

Northeastern University

CS6020: Collecting, Storing, and Retrieving Information

Data Collection Through Web Scraping

Data Collection Through Web Scraping

WEB SCRAPING CONCEPTS

Lesson Objectives

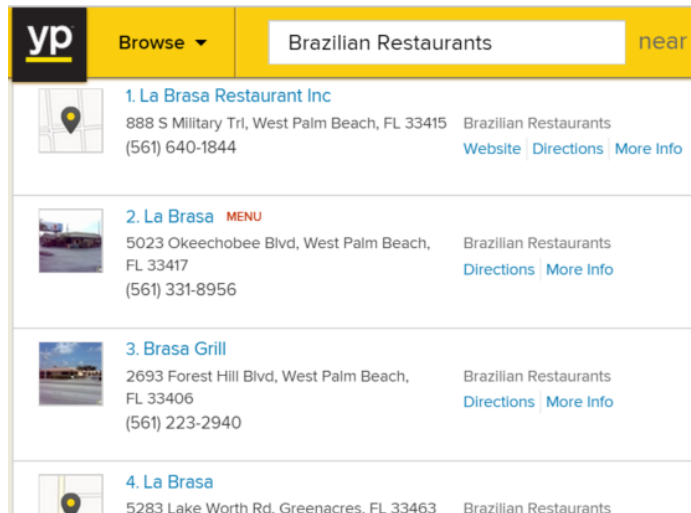
- After completing this lesson, you are able to:
 - know when web scraping is appropriate
 - describe the process of web scraping
 - list web scraping toolkits

Web Scraping

- As the Internet continues to grow the amount of data available has increased substantially.
- However, availability of data does not translate to accessibility.
- **Web scraping** (or **web harvesting** or **web data extraction**) extracts data from websites when the data is not available in text file format, such as CSV.

Example: Scraping YellowPages

- For example, [YellowPages](#) does not make its data available as a file or a Web API, so extracting or “scraping” the data from their web sites is required.



To extract a list of the restaurants use a GET request to specify the search and then “parse” or “scrape” the HTML page that is returned. This would be done programmatically.

http://www.yellowpages.com/search?search_terms=Brazilian+Restaurants&geo_location_terms=West+Palm+Beach%2C+FL

Web Scraping

- While data could be manually copied from web sites, it would be too tedious and time consuming for anything but very small amounts of data.
- “**Web Scraping**” automates this process, so that instead of manually copying the data from websites, a program does the copying.

Web Scraping Process

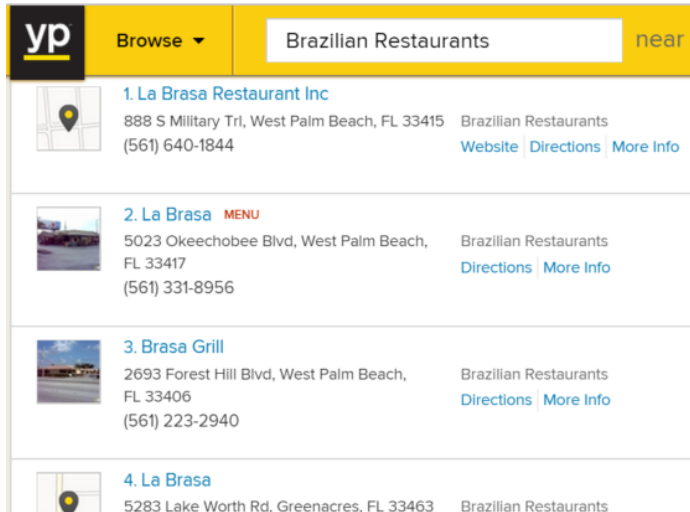
- Web Scraping essentially transforms the raw unstructured data in HTML into structured data that can be cleaned, stored, shaped, and used in analysis.



Web Scraping Toolkits

- Instead of building a custom program to scrape data, there are toolkits that simplify a web scraping effort:
 - Kimono
 - import.io
 - 80legs
 - scrapy

Example of a Web Scrape



yp Browse ▾ Brazilian Restaurants near

1. La Brasa Restaurant Inc
888 S Military Trl, West Palm Beach, FL 33415 Brazilian Restaurants
(561) 640-1844 Website Directions More Info

2. La Brasa MENU
5023 Okeechobee Blvd, West Palm Beach, FL 33417 Brazilian Restaurants
(561) 331-8956 Directions More Info

3. Brasa Grill
2693 Forest Hill Blvd, West Palm Beach, FL 33406 Brazilian Restaurants
(561) 223-2940 Directions More Info

4. La Brasa
5283 Lake Worth Rd, Greenacres, FL 33463 Brazilian Restaurants

import.io took the URL of the web page and transformed it into downloaded text data.

http://www.yellowpages.com/search?search_terms= GET DATA Examples

#	n_link	directions_link	categories_li...	link_1
1	1. La Bras...	Directions	Brazilian Restaurants	Website
2	2. La Brasa	Directions	Brazilian Restaurants	Directions
3	3. Brasa G...	Directions	Brazilian Restaurants	Directions
4	4. La Brasa	Directions	Brazilian Restaurants	Directions
5	5. Texas d...	Directions	Restaurants; Brazilian Restaurants	Website

COPY TABLE DOWNLOAD... GET API...

Legal Issues

- Some websites do not allow web scraping as part of their terms of use.
- Additionally, the scraped data may be private or copyrighted and may be prohibited from being copied or used for commercial purposes.

Web Scraping Strategies

- Current web scraping approaches range from ad-hoc, manual scraping, to fully automated parsing.
 - Manual Copy-and-Paste
 - Regular Expression Matching
 - HTTP Retrieval with HTML Parsing
 - DOM Parsing
 - Web Scraping Toolkits
 - Vertical Integration Platforms
 - Metadata & Semantic Markup Recognition
 - Machine Learning Based Visual Scanning

Summary

- In this lesson, you learned that:
 - web scraping is used to extract data from the web that is not available in any other format
 - web scraping requires parsing HTML
 - web scraping parsing fails if the web page structure changes substantially
 - there are many web scraping toolkits available that simplify the process of scraping data from web sites and transform them into structured data



Summary, Review, & Questions...