A LAZY PROGRAMMER'S GUIDE TO WEB SCRAPING

(WEB SCRAPING USING PYTHON)

By : Pradhvan Bisht

User.py

```
def Noob_user(Name,nick,Status):
    print( "Name:",  Name)
    print ("Internet_nick :",  nick)
    print ("Status: ",  Status)
    return;

Noob_user(Name = "Pradhvan Bisht ", nick = “Pradhvan",
Status = "Student")
```

Prerequisite

- Familiarity with python language.
- Knowledge of HTML and CSS.
- Understanding of HTTP mainly GET and POST.

HTTP - The what ? The How ?

What is web scraping ?

Web scraping is a technique to extract large amounts of data from websites whereby the data is extracted and saved to a local file in your computer.

The data can be used for several purposes like displaying on your own website and application, performing data analysis or for any other reason.


DATA ! DATA ! EVERYWHERE

Why should you scrape ?

- API may not provide what you need
- No rate limit
- Take what you really want!
- Reduces manual effort
- Swag!

Things that might come handy

-Python libraries

-XPATH

-Regular Expressions

-Selenium Webdriver

How it's done ?

Broadly a Three Step Process

1. Getting the content (in most cases HTML)
2. Parsing the response.
3. Optimizing/Improving the performance and preserving the data

GETTING THE CONTENT

- Getting the URL.
- Using HTTP libraries for making a request
- Involves GET/POST request to the server.
- The response contains the information to be extracted.
- Sometimes not as easy as it may seem.

Extracting The Data - The fun part

1. Using Regular Expression and Basic python

   Tricky, complex and kind of fragile.

2. Using Parsing Libraries

   - Two different approaches possible -- Simple Parsing and Search Tree parsing.
   - Some popular libraries are BeautifulSoup and Lxml.
   - Each modules has its own techniques and thus its own pros and trade-offs

WHAM ! EXTRACTING DATA WITHOUT AN API

#SWAG

Comparing Parsers

BEAUTIFUL SOUP

Beautiful Soup is a Python library for pulling data out of HTML and XML files.

LXML

LXML is a reference to the XML toolkit n a pythonic way which provides bindings for C library libxml2, and libxslt without sacrificing speed.

BEAUTIFUL SOUP

- Easy to get started with
- Can handle broken markup
- Purely in python
- Slow

LXML

- Very Fast
- Not purely in Python
- Works great with large project

Preserving The Data

1. Writing to a file.

2. Exporting as csv or excel file.

3. Storing in a database.

Example

Example 1 : Using **Selenium** to login and fetch job listings from HasJob which uses Dynamic HTML.

Let’s talk CODE !


