Web scraping for Financial Data

Difficulty Level : Easy
Duration : 30-45min
Prior Experience : Basic knowledge of python or basic coding experience

Description:

In this workshop, we will be learning how to use Python and LXML to scrape stock market data from publicly available websites. Then, we will be using a popular financial data API to compare current web scraped priced with older data. This is a great way to combine a commonly used technique in computer science and data science while also learning about some key metrics in finance. Participants will be taught the basics of web scraping including concepts such as crawl rate, terms of service, and robot.txt. We will also talk about the legality of web scraping and how and when to do it. Next steps will include analyzing and graphing the data and even integrating it with your email.

After this workshop, students should be able to apply the basic principles of web scraping and use the given code snippets on other public websites to get similarly formatted data. Participants can then use the data for things like statistical analysis, graphing, and automatically compiling reports.

What you need to start :

- Text editor (for this tutorial, I will be using Atom, but any other works as well!)
- Terminal

Things to do BEFORE the workshop:

- Install Python 3 (Instructions here and here)
- Compile a list of stocks and their ticker ID's (5-10)
- go to https://www.quandl.com/tools/python and click "Sign Up Now For A Free API Key" at the bottom.

Further instructions also at bottom of file!

Terms to Know:

- Web scraping – is a term for various methods used to collect information from across the Internet
- Bid - is the highest price that a buyer is willing to pay for a good
- Market Cap - it is the market value of a company's outstanding shares
- Beta - is a measure of a **stock's** volatility in relation to the market.
- PE Ratio- is the **ratio** for valuing a company that measures its current share price relative to its per-share earnings.

The Steps (a very high overview) & the Structure :

- Talk about web scraping (explanation + other common applications)
- Pick stocks and website to scrape (we will most likely be using yahoo)

- URL constructions & explanation
- Walk through of code
- Download and parse HTML
- Save data to JSON files
- More python features

Students can code in individually, groups, or just follow along. Full solution with detailed instructions will be released after the workshop!

How to set up

```
$ /usr/bin/ruby -e "$(curl -fsSL
https://raw.githubusercontent.com/Homebrew/install/master/install)"
$ brew install python3 git tree wget
$ python3 -m venv env
$ echo $PYTHONPATH  # Output isn't blank, problem!
/Users/xxx/anaconda/lib/
$ rm -rf env  # Remove environment from previous step and start over
$ unset PYTHONPATH
$ python3 -m venv env
$ source env/bin/activate
##### a working python env! ######


#####now add some packages while env is active ######
$ git clone git://github.com/requests/requests.git # add some python packages
$ cd requests
$ pip install .
$ cd ..
$ sudo port install py27-lxml   # add some python packages
$ pip install lxml
$ pip install pandas-datareader  # add some python packages
```

ALSO: go to https://www.quandl.com/tools/python
And click “Sign Up Now For A Free API Key” at the bottom. We will need it for the API portion of workshop