Group 2: Web Scraping

MIE1624 Intro to Data Analytics
January 23, 2017

Mohammed Bubshait
John de Vera
Abdulrahman Fnais
Madeh Piryonesi
Jose Vera Aray

Agenda

1. Introduction
    a. Problem overview
    b. What is HTML and Web Scraping ?
2. Python packages
    a. Python packages overview
    b. Data Scraping with Beautiful Soup
3. Program walkthrough
4. Real world applications
5. Lessons learned
6. Q/A

Problem Overview

Websites with HTML Pages

Web Scraping Technology

Structured Data

http://webdata-scraping.com/ http://blog.myetv.tv/

What is HTML?

HTML (Hyper Text Markup Language) is the standard markup language for creating Web pages

What is Web Scraping?

An automatic software technique for extracting information

Identify Project Information Requirements

Investigate Web pages Structure

https://www.webharvy.com/articles/what-is-web-scraping.html

Python Packages

Python Packages

- Beautiful Soup
  - Python library for extracting HTML or XML data
  - BeautifulSoup functions helps to select common HTML elements
  - https://www.crummy.com/software/BeautifulSoup/bs4/doc/
- URLLIB2
  - Python library for opening URLs
  - https://docs.python.org/2/library/urllib2.html
- CSV
  - Comma Separated Value, a format used for representing spreadsheets and databases
  - https://docs.python.org/2/library/csv.html

Data Scraping w/Beautiful Soup Python Library

amazon.com

Term	Purpose	Soup Analogy
Amazon.com	Main website	Campbell
Amazon.com Books	Website focus	Campbell's Alphabet Soup
URLLIB2	Open URL	Can opener
Beautiful Soup	HTML data file	Soup in a bowl
Parser	Extract data (LXML/HTML)	Spoon / Fork
Objective: Why did we choose this?	Display top 100 books	Display alphabet (A, B, C..) in order

Program Walkthrough

Program Walkthrough

General Outline

1. Read URL of parent web page and store it.
2. From this web page, get the hyperlinks of all child webpages where the items are displayed
3. Initialize empty dictionary for storing data
4. For each web page, read and store it
    a. Select all relevant items (By looking the HTML code, the items I'm interested are in a single "div" element of class= "zg_itemImmersion")
    b. For each item we extract the data we are interested and store it in the dictionary
5. Write CSV file from the data stored in the dictionary

Real World Applications

Real World Applications

Demand Analysis

Pricing Strategy

e-commerce

Meta-Search Engine

tripadvisor
trivago
hotelscombined
KAYAK

Campaign Monitoring

For Real Estate

- Listings Gathering
- Agent contact details

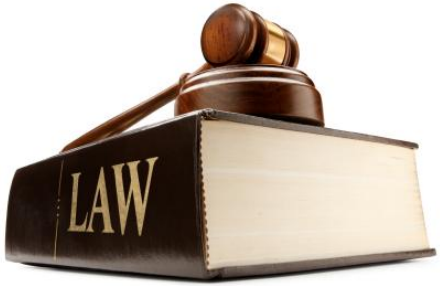
For Marketing

Gather contact details of Businesses and individual

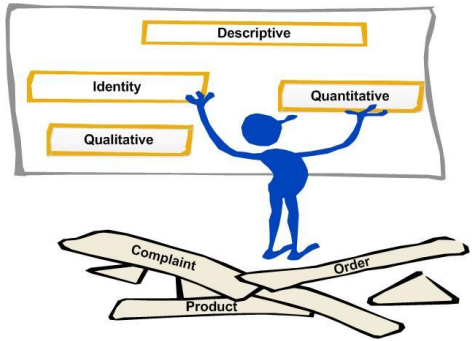
Lessons Learned

Lessons learned

- Make sure you read website's terms of use

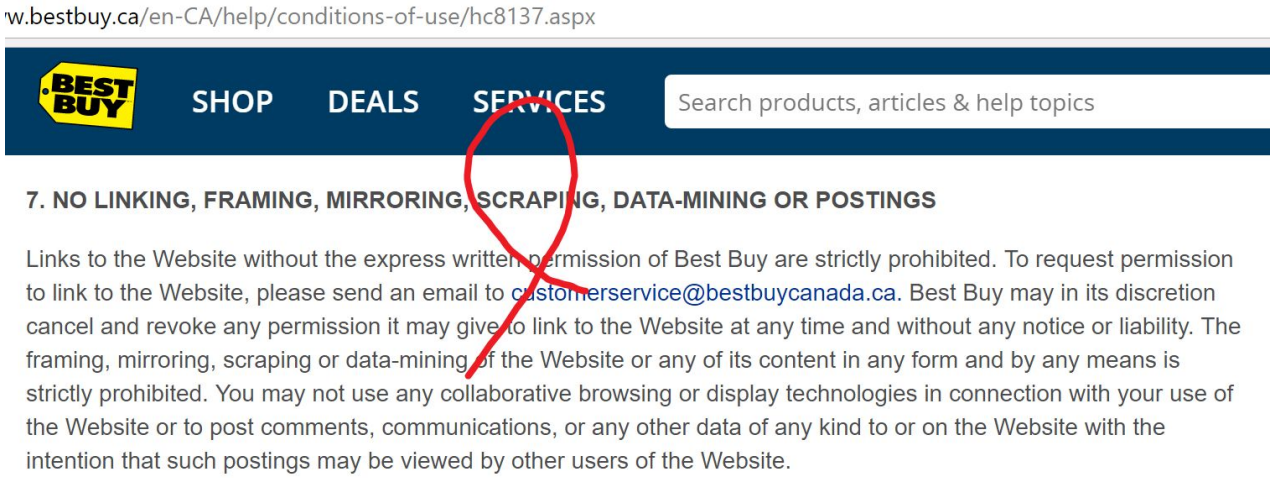


- Understand your software requirements
- Test different conditions, don’t just assume since you program runs, its producing the right results



Lessons learned

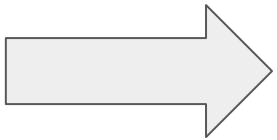
- Legal issues
- Different countries have different regulations.
- Make sure you read ***terms of use***



Lessons learned

- Make sure that you check the data type and format of extracted data
- Extracted data might need need to be transformed

Cost (extracted as string)
$30
$128.99
$25



Cost (cast as float)
30
128.99
25

Thank You

google.com/+ProgrammingCom/posts/Tpj2bD8x7hX