Collecting Price Data Through Web Scraping

Majed Skaini
UN-ESCWA
10 June 2021

UNITED NATIONS
الاسكوا
ESCWA

Shared Prosperity **Dignified Life**

What is Web Scraping?

- Web scraping is a process employed to automatically extract large amounts of data from websites, whereby the data extracted is saved to a local file or to a database in table (spreadsheet) format.
- It helps acquire data from multiple sources in a noticeably short period of time.
- It keeps track of any online changes in data.
- It aids in data archiving.

Development of the Idea in the Arab Region

ESCWA aimed to modernize price statistics in the region by introducing technology and Big Data.

We started looking into the application of Big Data tools in data collection, such as web scraping and scanner data.

We developed our newest initiative: the application of web scraping for automatic extraction of price data as an innovative data collection initiative.

Web scraping represents a complementary data collection method to the traditional field collection and does not intend to replace it.

Launching of the Pilot Stage

ESCWA chose 4 pilot countries to launch its initiative: Bahrain, Kuwait, Qatar and Lebanon.

The pilot phase consisted in implementing web scraping techniques for the price collection of Household Consumption Fast Evolving Technology Items.

Training was conducted for Bahrain and Kuwait in January and February 2020 respectively, while virtual sessions were conducted for Qatar and Lebanon in April 2020.

ESCWA developed scrapers for a number of reliable outlets with up-to-date websites in each country.

The initiative was successful and positive feedback was received from all 4 pilot countries.

Following the COVID-19 related lockdown measures, ESCWA conducted a survey for its member states in April 2020 to assess the effect of the pandemic on price statistics.

85% of respondent national statistical offices reported that the pandemic had impacted their statistical work, with the effects mostly felt in data collection.

77% of respondents stated that their traditional field data collection was affected by the pandemic.

The survey revealed the need for implementing alternative data collection methods to sustain price data collection and other statistical programs.

The Impact of COVID-19 on Price Data Collection in the Arab Region and the Need for Alternative Data Sources

The Aftermath

Given the need for implementing alternative data sources in the region in order to sustain price data collection following the pandemic, and due to the success of the pilot stage, ESCWA expanded its web scraping initiative to include all its member countries and the entire household consumption list.

Methodology

- ESCWA is developing scrapers for reliable and large outlets with online presence in each of its member countries for the entire household consumption list.
- We are writing the software from scratch in our office using programming language, developing the scrapers on Python using Anaconda’s Jupyter Notebook.
- Each outlet webpage presents different challenges which we tackle on an individual basis.
- We are currently conducting virtual national trainings on the use of the developed scrapers for each member country at a time.

Methodology (cont'd)

1. After selecting a website, we check its "terms of use" by adding "/robots.txt" to the root URL we intend to scrape to follow ethical standards and avoid legal repercussions.

2. We find the URL we want to scrape – i.e., the page of interest within the main website.

3. We inspect the page.

4. We find the data we wish to extract, for example price, name and additional description.

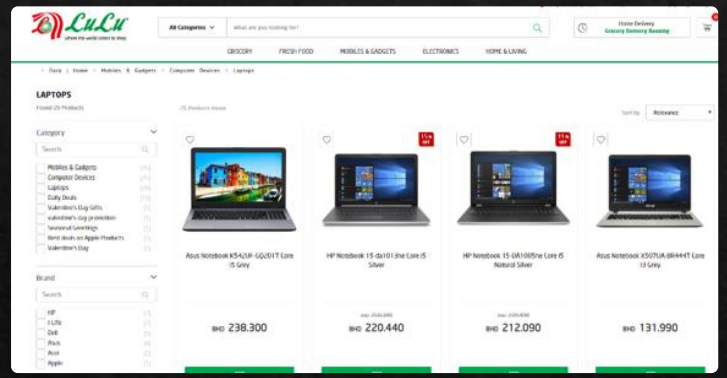
5. We write the code using Python language (other languages can be used).

