Scraping information

A few notes...

- Before you get started, take a few seconds to see if there are any public apis for the information that you want.
- If you’re trying to “collect data” from any websites then this lecture is absolutely for you.

Public APIs

- If websites benefit from you using their data/service or are just feeling nice they often release a public API
  - What is an API?
    - Application Programming Interface
    - Easy way to “communicate” with a website’s application/service
      - Websites define APIs and their functionality and you communicate with it within the confines of their defined functionality
    - Scraping for information is inconvenient!

USE APIS!!!!

- Berkeley API for classes: https://developer.berkeley.edu/
- Amazon Product information: https://affiliate-program.amazon.com/gp/advertising/api/detail/main.html
- Twilio API: http://www.twilio.com/docs/api/rest
- Instagram API: http://instagram.com/developer/

USE APIS!!!

- Facebook API: https://developers.facebook.com/
- Edmunds API: http://developer.edmunds.com/
- Zillow API: http://www.zillow.com/howto/api/APIOverview.htm
- Google Maps API: https://developers.google.com/maps/

USE APIS!!!!

- ESPN API: http://developer.espn.com/docs

How to API: Overview

- Need API key to access an API
  - Necessary for companies to keep track of what you’re looking up.
  - Data is valuable: if the company just dumps ALL of the information to you, their data isn’t valuable anymore.
- After obtaining key, everything is easy from there :)

How to API: Twilio Walkthrough

1. Get Twilio Python API set up on your computer (hard part)
    a. Put export PYTHONPATH=$HOME/local into the file .bash_profile in your home directory
    b. Terminal: mkdir ~/local
    c. Terminal: easy_install -d ~/local twilio
2. Register for a Twilio account to get a number (write the number down) and find the account SID and the API key here https://www.twilio.com/user/account next to “account sid” and “auth token”
3. Read documentation and code away :)

Twilio example code

from: https://www.twilio.com/docs/api/rest/sending-messages

Replace account sid and auth_token with your account_sid and auth_token.

Replace “to” with your own number you signed up with

Replace “from” with number twilio gave you

```
from twilio.rest import TwilioRestClient
account_sid = "AC3a8d401672272148f5ec2e9c299f4814"
auth_token  = "insert your token here"
client = TwilioRestClient(account_sid, auth_token)
message = client.messages.create(body="Test!!",
    to="+16503195873",
    from_="+14083296380")
```

How to API: Zillow API

1. Register to get a key at https://www.zillow.com/webservice/Registration.htm
2. Read documentation at http://www.zillow.com/howto/api/GetSearchResults.htm and start coding!

Zillow example code

```
html = urllib2.urlopen('http://www.zillow.com/webservice/GetSearchResults.htm?zws-id=<insert_zwsid_here>&address=2114+Bigelow+Ave&citystatezip=Seattle%2C+WA').read()
print html
```

Like scraping, but info returned directly to you

All other APIs should be similar

Create API Key, read documentation, and write code! Simple as that :)

Scraping

- Scraping is going to be a little more complicated.
- BUT you've learned everything you need to know to scrape anything
  - You Just need practice, practice, practice… http://norvig.com/21-days.html
- Start by finding PATTERNS
- Remember string manipulation (string.split(' '), string.find ('find_this'), string slicing)
- Sometimes helpful to look at html lines one by one:

```
for line in html.split('\n'):
    # do something with line here
```

Meet with each of you individually to talk about using scraping in your projects