china-discontent-among-the-normally-faithful.html`

Random and not-so random questions for me that open data can help answer

Reading the news, world news, local news, tech news, understanding new contexts, deepening old interests, controlled serendipity.

- What music to listen to
- What book to read?
- When can I see the next episode of White Collar?
- When was BWV 156 (Ich steh mit einem Fuß im Grabe) first performed?
- How to invest our money?
- What programming language to learn next?
- What charity to give money to?
- What restaurant to try?
- What to cook?
- How should we take care of our physical health?

Some Big Questions for the Course

- What are the essential characteristics of open data?
- What are the costs and benefits of open data?
- How to map the universe of open data? What's out there? What data is not available in open form?
- What can we learn from open data?
- What business models?
- What are people doing with open data?
- What are the different common formats used to represent open data? (e.g., CSV, XML, KML, SHP in data.gov) – and how can we use Python to process those formats?
- What are the issues that we face in combining open data with closed data

Related upcoming events in the area

- Streams, Gardens, and Clouds: Visualizing Dynamic Data for Engagement, Education and the Environment: A CITRIS Data and Democracy Event to Celebrate Data Innovation Day (Thursday, Jan 24)
- Bay Area Wikipedia Women's Edit-a-Thon 4! Saturday, Jan 26
- OSM Data Edit-a-thon via Come out and map: January 26 #editathon
- Big Money, Big Data, and Datafest Feb 2, 3

Homework

Answer the Getting to Know You survey in bSpace

Readings

- read `PfDA`, Chap 1 Preliminaries, especially the installation instructions for EPD Free for your computer platform. I want you to try installing EPD Free (or EPD Academic) before class on Thursday.
- read `PfDA`, Chap 3
- skim `PfDA`, Appendix: Python Language Essentials – to help remind yourself of key elements of standard Python
- skim `PfDA`, Chap 2 Introductory Examples