Intro to Web Scraping

@vicohyeh | vicyeh.me

Overview

- What is Data?
- Web Scraping
- Python Beautiful Soup

What is Data?

“Information used to calculate, analyze or plan something” ~ Merriam-Webster

Data

- ex. survey results, purchase history

#	Purchase Code	Supplier	Product	Quantity	Unit Price	Payment	Method	Date
1	bb6b7e1b3d46bfe	Chantale Stone	Iphone 6	5	500	2500	cash	29-01-2015
2	4df9d5ab3a97733	Colette Gill	Iphone 6 plus	2	700	1400	cheque	21-01-2015
3	5699fc129839967	Chantale Stone	Samsung Galaxy Note 4	3	300	900	cash	29-01-2015
4	4f89260ee9d8854	Colette Gill	Casio CTK	3	1200	3600	cheque	27-01-2015
5	f8c487618a53f05	Keely Rowland	Macbook Pro	3	1100	3300	cash	29-01-2015
6	2b0f206835ebbb6	Keely Rowland	Ibanez Guitar	5	2000	10000	cheque	29-01-2015
7	0345b62c8fa32c8	Joana Ethen	Rayban Clubmaster	20	500	10000	card	29-01-2015
8	5ef83d75fe0db97	Joana Ethen	Casio CTK	3	500	1500	cheque	29-01-2015
9	02bae24b0b7ed98	Joana Ethen	Samsung Galaxy Note 4	4	300	1200	cheque	13-01-2015
10	6dee340a05a1145	Joana Ethen	Yamaha Electric Guitar	4	500	2000	cash	29-01-2015

howing 1 to 10 of 10 entries

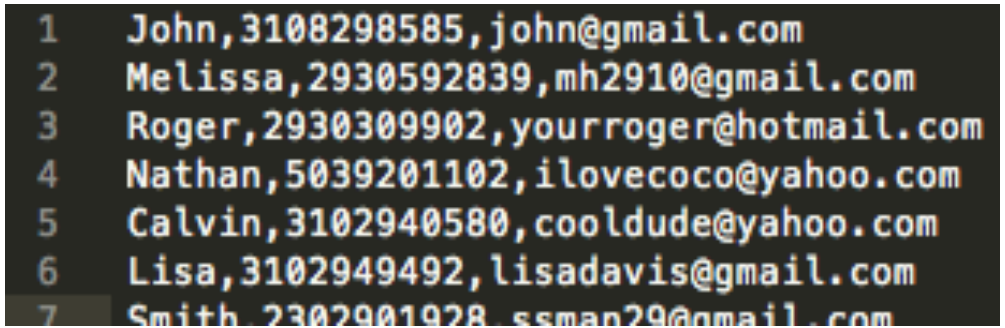
Previous 1 Next

Raw Data

- Data that have not been collected but not formatted/analyzed.

Data format

Comma-separated values (CSV)



```
1  John,3108298585,john@gmail.com
2  Melissa,2930592839,mh2910@gmail.com
3  Roger,2930309902,yourroger@hotmail.com
4  Nathan,5039201102,ilovecoco@yahoo.com
5  Calvin,3102940580,cooldude@yahoo.com
6  Lisa,3102949492,lisadavis@gmail.com
7  Smith,2302901928,ssman29@gmail.com
```

JSON

```
{ "users":[
            {
                "firstName":"Ray",
                "lastName":"Villalobos",
                "joined": {
                    "month":"January",
                    "day":12,
                    "year":2012
                }
            },
            {
                "firstName":"John",
                "lastName":"Jones",
                "joined": {
                    "month":"April",
                    "day":28,
                    "year":2010
                }
            }
    ]}
```

Web scraping?

Pulling data straight out of HTML

Rule of thumb

Any content that can be viewed on a webpage can be scraped.

Pros vs Cons

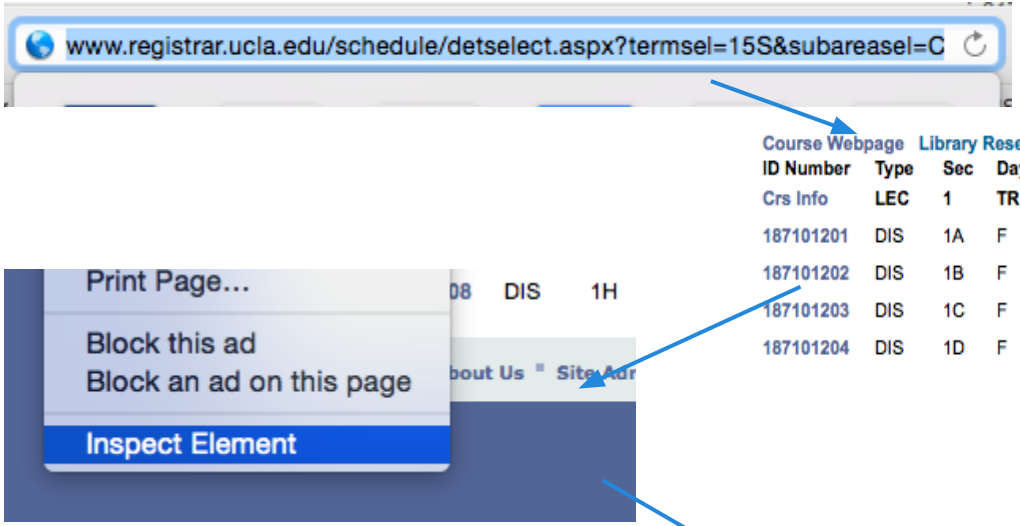
Pros

- No need to read lengthy API Docs
- Usually able to retrieve more data than an official API provides
- Make you sound smart

