

# ***Web Scraping***

Turning Websites Into Data



# Objectives

1. **Parse HTML and CSS elements in webpages**
2. **Use requests and BeautifulSoup to get and process webpage contents**
3. **Use ethics when scraping websites**

# HyperText **Markup** Language (HTML)

If the HTTP response can only contain strings, how does the browser know how to display a website?

## **Markup!**

- Everything the browser needs to know is embedded into one big string
- The string is structured with a hierarchy of **tags** that represent each component to be rendered

## HTML

```
<!DOCTYPE html>
<html>
<head>
<link rel="stylesheet" type="text/css" href="styles.css" />
</head>

<body>

<h1>This is a header</h1>

<p>Some text for my paragraph</p>

</body>
</html>
```

## CSS

```
h1 {
  font-size: 20pt;
  color: red;
}

p {
  font-size: 12pt;
}
```

## WEB BROWSER

This is a header

Some text for my paragraph

# Why Scrape?




ENTIRE GUESTHOUSE · 1 BED

**Small Bungalow w/Private Entrance**

\$75 per night · Free cancellation

★★★★★ 363 · Superhost

Government of Canada

Gouvernement du Canada

Jobs ▾

Immigration ▾

Travel ▾

Business ▾

Benefits ▾

Health ▾

Taxes ▾

More services ▾

Home → [Import, Export and Investment](#) → [Trade data online](#)

## Report - Trade Data Online

[Help](#) | [Return to Trade Data Online](#)

Report date: 2020-02-05

### Criteria

|             |                                       |
|-------------|---------------------------------------|
| Title       | Canadian total exports                |
| Industries  | Naics 11111 - soybean farming         |
| Origin      | Canada                                |
| Destination | All countries (total)                 |
| Period      | Latest 5 years                        |
| Units       | Value in millions of canadian dollars |

Change criteria

### Report

|                       | 2014  | 2015  | 2016  | 2017  | 2018  |
|-----------------------|-------|-------|-------|-------|-------|
| All Countries (Total) | 1,987 | 2,359 | 2,542 | 2,499 | 2,889 |

Data Source: Statistics Canada

[Share this page](#)

Save report as CSV

Save report as Excel

BarkBox Small Rainfurrest Dog Toy & Treat Bundle Assortment - Plush Toys, Chew Toys, Squeak Toys, All-Natural Treats/Chews Made in The USA

★★★★☆ ▾ 234

\$31<sup>86</sup>

✓prime

Get it by **TODAY, Feb 21**

FREE Shipping on eligible orders

Dog Chew Rope Toys Knotted Clean Teeth Cotton for Aggressive Chewers Pack of 3 (Blue-White)

★★★★☆ ▾ 57

\$9<sup>99</sup>

✓prime

Get it by **Sat, Feb 23**

FREE Shipping on eligible orders



| price | title   |
|-------|---|
| 31.86 | BarkBox Small Rainfurrest Dog Toy & Treat Bund... |
| 9.99  | Dog Chew Rope Toys Knotted Clean Teeth Cotton ... |

# ***Why Not Scrape?***



**Downloadable  
Dataset**



**Database  
Connection**



**Public  
REST API**

## ***Ethical Considerations***

- Terms of Service
- Denial of Service Attacks
- Confidentiality

[This article](#) discusses legal issues related to web scraping

*We are not lawyers - this does not constitute legal advice.*



# ***Python Tools for Web Scraping***

| Name          | Fetch HTML | Parse HTML | Notes                        |
|---------------|------------|------------|------------------------------|
| urllib        | X          |            | Python standard library      |
| requests      | X          |            | Lightweight, fast, easy      |
| BeautifulSoup |            | X          | Fast, can parse any XML file |
| scrapy        | X          | X          | Complex, powerful            |
| selenium      | X          | X          | Automates a real web browser |