6. We run the code and extract the data.

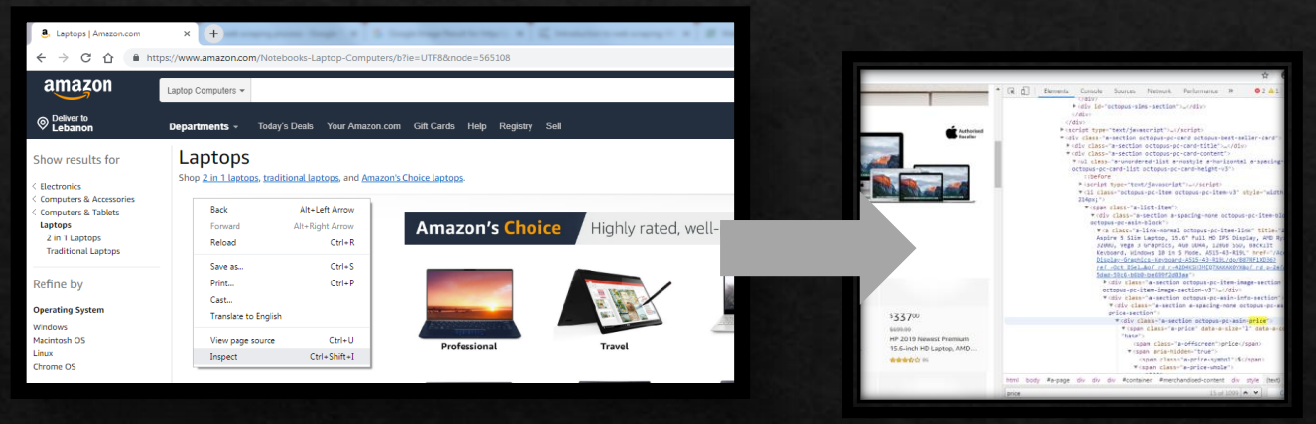
7. We store the data in the required format.

Web Scraping Process

Finding the URL we want to scrape



Inspecting the page



Writing the code

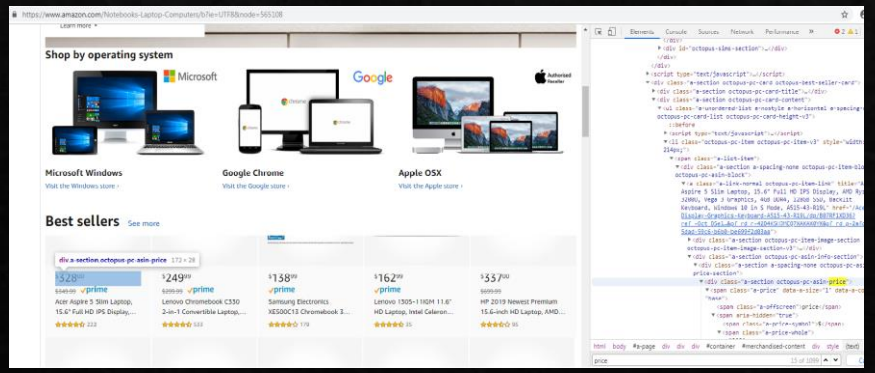
```
rt requests
 bs4 import BeautifulSoup
rt pandas as pd
rt re

scrape():
headers = requests.utils.default_headers()
headers.update({
'User-Agent': 'Mozilla/5.0 (X11; Ubuntu; Linux x86_64; rv:52.0) Gecko/20100101 Firefox/52.0',
})
prod_name_list = []
prod_price_list = []
prod_price_new_list = []
page = requests.get("https://www.luluhypermarket.com/en-bh/laptops-computer-devices-mobiles-gadgets/c/HY00214754
soup = BeautifulSoup(page.text, 'html.parser')
products = soup.find('div',class_='product__listing product__grid col-xs-12')

prod_list = products.find_all('div',class_='product-tile-main')
for prod in prod_list:
    prod_name = prod.find('div',class_='plp-prod-name').text
    price_box = prod.find('div',class_='product-pricing-section')

    try:
        prod_price_new = price_box.find('div',class_='price').text
        prod_price = price_box.find('div',class_='act-price').text
    except:
        prod_price = price_box.find('div',class_='act-price').text
```

