UNIVERSITY OF WARSAW

2400-DS1CA /WEB&SOCIAL MEDIA SCRAPING

2020-2021 SUMMER SEMESTER

TERM PROJECT: an online market website scraping

By using BeautifulSoup-Scrapy-Selenium

Submitted by:

Jabir Kangarli

Evrim Bilgen

Yenish Nurmuhammedov

Principal Investigators: Anna Lewczuk & Przemyslaw Kurek

1. **Introduction & Description**

Participants:

Evrim Bilgen – 434281

Jabir Kangarli – 428088

Yenish Nurmuhammedov - 381864

Short description of website: ShopClues.com is India's first online Managed Marketplace that connects buyers and sellers online and offers a trusted and safe online shopping environment.

Short description of the topic:

Before we start, we basically decided to;

- What domain we want to **start** scraping
- What domains you want to **allow** to scrape
- What information you want to **get** from each page

And after then for a basic market analysis, we choose main product categories (we focused on most selling product categories) which are:

- ***Mobile phone***
- ***Gaming consoles***
- ***Headphones***

In these categories, we can get our outcomes

(name-price-discount ratio-image)

and using this result as a proper format (eg. Csv) we can interpret the result.

Our scraping methods:

1- Jabir- Beautiful Soup

2- Evrim-Scrapy

3- Yenish-Selenium

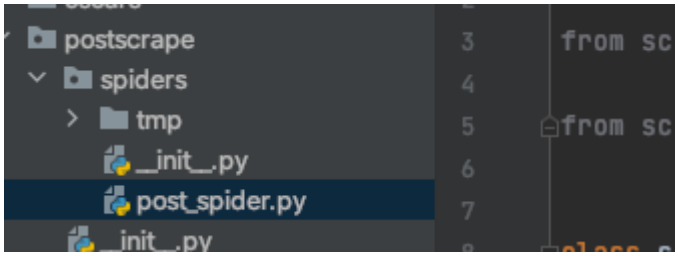
Detailed description about our methods:

In Beautifulsoup (Jabir's);

In the creation of the website, as you can see from the content, in the back-end part PHP has been used and because of this we cannot apply BeautifulSoup and get the accurate output. To see more in details, please, kindly follow the attached link to check which tools has been used within the creation of the website. It has mobile applications in Playstore as well which proves that website is totally dynamic one.

In Scrapy (Evrim's)

I created project via terminal and in the spider file, I created my scrap file (post_spider.py) , by this I am able to download HTML.



I imported scrapy framework-created a spider class and named it (shopclues) and for further needs, I also imported selector libraries.

After indicating allowed domains (which is the website itself), I put my start_urls

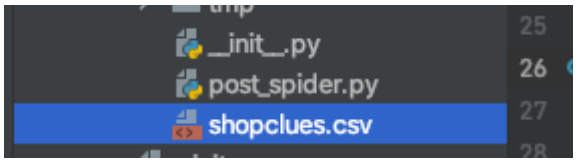
And then, I start scraping the data. I specified the data where it is in css file and bring them with using selectors when it parsed and processed the data

In the for loop, I extracted all the materials I need with their selectors. (My main selector is in a div section which is named col3.)

```
# extract product's name-orginal price-discount ratio-its image url
def parse(self, response):
    img = response.css('img::attr(data-img)').extract()
    for index, col3 in enumerate(response.css('div.col3')):
        yield {
            'name': col3.css('span.prod_name::text').extract_first(),
            'price': col3.css('span.p_price::text').extract_first(),
            'discount': col3.css('span.prd_discount::text').extract_first(),
            'image': img[index],
        }
```

In the end, I run the code below in the console, so I could get my result as a proper csv format.

```
-06 11:48:31 [scrapy.core.engine] INFO: Spider closed (finished)
ewrimmm@Evrims-MacBook-Air spiders % scrapy crawl shopclues -o shopclues.csv
```



