CS 2316 Data Manipulation for Engineers

Web Mining

Christopher Simpkins
chris.simpkins@gatech.edu

Web Mining

Two ways to mine data from the web

- The hard way, by web scraping
- The easy way, using web service APIs

We'll see brief examples of both.

The Hard Way: Web Scraping

Web scraping means getting the same HTML meant for a browser to render and scanning the text for snippets of interest. Here we get the page with most popular videos and find video titles using a regular expression.

```
from urllib.request import urlopen
import re

# Most Popular
url =
    "http://www.youtube.com/playlist?list=PLrEnWoR732-BHrPp_Pm8_VleD68f9s1
response = urlopen(url)
contents = response.read()
text = contents.decode('utf8')
for title in re.findall(r'<a
    class="pl-video-title-link.+?">(.+?)</a>', text, re.S):
    print(title.strip())
```

YouTube Most Popular

Here are the results of running youtube_hard.py:

```
$ python3 youtube_hard.py | head
Last Week Tonight with John Oliver: Wealth Gap (HBO)
The official full length TV launch trailer - Doctor Who Series 8 2014
    - BBC One
Andrew Wiggins Ridiculous Pregame Dunk
Old Spice | Soccer
A sudden hail storm in Novosibirsk (Russia) 12.07.2014 |
10 Deadliest Wars in History
7 Unbelievable Trick Shot Videos
Argentina Fans Sad but Proud in Loss
Moore&#39;s Law and The Secret World Of Ones And Zeroes
-Maru is sleeping in the box.-
```

Why Web Scraping is Hard

- Have to read through messy HTML to figure out how to scrape. Consider:

```
 </span><img src="http://s.ytimg.com/yts/img/pixel-vfl3z5WfW.gif"
    alt="" class="yt-uix-button-arrow yt-sprite"></button>
</a></td>  <td class="pl-video-title">
    <a class="pl-video-title-link yt-uix-tile-link
    yt-uix-sessionlink  spf-link " dir="ltr"
    href="/watch?v=vd7U3OYziHY&amp;list=PLrEnWoR732-BHrPp_Pm8_VleD68f
    data-sessionlink="ei=vefEU-F40dOoBciggvgL&amp;feature=plpp_video"
      Should You Post A Selfie?
    </a>
  </td>
```

- Have to fiddle with the regex to get it right.
- If web site changes HTML output, your web scraper breaks.
- Most HTML is UI markup, not semantic markup. So you often don't find html elements like `<li class="video-title">Some Title</li>`, it may be more like `<li class="red underline`

The Easy Way: Web Service APIs

The easy way to mine the web is using APIs. Here we use Google's feeds API that doesn't require a key.

```
from urllib.request import urlopen
import xml.etree.ElementTree as et

url = "https://gdata.youtube.com/feeds/api/standardfeeds/top_rated"
response = urlopen(url)
contents = response.read()
text = contents.decode('utf8')
root = et.fromstring(text)
for vid in root.findall('{http://www.w3.org/2005/Atom}entry'):
    print(vid.find('{http://www.w3.org/2005/Atom}title').text)
```

This URL returns an XML document that we can parse just like we have before ... almost. The weird tag name includes a namespace. ElementTree doesn't play well with namespaces.

We could also get the result as JSON.

YouTube Top Rated

Here are the results of running youtube_easy.py:

```
$ python3 youtube_easy.py
Charlie bit my finger - again !
Evolution of Dance - By Judson Laipply
Justin Bieber - Baby ft. Ludacris
Cut Chemist feat. Hymnal "What's the Altitude" Music Video
Nicki Minaj - Super Bass
Miley Cyrus - 7 Things
LMFAO - Party Rock Anthem ft. Lauren Bennett, GoonRock
Timbaland - Apologize (feat. One Republic)
Potter Puppet Pals: The Mysterious Ticking Noise
Lady Gaga - Bad Romance
Chris Brown - Look At Me Now ft. Lil Wayne, Busta Rhymes
The Sneezing Baby Panda
Jonas Brothers - SOS Music Video - Official (HQ)
Vanessa Hudgens Say Ok Music Video (Official with Zac Efron)
Basshunter : Now You're Gone
"Britains Got Talent or Americas Got Talent Connie Talbot WOWs Simon
    Cowell !"
My Chemical Romance - "Teenagers" [Official Music Video]
Jennifer Lopez - On The Floor ft. Pitbull
The Rej3ctz - Cat Daddy (Starring Chris Brown)
```

Why Web Services are Easier

- Using a defined API that the service provider commits to
- Result is in a format designed for carrying data (XML, JSON, etc), as opposed to visual markup
- Web service APIs are controlled by the service provider, so if you use within their terms your use is ethical

Closing Thoughts

Lots more to web mining. For real-world mining and scraping, have a look at

- Beautiful Soup
- Requests

Be sure to read DMSI: Reading The Web