GENERAL ASSEMBLY

DATA SCIENCE
CLASS 2: GETTING DATA

I. GETTING DATA
II. REGEX / REQUESTS
III. API / WRAPPERS
IV. ETHICS

I. GETTING DATA

Data lives all over the internet

The question is whether or not the author of the data makes it easy for us to grab it.

We will look at three different ways of getting data

1. Using an HTML Parser
2. Using an API
3. Using an API wrapper

Using an HTML Parser

- Pattern recognition
- Regular Expressions

Using an API

- API Documentation
- JSON vs. XML
  - JSON – Javascript Object Notation (more common)
  - XML – Extensible Markup Language

Using an API wrapper

- Wrapper Documentation
- Usually hosted on Github
- Still will probably use JSON
- Example: https://github.com/tweepy/tweepy

II. REGEX / REQUESTS


WHENEVER I LEARN A NEW SKILL I CONCOCT ELABORATE FANTASY SCENARIOS WHERE IT LETS ME SAVE THE DAY.
OH NO! THE KILLER MUST HAVE FOLLOWED HER ON VACATION!
BUT TO FIND THEM WE'D HAVE TO SEARCH THROUGH 200 MB OF EMAILS LOOKING FOR SOMETHING FORMATTED LIKE AN ADDRESS!
IT'S HOPELESS!
EVERYBODY STAND BACK.
I KNOW REGULAR EXPRESSIONS.
PERL!
TAP TAP

REGular EXpressions

are how we capture patterns in text

Google AAPL

Finance **Apple Inc.** (NASDAQ:AAPL) Add to portfolio

Company
- Summary
- News
- Option chain
- Related companies
- Historical prices
- Financials

Markets

112.82

+1.04 (0.93%)

Real-time: 3:18PM EST

NASDAQ real-time data - Discla

Currency in USD

Range	111.97 - 113.49	Div/yield	0.47/1.67
	19.75	EPS	6.43
	12.16	Shares	5.86B
	.32M	Beta	0.92
	.48B	Inst. own	62%
	17.54		

g+1 7.6k

Dow
Nas
Tech
AAP

Back
Forward
Reload
Save As...
Print...
Translate to English
View Page Source
View Page Info
Buffer This Page
Inspect Element

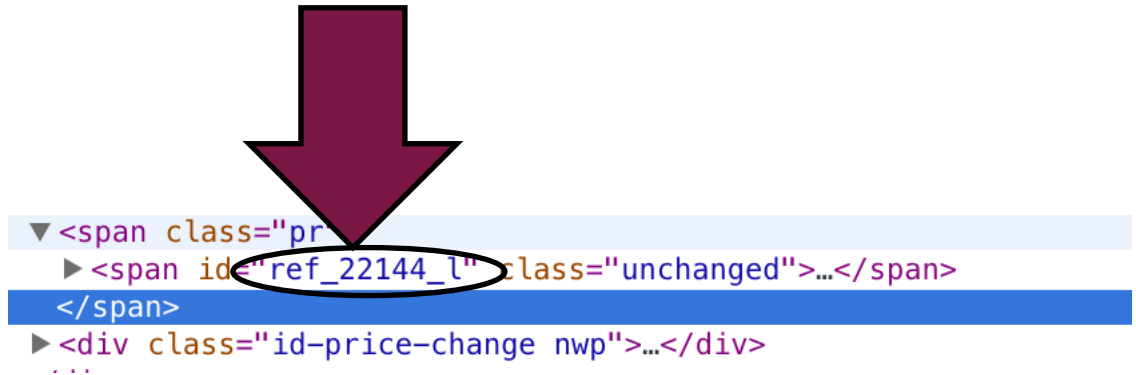
Compare: Enter ticker here es Nasdaq MSFT SNDK SSNNF more »

Elements Network Sources Timeline Profiles

```
▶ <div class="g-section sfe-break-bo
▶ <script>…</script>
▼ <div class="g-section g-tpl-right-
  ▶ <div class="g-unit g-first">…</d
  ▼ <div class="g-unit">
    ▼ <div id="market-data-div" class="id-market-data-div nwp g-floatfix">
      ▼ <div id="price-panel" class="id-price-panel goog-inline-block">
        ▶ <div>…</div>
        ▶ <div>…</div>
      </div>
      ▶ <div class="snap-panel-and-plusone">…</div>
    </div>
    <script>google.finance.renderMarketData();</script>
    ▶ <div id="sharebox-data" itemscope="itemscope" itemtype="http://schema.org/Intangible/FinancialQuote">…</div>
  </div>
```

html body div #fjfe-real-body #fjfe-click-wrapper div #app #gf-viewc div div div div #market-data-div #price-panel div

Console Search Emulation Rendering



What regex can we use to capture this?

BEAUTIFULSOUP

Is a python based HTML parser.

BEAUTIFULSOUP

Is a python based HTML parser.

Let's try it!

WEB CRAWLERS

We just built one!

WEB CRAWLERS

We just built one!

But be careful….

Hacking OKCupid: http://www.wired.com/2014/01/how-to-hack-okcupid/all/

III. APIS AND WRAPPERS

HTML Parsing vs. API

Must call using requests and BeautifulSoup (imitate human behavior)

Makes the call for us (the author is "allowing us" to access the data)

http://www.pythonforbeginners.com/api/list-of-python-apis

API (n):
Application Programming Interface

Easing access into a web based software

Examples of API's:

- Amazon (price data)
- Twitter (tweets)
- Facebook (social network)
- Sentiment Analysis

Examples of API's:

- Amazon (price data)
- Twitter (tweets)
- Facebook (social network)
- Sentiment Analysis

Mashape.com has an extensive collection

API vs. API wrapper

May still be a bit confusing how to call the right page

Puts the API into a specific programming language. Gives us python functions.

http://www.pythonforbeginners.com/api/list-of-python-apis

IV. ETHICS

Facebook Experiment vs. Dunkin Donuts

Facebook running psychological experiments on us

Dunkin Donuts offering promos to areas with negative tweets

Conclusion

Data is all over the web, but we must be polite and conscious of what data is available to us.