name	price	discount	image
TSV AK -16 Gamepad Gaming Joystick PUBG Game Controlle	₹359,00	64% Off	https://cdn.shopclues.com/images1/thumbnails/111934/280/1/151379128-111934049-1604730314.jpg
Inext Tv video Game with 22 Inbuilt Games	₹1,62	37% Off	https://cdn.shopclues.com/images1/thumbnails/107977/280/1/150138423-107977538-1592303111.jpg
R-11 Mobile Game Controller Trigger for Playing PUBG,	₹338,00	66% Off	https://cdn.shopclues.com/images1/thumbnails/111934/280/1/151379147-111934072-1604730382.jpg
Inext Tv Video Game (With 22 Inbuilt Games)	₹1,61		https://cdn.shopclues.com/images1/thumbnails/102449/280/1/147581381-102449486-1566456170.jpg
R11 PUBG Trigger Fire Free Button	₹338,00	66% Off	https://cdn.shopclues.com/images1/thumbnails/111934/280/1/151379153-111934078-1604730393.jpg
AK-16 Blue Red Mobile Gaming Trigger Fire Button Aim K	₹361,00	63% Off	https://cdn.shopclues.com/images1/thumbnails/109207/280/1/150594654-109207763-1598010452.jpg
SR-2000 Controller Shooter Gaming Button Gamingpad	₹632,00	66% Off	https://cdn.shopclues.com/images1/thumbnails/109071/280/1/150542233-109071590-1597063613.jpg
Game Controller PUBG Gamepad Joystick Metal L1 R1 Trigg	₹455,00	54% Off	https://cdn.shopclues.com/images1/thumbnails/111934/280/1/151379397-111934459-1604731772.jpg
TSV AK-16 Pubg Game Metal (Game Controller)Lightweig	₹359,00	64% Off	https://cdn.shopclues.com/images1/thumbnails/111934/280/1/151379197-111934211-1604730666.jpg
TSV AK-16 Joystick Fire Buttons L1 R1 Trigger Controll	₹359,00	64% Off	https://cdn.shopclues.com/images1/thumbnails/111934/280/1/151379193-111934207-1604730654.jpg
TSV AK16 Game Controller L1 R1 Wireless Controller Tri	₹359,00	64% Off	https://cdn.shopclues.com/images1/thumbnails/111934/280/1/151379194-111934208-1604730658.jpg

Performance: In addition, as known scrapy's has the biggest advanced is speed.

```
'scheduler/dequeued': 4,
'scheduler/dequeued/memory': 4,
'scheduler/enqueued': 4,
'scheduler/enqueued/memory': 4,
'spider_exceptions/IndexError': 1,
'start_time': datetime.datetime(2021, 5, 9, 17, 41, 37, 511720)}
2021-05-09 20:41:40 [scrapy.core.engine] INFO: Spider closed (finished)
(base) ewrimmm@Evrims-MacBook-Air spiders %
```

In Selenium (Yenish's)

If speed isn't a top priority, Selenium will be a good option.

I did use the Selenium WebDriver with the featured browser automation APIs.

I found Selenium' functionality is very useful in web scraping because a lot of today's modern web pages make extensive use of JavaScript to dynamically populate the page.

Code Overview:

Data set and impotrted libraries

Selenium requires a driver and I have chosen Firefox browser.

```
from selenium import webdriver
import time
import pandas as pd
gecko_path = 'C:\\Users\\yenis\\Desktop\\Web_SMS\\project
WEB_SCPY\\geckodriver.exe'

url = 'https://www.shopclues.com/'
options = webdriver.firefox.options.Options()
options.headless = False

driver = webdriver.Firefox(options = options, executable_path =
gecko_path)

driver.get(url)

dil = pd.DataFrame({"Brand":[], "Price":[], "Discount":[]})
```

You can find link of the website which informations were taken “https://www.shopclues.com/branded-deals.html”

```
links=[link.get_attribute('href') for link in shows]
print(links)
for link in links:
    print(link)
    print("####################")
    driver.get(link)
    time.sleep(1)


    try:

        brand =
driver.find_element_by_xpath("//div//li//a[@class='Brand']",
'https://www.shopclues.com/').text
        print(brand)
    except:

        print("Not Found")

    try:

        price =
driver.find_element_by_xpath("//div//li//a[@class='Price']",
'https://www.shopclues.com/').text
        print(price)
    except:
        print("Not Found")

    try:

        discount =
driver.find_element_by_xpath("//div//li//a[@class='Discount']",
'https://www.shopclues.com/').text
        print(discount)
    except:
        print("Not Found")


        shop = {"Brand": brand, "Price": price, "Discount":discount}

        dil=d.append(shop, ignore_index = True)

dil.to_csv('SSS.csv')
```

Find elements was by: xpath

Start time mechanism.

```
start = time.time()
print("Running time: ",time.time() - start)
```

