# Web Scraping by Python feat. Beautiful Soup

Marcos R. Pesante Colón ACM – CSE Chapter August 8, 2020

## Web Scraping

Introduction & Applications

### Web Scraping

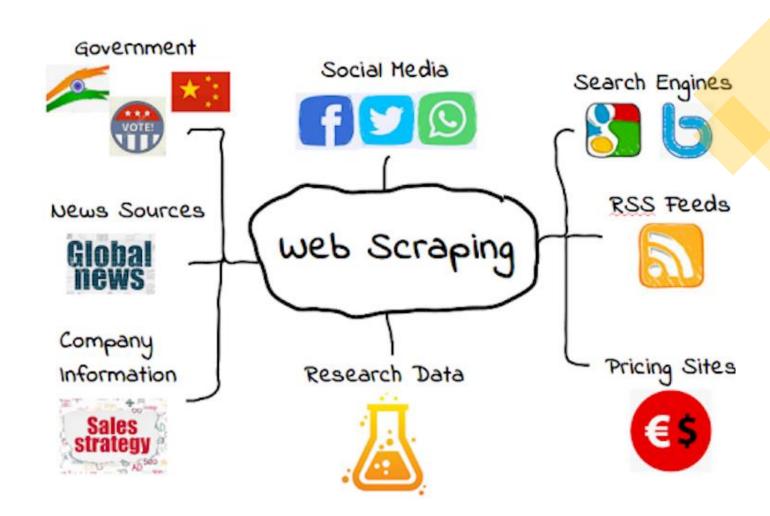
# The extraction and copying of data from a website into a structured format using a computer program

https://www.dictionary.com/browse/web-scraping

\* Useful for when you want to automate data extraction from a website and they do not provide APIs or other services with this object

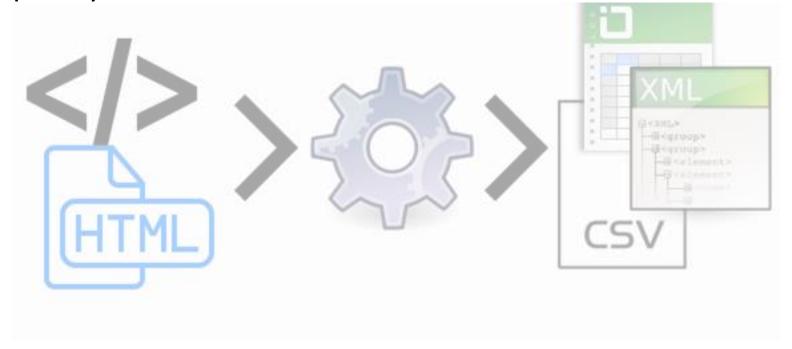
#### **Applications**

- Google Web Crawling to find new websites
- Social Media Sentiment Analysis
- Price monitoring
- Obtain Data to feed Machine Learning Models



### Before Beginning

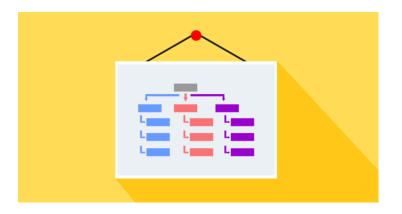
- Establishing goals
  - 1. What information do you want to extract?
  - 2. Who is going to use this data and how do they want it to be stored?
  - 3. How frequently will the data be "reset"?



### Before Beginning

- robots.txt
  - List of restricted parts of the website
  - You cannot scrape these parts of the website
    - https://moz.com/learn/seo/robotstxt
- sitemap.xml
  - List of most, if not all, of the links in a website
  - Sometimes can even refer to even more sitemaps





 $\mathsf{HTML}$ 

Introduction / Review

#### HTML

- Hyper Text Markup Language
- Standard language to create and **structure** Web Pages
- Accomplishes this through the use of nested tags
  - <h1></h1>
  - •
  - <img>
  - <a href="#"></a>
  - <div></div>