Finding the data we want to extract

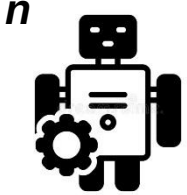


Storing the data

	Name	Initial Price	Final Price
1	Apple MacBook - Core M3 1.2GHz 8GB 256GB Shared Silver Arabic	BHD593.250	BHD349.990
2	Apple MacBook Air (2019) - Core i5 1.6GHz 8GB 128GB Shared 13.3inch Space Grey Arabic	BHD521.745	BHD499.990
3	Apple MacBook Pro 13 with Touch Bar (2019) - Core i5 1.4GHz 8GB 128GB Shared 13.3inch Space Grey Arabic	BHD607.950	BHD569.990
4	Apple MacBook Pro 13 with Touch Bar (2019) - Core i5 2.4GHz 8GB 512GB Shared 13.3inch Space Grey Arabic		BHD917.700
5	Apple MacBook Pro 13 with Touch Bar (2019) - Core i5 1.4GHz 8GB 256GB Shared 13.3inch Space Grey Arabic		BHD702.450
6	Apple MacBook Pro 13 with Touch Bar (2019) - Core i5 2.4GHz 8GB 256GB Shared 13.3inch Space Grey Arabic		BHD823.200
7	Apple MacBook Pro 13 with Touch Bar (2019) - Core i5 2.4GHz 8GB 512GB Shared 13.3inch Space Grey Arabic		BHD917.700
8	Apple MacBook Pro 13 with Touch Bar (2019) - Core i5 2.4GHz 8GB 256GB Shared 13.3inch Space Grey Arabic		BHD823.200
9	Dell Inspiron 15 5584 Laptop - Corei7 1.8GHz 8GB 1TB+128GB 4GB 15.6inch FHD Silver	BHD383.824	BHD362.824
10	Dell Inspiron 15 3581 Laptop - Core i3 2.3GHz 4GB 1TB Shared Win10 15.6inch FHD Black		BHD194.562
11	Dell Inspiron 14 3480 Laptop - Core i5 1.6GHz 4GB 1TB 2GB Win10 14inch HD Silver		BHD237.822
12	Dell G3 15 Gaming Laptop - Core i7 2.6GHz 16GB 1TB+256GB 4GB Win10 15.6inch FHD Black	BHD488.297	BHD472.990
13	Dell G5 15 Gaming Laptop - Core i7 2.6GHz 16GB 1TB+256GB 6GB Win10 15.6inch FHD Black		BHD633.218
14	Lenovo 300e Convertible Touch Laptop - Celeron 1.1GHz 4GB 128GB Shared Win10 11.6inch HD Black		BHD159.990
15	Lenovo Ideapad 330 Laptop - Celeron 1.8GHz 4GB 500GB Shared Win10 15.6inch HD Onyx Black		BHD109.990
16	Lenovo Ideapad 130 Laptop - Core i3 2.3GHz 4GB 1TB Shared Win10 15.6inch HD Black		BHD139.990
17	Lenovo Ideapad L340 Gaming Laptop - Core i7 2.6GHz 16GB 1TB+128GB 4GB Win10 15.6inch FHD Black		BHD429.990
18	Lenovo Legion Y540 Gaming Laptop - Core i7 2.6GHz 16GB 1TB+256GB 6GB Win10 15.6inch FHD Black		BHD589.990

Benefits

Automation

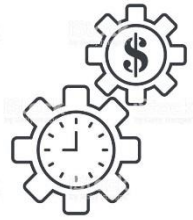


Web Scraping extracts information automatically, instead of looking for prices and pasting them manually into spreadsheets

Accuracy

Data is accurately extracted from websites into spreadsheets eliminating any error in comparison to traditional collection process

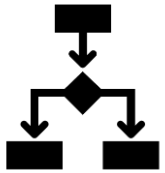
Efficiency



Web scraping is less resource-intensive than traditional collection

Benefits (cont'd)

Multi-purpose



Web scraping can be applied for both CPI and ICP price data collection

Frequency

Web scraping enables more frequent data collection

Validation

Web scraping can be used as an additional validation tool for revision of field-collected data

Challenges and Solutions

Web designs are always evolving, making it harder to label data for scraping

- **Customize and update the script to handle complex web designs**

Web scrapers can be blocked when scraping from the same **IP address**

- **Contact websites to gain permission for web scraping**

Certain websites may prohibit data extraction or use of bots fearing increase in traffic

- **Contact websites and inform them of the actual intentions to avoid legal issues**

Websites that use **CAPTCHA** or similar anti-bot systems

- **Unfortunately, no procedure is set to solve such issue**

Thank you!