Comparison table of our libraries

	Scrapy	Beautiful Soup	Selenium
What is it?	Web scraping framework	Library	Library
Purpose	Complete web scraping solution	Data parser	Scriptable web browser to render javascript
Ideal use case	Development of recurring or large scale web scraping projects	Simple non-recurring web scraping tasks	Small-scale web scraping of javascript heavy websites
Built-in Data Storage Supports	JSON, JSON lines, XML, CSV	Need to develop your own	Customizable
Available selectors	JCSS & Xpath	CSS	CSS & Xpath
Asynchronous	Yes	No	No
Javascript support	Yes,via Splash library	No	Yes
Documentation	Excellent	Excellent	Good
Learning curve	Easy	Very easy	Easy
Ecosystem	Large ecosystem of developers contributing projects and support on Github and StackOverflow	Few related projects or plugins	Few related projects or plugins

Data analysis:

By using these methods, we have a basic market analysis result. So, we can observe which product is pricey or cheaper. Also, we can analyse the discount ratio throughout all products, hence understand what kind of products have the discount. By these results, we will estimate the customer's choices and build a competitive pricing structure, advertising plans for further studies.

Our basic result format is csv and it is including these columns;

- **Name**
- **Price**
- **Discount**
- **image**

For more advanced further analysis;

we can expand our results by scraping more data; expanding more product categories, location information, choice of bands, payment options.

Collecting that kind of data, we can build our data model. And for example, with using ***Pandas*** we can build and plot some estimated models (***matplotlib, ggplot, seaborn***) turning into meaningful visualizations such as

- ***customer's choice model or***
- ***which brands can prefer or***
- ***price analysis***

For example: Using ascending format on price column, so we can observe which products are cheaper for further raise or vice versa

	name	price
2	Hitman Blood Money - Pc	₹1,28
3	Hitman Codename 47	₹1,28
4	HP H2800 Headset Black with in-line Microphone Headset	₹1,57
5	Inext Tv Video Game (With 22 Inbuilt Games)	₹1,61
6	Inext Tv video Game with 22 Inbuilt Games	₹1,62
7	video game king game2 inbuilt games 8 bit games for all	₹1,70
8	Hitman 2 Silent Assassin - PC	₹2,14
9	Tiitan Active Noise Cancellation Wireless Over the Ear	₹2,24
10	Hitman Collection	₹2,57
11	Ikall K1 1 GB RAM 8 GB ROM Smartphones	₹3,61
12	I KALL K8Plus 5.5 Inch Display 2 GB RAM 16 GB ROM Smart	₹3,71
13	IKall K1 (Dual Sim,1 GB, 8 GB) + Neckband Music Player	₹3,71
14	I Kall K600 Smartphone (5 Inch Display, 2GB RAM, 16GB I	₹3,90
15	IKall K4 Smart Phone 4G VoLTE 2 GB 16 GB with Bluetoot	₹3,94
16	I Kall K5 2 GB RAM 16 GB ROM (Blue) Smart Phone	₹4,08
17	I Kall K8 New (5.5Inch Display, 4G, 2GB RAM, 16GB) Mob	₹4,08
18	I KALL K8 (5.7 inch, Dual Sim, 4G) Mobile Phone with M	₹4,08
19	I Kall K9 5.99 Inch Display 4G Smartphone Blue (2GB RAM	₹4,39
20	IKall K250 6.53 Inch HD+ (4GB, 64GB, 4G Volte) Blue	₹5,85
21	I Kall K320 6.53 Inchs Display 4+64 Smartphone	₹5,88

Gathering crucial information from these analyses, we can also offer some suggestions to the company. For example;

- The website can show more advertisements about discount days including the brands' information.
- If there are less popular but yet cheaper products for those who prefer the same brand, the website can show some recommendation models using the methodologies that we mentioned above.
- Regarding the correlation between price and product, the website can use some advertising boost with eye-catching thoughts

Here are some analyzing models that we built

```
import numpy as np
import pandas as pd
```

```
df = pd.read_csv (r'/Users/ewrimmm/Desktop/webScrapping/project/shopclues.csv')
print (df)
```

```
[73 rows x 4 columns]
```

5]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 73 entries, 0 to 72
Data columns (total 4 columns):
 #   Column    Non-Null Count  Dtype 
---  ------    --------------  ----- 
 0   name      73 non-null     object
 1   price     73 non-null     object
 2   discount  71 non-null     object
 3   image     73 non-null     object
dtypes: object(4)
memory usage: 2.4+ KB
```