TOO MANY WAYS
WHAT TO USE WHEN ?
imgflip.com

WHAT TO USE WHERE

1. **Handling dynamically generated html**

Solutions: Selenium

2. **Cookie based Authentication**

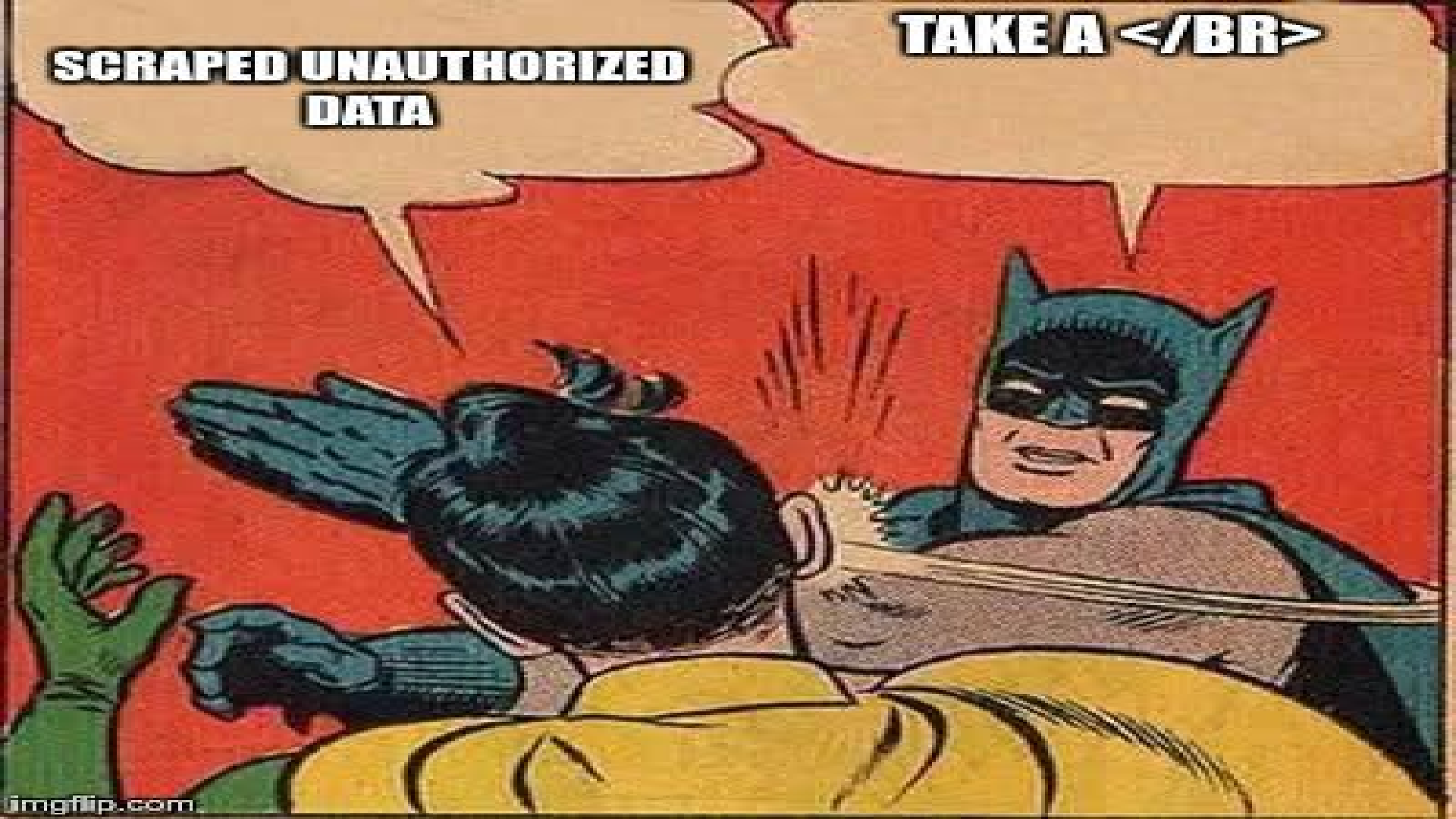
Solution : Requests module.

3. **Simple scraping**

Solutions: BeautifulSoup+Requests/Urllib, Selenium

Scraping hacks

1. Overcoming captchas
2. Per IP address query limit
3. Write tests


SCRAPED UNAUTHORIZED DATA
TAKE A </BR>
imgflip.com

Ethics Of Scraping

Exceeding authorized use of the site

Means doing anything that is prohibited in the Terms of Use (See CFAA, breach of contract, unjust enrichment, trespass to chattels, and various state laws similar to CFAA)

Copyright Issues

If the material you are scraping is not factual, but something that required some amount of creativity to create, you have copyright to worry about.

QuickTip -- Conform to the the ***robots.txt*** file.

What Next ?

Massive Scraping - SCRAPY

Find the right tool and write tests.

Practice ! Practice ! Practice !


ALWAYS EASY THE BRUTE-FORCE
WAY IS NOT. YEESSSSSSS.

THANK YOU

@bishtpradhvan

http://bit.ly/Pyconf-hyd