

# Group 2: Web Scrapping

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 BeautifulSoup

## 3. Program walkthrough

## 4. Real world applications

## 5. Lessons learned

## 6. Q/A

# Problem Overview



Websites with HTML  
Pages



Web Scraping  
Technology



Structured Data

1. **Description of Service:** The Service provides consumer content regarding insurance products and services, for informational purposes only. The Service does not sell any type of insurance and is not an insurer or insurance broker, nor does it recommend, support or endorse any particular insurance plan. The Service enables you to request to receive insurance or discount program quotes from a network of insurance companies, agents, brokers, discount program representatives and other providers (the "Insurance Representatives"). Through the Service, you choose to provide information about yourself and your insurance preferences ("User Information") which is in turn used to attempt to match you with Insurance Representatives who may be able to follow-up on your request. If you use the Service, we cannot guarantee that any of the Insurance Representatives to whom we forward your information will contact you or agree to offer you coverage. We also cannot guarantee the carrier affiliation of any Insurance Representative who may contact you. We have no control whatsoever for the conduct of any of the Insurance Representatives who may contact you. We have no control whatsoever for the conduct of any of the Insurance Representatives who may contact you.
2. **Terms of Services**  
Did you ever read them?
3. **Third Party Information and Trademarks:** All content provided on this site about third parties (including companies and brokers) is provided for informational purposes only. Such content is not an endorsement of or a recommendation for any third party. It does not imply, directly or indirectly, any sponsorship or affiliation with such third parties, and no guarantees regarding the same are made herein. All third party trademarks are the property of their respective

# What is HTML?

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

```
▼<div class="footer-columns">
  ▼<div id="foot-column-01" class="foot-column">
    <h3>TESTING 123 123</h3>
  </div>
  <!-- /#foot-colomn-01 -->
  ▶<div id="foot-column-02" class="foot-column">...</div>
  <!-- /#foot-colomn-02 -->
  ▶<div id="foot-column-03" class="foot-column">...</div>
  <!-- /#foot-colomn-03 -->
</div>
<!-- /.footer-columns -->
</div>
```

# What is Web Scrapping?

An automatic software technique for extracting information

Identify  
Project  
Information  
Requirements

Investigate  
Web pages  
Structure



# 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

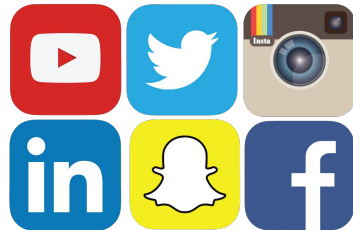


e-commerce

Meta-Search Engine



Campaign Monitoring



For Real Estate



Listings Gathering

Agent contact details

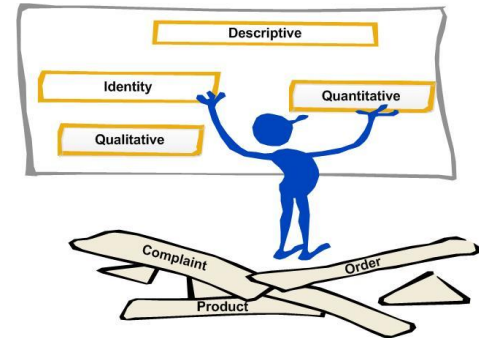
For Marketing

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***

[w.bestbuy.ca/en-CA/help/conditions-of-use/hc8137.aspx](http://w.bestbuy.ca/en-CA/help/conditions-of-use/hc8137.aspx)



## 7. NO LINKING, FRAMING, MIRRORING, SCRAPING, DATA-MINING OR POSTINGS

Links to the Website without the express written permission of Best Buy are strictly prohibited. To request permission to link to the Website, please send an email to [customerservice@bestbuycanada.ca](mailto:customerservice@bestbuycanada.ca). Best Buy may in its discretion cancel and revoke any permission it may give to link to the Website at any time and without any notice or liability. The framing, mirroring, scraping or data-mining of the Website or any of its content in any form and by any means is strictly prohibited. You may not use any collaborative browsing or display technologies in connection with your use of the Website or to post comments, communications, or any other data of any kind to or on the Website with the intention that such postings may be viewed by other users of the Website.

# 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

