# Web Scraping

**Turning Websites Into Data** 



# **Objectives**

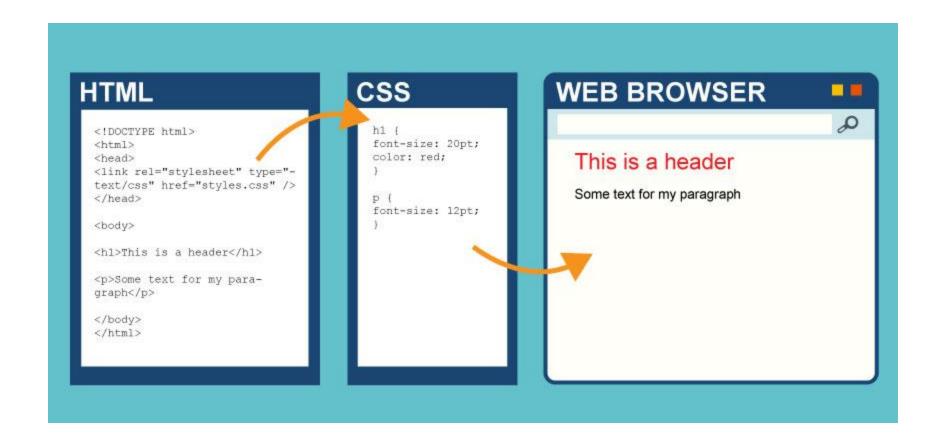
- Parse HTML and CSS elements in webpages
- 2. Use requests and BeautifulSoup to get and process webpage contents
- 3. Use ethics when scraping websites

## 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



#### Why Scrape?

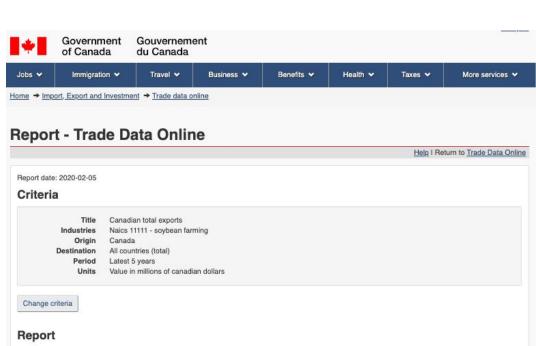


**ENTIRE GUESTHOUSE · 1 BED** 

Small Bungalow w/Private Entrance

\$75 per night · Free cancellation

\*\*\*\* 363 · Superhost



11	2014	2015	2016	2017	2018
All Countries (Total)	1,987	2,359	2,542	2,499	2,889

Data Source: Statistics Canada

C 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** 

\$3186

√prime

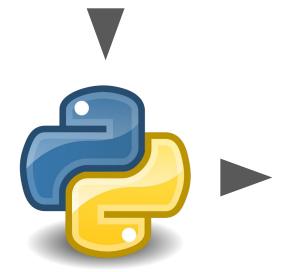
Get it by TODAY, Feb 21 FREE Shipping on eligible orders Dog Chew Rope Toys Knotted Clean Teeth Cotton for Aggressive Chewers Pack of 3 (Blue-White)

★★☆☆☆ ▼ 57

\$999

√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?





**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