```
df.head(10)
```

	name	price	discount	image
0	TSV AK -16 Gamepad Gaming Joystick PUBG Game ...	₹359	64% Off	https://cdn.shopclues.com/images1/thumbnails/1...
1	Inext Tv video Game with 22 Inbuilt Games	₹1,615 ...	37% Off	https://cdn.shopclues.com/images1/thumbnails/1...
2	R-11 Mobile Game Controller Trigger for Playin...	₹338	66% Off	https://cdn.shopclues.com/images1/thumbnails/1...
3	Inext Tv Video Game (With 22 Inbuilt Games)	₹1,606 ...	NaN	https://cdn.shopclues.com/images1/thumbnails/1...
4	R11 PUBG Trigger Fire Free Button	₹338	66% Off	https://cdn.shopclues.com/images1/thumbnails/1...
5	AK-16 Blue Red Mobile Gaming Trigger Fire But...	₹361	63% Off	https://cdn.shopclues.com/images1/thumbnails/1...
6	SR-2000 Controller Shooter Gaming Button Gamin...	₹632	66% Off	https://cdn.shopclues.com/images1/thumbnails/1...
7	Game Controller PUBG Gamepad Joystick Metal L1...	₹455	54% Off	https://cdn.shopclues.com/images1/thumbnails/1...
8	TSV AK-16 Pubg Game Metal (Game Controller)...	₹359	64% Off	https://cdn.shopclues.com/images1/thumbnails/1...
9	TSV AK-16 Joystick Fire Buttons L1 R1 Trigger...	₹359	64% Off	https://cdn.shopclues.com/images1/thumbnails/1...

```
# Sort the products by their discount range...
df_result = df.sort_values('discount', ascending=True)
# ...and display the top 15
display(df_result.head(15))
```

	name	price	discount	image
56	Infinix Hot 10 (Amber Red, 64 GB) (4 GB RAM)	₹10,469 ...	12% Off	https://cdn.shopclues.com/images1/thumbnails/1...
49	Infinix Hot 10 (Moonlight Jade, 64 GB) (4 GB ...	₹10,469 ...	12% Off	https://cdn.shopclues.com/images1/thumbnails/1...
24	HP H2800 Headset Black with in-line Microphone...	₹1,567 ...	21% Off	https://cdn.shopclues.com/images1/thumbnails/1...
54	Redmi 9A 2GB 32GB (Midnight Black)	₹6,588 ...	22% Off	https://cdn.shopclues.com/images1/thumbnails/1...
51	Redmi 9i 4 GB 64 GB RAM	₹8,550 ...	22% Off	https://cdn.shopclues.com/images1/thumbnails/1...
66	Realme Narzo 20 4 GB 64 GB Glory Silver	₹10,090 ...	22% Off	https://cdn.shopclues.com/images1/thumbnails/1...
60	Redmi 9 (Carbon Black, 64 GB) (4 GB RAM)	₹8,170 ...	25% Off	https://cdn.shopclues.com/images1/thumbnails/1...
61	I Kall K9 5.99 Inch Display 4G Smartphone Blue...	₹4,389 ...	26% Off	https://cdn.shopclues.com/images1/thumbnails/1...
68	Realme C15 4 GB 64 GB Power Silver	₹9,569 ...	26% Off	https://cdn.shopclues.com/images1/thumbnails/1...
70	Redmi 9A 3GB 32GB (Nature Green)	₹6,922 ...	27% Off	https://cdn.shopclues.com/images1/thumbnails/1...
69	Tecno Spark Go 2020 (Ice jadeite, 32 GB) (2 G...	₹6,520 ...	27% Off	https://cdn.shopclues.com/images1/thumbnails/1...
15	Hitman 2 Silent Assassin - PC	₹2,138 ...	27% Off	https://cdn.shopclues.com/images/thumbnails/54...
62	Ikall K1 1 GB RAM 8 GB ROM Smartphones	₹3,610 ...	27% Off	https://cdn.shopclues.com/images/thumbnails/86...
14	Hitman Collection	₹2,565 ...	27% Off	https://cdn.shopclues.com/images/thumbnails/50...
12	Hitman Blood Money - Pc	₹1,283 ...	29% Off	https://cdn.shopclues.com/images/thumbnails/50...

for processing time

```
[34]: from timeit import default_timer
      beginning = default_timer()
      my_list = list(df)
      ending = default_timer()

      print((ending - beginning)*1000)
```

```
0.3272909998486284
```