Cons

- Less intuitive
- If HTML changes, web scraper has to change

Strategy



Course Webpage Library Reserves Textbooks

ID Number	Type	Sec	Days	Start	Stop	Building	Room	Res't	#En	EnCp	#WL	WLCp	Status
Crs Info	LEC	1	TR	2:00P	3:50P	WGYOUNG	CS50	Yes	242	280	0	0	Open
187101201	DIS	1A	F	10:00A	11:50A	PUB AFF	1337	Yes	30	35	0	0	Open
187101202	DIS	1B	F	10:00A	11:50A	BOELTER	3400	Yes	27	35	0	0	Open
187101203	DIS	1C	F	10:00A	11:50A	HUMANTS	135	Yes	28	35	0	0	Open
187101204	DIS	1D	F	10:00A	11:50A	PAB	1749	Yes	33	35	0	0	Open

```
<td class="dgdClassDataColumnSpacer"> </td>
▼<td class="dgdClassDataEnrollTotal">
  ▼<span id="ctl00_BodyContentPlaceHolder_detselect_ctl02_ctl02_EnrollTotal">
      <span class="bold">242</span>
    </span>
  </td>
  <td class="dgdClassDataColumnSpacer"> </td>
▼<td class="dgdClassDataEnrollCap">
```

Python Beautiful Soup

Python library for pulling data out of HTML and XML

- Support popular HTML parsers
- Easy to navigate, search, and modify the parse tree

Installation

Linux

apt-get install python-bs4

apt-get install python-lxml

Mac

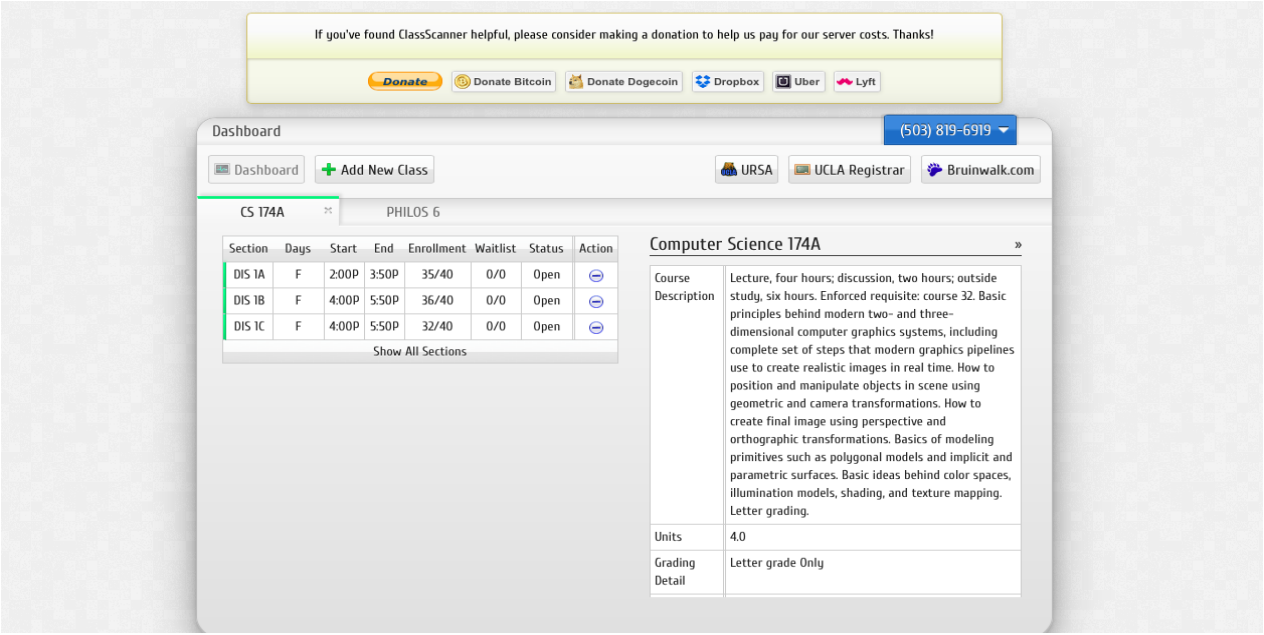
pip install beautifulsoup4

pip install lxml

Hack Time!

Applications

ex. Class Scanner



Further Reading

- Beautiful Soup Documentation

http://www.crummy.com/software/BeautifulSoup/bs4/doc/

- *“Gaming the UCLA Enrollment Process”* by Shahid Chohan

http://shahidchohan.com/gaming-the-ucla-enrollment-process/

Questions?

Contact me at vic.yeh@ucla.edu

Or tweet me @vicohyeh :)