Final Project Instructions

DS-GA-1007

Project Requirements

1. Students can work individually or on teams of up to 3. Team members must be identified in the project proposal. All team members are expected to contribute to the project, and will receive the same grade.

2. A brief project proposal providing a description of the intended project, along with a list of the datasets that will be used, must be submitted before the project is undertaken.

3. Use publicly available datasets (see samples below.) You can either download a dataset and use the downloaded version, or access a dataset directly via its supported APIs.

4. The project must demonstrate a good understanding of Python, pandas, matplotlib, and NumPy. In particular:

    a. Loading a non-trivial dataset into pandas objects (DataFrame and Series)
    b. Perform some kind of meaningful analysis of the data using pandas and/or NumPy computational and data analysis tools.
    c. Display the results of the analysis using matplotlib.
    d. Allow the user to interactively control the analysis and display of the data.

5. The project must include a user guide that describes how to run the program.

Grading

Grading will be based on:

- Program correctness
- Readability and maintainability of the code
- Program and module structure
- Quality of documentation
- Demonstrated understanding of Python
- Demonstrated understanding of pandas, NumPy, and matplotlib

Examples of User Interaction

There are many ways to interact with users. Choose whatever you feel most comfortable with, and that you think can be achieved in the time available. The following are suggestions, but use something else if you think it is more appropriate.

Obtaining Input From The User

- Command line
- Reading directly from terminal
- Input file
- Graphical User Interface (TK, Qt, Web form)

Presenting Output To The User

- A single matplotlib plot saved in a pdf/png file
- Animated GIFs
- Terminal text output + matplotlib pop-up + matplotlib plot to file
- HTML file (report) including text and figures
- Graphical User Interface (TK, Qt)
- HTML5
- HDF (data format)

Sample Projects

1. Datasets: City of NY Taxi; NOAA weather

   Allow the user to enter a location in Manhattan (e.g. address or coordinates) and a day of the year. The program would plot a graph of the average, min, and max wait time for a taxi over the 24 hour period, assuming that a taxi must be in the same block, and empty, for it to be available. The program would also provide the option to view the same information depending on the weather, such as raining, hot, cold, etc.

2. Datasets**:** Historical financial; a managed fund (mutual, hedge, etc.)

   Allow the user to enter a date range, and plot a graph comparing the fund's daily performance against the actual market data over this period. Provide a statistical analysis of how the fund performed over the whole period. Allow the user to provide additional datasets to perform a multi-way comparison.

3. Datasets**:** NOAA weather data

   Allow the user to enter a date, then generate an animated gif showing an isosurface representation, using color to represent intensity, of weather data over a 100 mile radius centered on Manhattan. The gif frames should be for each hour over the subsequent 7 days. The user should be able to choose the type of data to be displayed, such as rainfall, cloud cover, etc., depending on what is available in the dataset. The user should also be able to choose to display the min, max, or average of the data over the same period.

Example data sets

- One of the datasets provided by the city of New York (http://data.cityofnewyork.us)
- Financial data supported by Pandas (Yahoo! Finance, Google Finance, World Bank, etc.)
- NOAA weather and climate data (http://www.ncdc.noaa.gov/data-access)
- Federal Election Commission Database (http://www.fec.gov/portal/download.shtml)
- UCI ML repository (http://archive.ics.uci.edu/ml/)
- Another overview of datasets geared towards Machine Learning (https://github.com/datasciencemasters/go/blob/master/datasets.md)
- Yelp dataset (https://www.yelp.com/academic_dataset)
- Amazon Web Services public datasets - might be on the bigger side, like the 500TB common crawl corpus (https://aws.amazon.com/datasets)
- List of datasets on IKANOW (http://www.ikanow.com/where-can-i-get-the-best-free-data-for-new-analysis-projects/)
- List of datasets on reddit (http://www.reddit.com/r/data/top/?sort=top&t=all)

Sources of Inspiration

These are some beautiful visualizations. Most of them are using interactive plotting and javascript libraries, which might be a bit too advanced for most (if you're motivated to do this, by all means go for it though!) Use them as inspiration of people using publicly available data to analyze and visualize.

- City of New York (http://data.cityofnewyork.us)
- US Budget visualization: http://solomonkahn.com/us_budget/
- Hans Rosling and Gapminder: http://www.gapminder.org/ - you might have seen his Ted talk. Also has data on the gapminder website!
- This famous collection of ipython notebooks, hosted on github. If you look at the "Scientific computing and data analysis" section, you'll definitely find many examples of

people using pandas+numpy+matplotlib to do cool things. https://github.com/ipython/ipython/wiki/A-gallery-of-interesting-IPython-Notebooks

- The New York Times published a lot of great visualizations. Google around or look at this overview of 2013: http://www.nytimes.com/newsgraphics/2013/12/30/year-in-interactive-storytelling/#dataviz and 2012: http://www.nytimes.com/interactive/2012/12/30/multimedia/2012-the-year-in-graphics.html

Extra Resources

- Prettyplotlib - pretty defaults for matplotlib
- Google charts api to easily create interactive plots (https://developers.google.com/chart/)
- If you are creating interactive plots with javascript / google charts, consider using a templating engine (http://jinja.pocoo.org)