#### Example

```
<!DOCTYPE html>
<html lang="en">
<head>
 <meta charset="UTF-8">
 <title></title>
</head>
<body>
 <h1 id="header1">This is a Heading</h1>
 <div class="card">
   Paragraph 1
   Paragraph 2
   </div>
 <div class="card">
   Paragraph 3
   Paragraph 4
   </div>
</body>
</html>
```

### This is a Heading

Paragraph 1

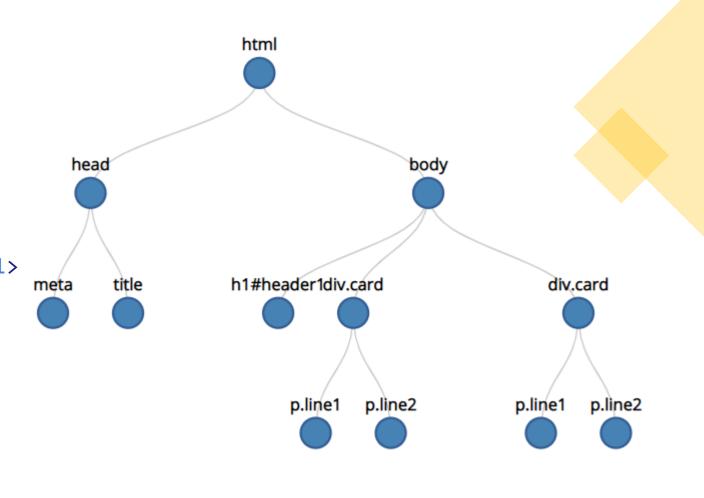
Paragraph 2

Paragraph 3

Paragraph 4

#### Example

```
<!DOCTYPE html>
<html lang="en">
<head>
 <meta charset="UTF-8">
 <title></title>
</head>
<body>
 <h1 id="header1">This is a Heading</h1>
 <div class="card">
   Paragraph 1
   Paragraph 2
   </div>
 <div class="card">
   Paragraph 3
   Paragraph 4
   </div>
</body>
</html>
```





## Being picky about our HTML

**CSS Selectors** 

#### **CSS Selectors**

- While HTML only indicates structure of a website's content, CSS decorates and styles it through assignment of colors and animations to specific elements
- Elements must be selected through the use of *ids*, *classes*, and other criteria such as attributes



CSS Selectors Reference

w3schools.com

#### Examples

<u>Duck Duck Go</u> -Select all search titles using CSS



• <u>Merriam-Webster</u> - Select all definitions using CSS







Using Requests and Beautiful Soup





#### Links to Images Used

- https://qph.fs.quoracdn.net/main-qimg-7dffea772ea85c4918a186073201e4ea
- https://roboticsandautomationnews.com/wp-content/uploads/2020/04/web-scraping-2.png
- https://www.antevenio.com/usa/wp-content/uploads/2019/12/web-scraping-service.png
- https://cdn.searchenginejournal.com/wp-content/uploads/2019/09/7-reasons-why-an-html-sitemap-is-a-must-have-1520x800.png
- https://sitechecker.pro/wp-content/uploads/2017/12/robots.txt.png
- https://upload.wikimedia.org/wikipedia/commons/thumb/6/61/HTML5\_logo\_and\_wordmark.svg/512px-HTML5\_logo\_and\_wordmark.svg.png
- https://blog.codepen.io/wp-content/uploads/2012/06/Button-Fill-Black-Large.png
- https://upload.wikimedia.org/wikipedia/commons/thumb/3/3e/W3Schools\_logo.png/800px-W3Schools\_logo.png
- https://upload.wikimedia.org/wikipedia/commons/thumb/3/32/Merriam-Webster\_logo.svg/1024px-Merriam-Webster\_logo.svg.png
- https://upload.wikimedia.org/wikipedia/en/9/90/The\_DuckDuckGo\_Duck.png
- https://upload.wikimedia.org/wikipedia/commons/a/aa/Requests\_Python\_Logo.png
- https://funthon.files.wordpress.com/2017/05/bs.png
- https://upload.wikimedia.org/wikipedia/commons/thumb/c/c3/Python-logo-notext.svg/600px-Python-logo-notext.svg.png