

CSE-3024 Web Mining

Lab Assignment 4b (Using Scrapy)

Alokam Nikhitha

19BCE2555

Question

Experiment 4 b

1. Use BeautifulSoup or Scrapy to crawl any one of the E-commerce website of your choice and perform the same. The following information needs to be extracted from the page: (Choose any one product : e.g laptop, smartphone ... etc)

- a) Product Name
- b) Product price
- c) Product discount
- d) Product image

Problem statement:

To Crawl any of the E-commerce website and extract the data from the page like Name, Price, Discount and image of the Product **using only Scrapy**.

Procedure:

- Firstly we install scrapy package with “pip install scrapy” in anaconda prompt
- Later, we can start Shell by “scrapy shell”
- Then Crawler run in the shell by use of the fetch and using view(response) to view fetched data.
- An object should be created for the scrapper by “scrapy startproject mobile”
- Create folder named “mobile” and move to that particular folder using command “cd mobile”
- Create a python(.py) file inside the “spider” folder by using the command “scrapy genspider ..url..”
- Here I scrapped data of amazons mobile as by product so the same url is pasted here.
- Then python code is written in the file.
- We can view the output in the terminal on typing “scrapy crawl ..name.. “ on terminal
- Finally it is exported as csv file using command “scrapy crawl mob -o data.csv”.

Installing Scrapy in Anaconda .

```
Anaconda Prompt (anaconda3) - scrapy shell

'scrappy.spidermiddlewares.offsite.OffsiteMiddleware',
'scrappy.spidermiddlewares.referer.RefererMiddleware',
'scrappy.spidermiddlewares.urllength.UrlLengthMiddleware',
'scrappy.spidermiddlewares.depth.DepthMiddleware']
2020-07-26 13:41:53 [scrapy.middleware] INFO: Enabled item pipelines:
[]
2020-07-26 13:41:53 [scrapy.extensions.telnet] INFO: Telnet console listening on 127.0.0.1:6023
2020-07-26 13:41:54 [asyncio] DEBUG: Using selector: SelectSelector
[s] Available Scrapy objects:
[s] scrapy scrapy module (contains scrapy.Request, scrapy.Selector, etc)
[s] crawler <scrapy.crawler.Crawler object at 0x0000016108E754C8>
[s] item {}
[s] settings <scrapy.settings.Settings object at 0x0000016108E755C8>
[s] Useful shortcuts:
[s] fetch(url[, redirect=True]) Fetch URL and update local objects (by default, redirects are followed)
[s] fetch(req) Fetch a scrapy.Request and update local objects
[s] shelp() Shell help (print this help)
[s] view(response) View response in a browser
2020-07-26 13:41:54 [asyncio] DEBUG: Using selector: SelectSelector
In [1]: fetch('https://www.amazon.in/s?k=mobile&ref=nb_sb_noss_2')
2020-07-26 13:42:01 [scrapy.core.engine] INFO: Spider opened
2020-07-26 13:42:02 [scrapy.downloadermiddlewares.redirect] DEBUG: Redirecting (301) to <GET https://www.amazon.in/mobile/s?k=mobile> from <GET https://www.amazon.in/s?k=mobile&ref=nb_sb_noss_2>
2020-07-26 13:42:02 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.amazon.in/mobile/s?k=mobile> (referer: None)

In [2]: view(response)
Out[2]: True
```

#creating scrapy project as name mobiles:

```
Anaconda Prompt (anaconda3)











(base) C:\Users\Dell>scrapy startproject mobile
New Scrapy project 'mobile', using template directory 'c:\users\dell\anaconda3\lib\site-packages\scrapy\templates\project', created in:
C:\Users\Dell\mobile

You can start your first spider with:
cd mobile
scrapy genspider example example.com

(base) C:\Users\Dell>cd mobile

(base) C:\Users\Dell\mobile>scrapy genspider mobile www.amazon.in/s?k=mobile&ref=nb_sb_noss_2
Cannot create a spider with the same name as your project
'ref' is not recognized as an internal or external command,
operable program or batch file.

(base) C:\Users\Dell\mobile>scrapy genspider mobiles www.amazon.in/s?k=mobile&ref=nb_sb_noss_2
Created spider 'mobiles' using template 'basic' in module:
mobile.spiders.mobiles
'ref' is not recognized as an internal or external command,
operable program or batch file.
```

 __pycache__	7/26/2020 1:44 PM
 spiders	7/26/2020 1:45 PM
 __init__	7/25/2020 11:27 AM
 items	7/26/2020 1:43 PM
 middlewares	7/26/2020 1:43 PM
 pipelines	7/26/2020 1:43 PM
 settings	7/26/2020 1:43 PM
 __pycache__	7/26/2020 1:44 PM
 __init__	7/25/2020 11:27 AM
 mobiles	7/26/2020 1:45 PM

Code:

#mobiles.py:

```
# 19BCE2555
import scrapy

class MobilesSpider(scrapy.Spider):
    name = 'mobiles'
    allowed_domains = ['www.amazon.in/s?k=mobile']
    start_urls = ['http://www.amazon.in/s?k=mobile/']

    def parse(self, response):
        i = 0
        image = response.css(".s-image-fixed-height .s-image::attr(src)").extract()
        discount = response.css(".a-letter-space+ span::text").extract()
        name = response.css(".a-color-base a-text-normal::text").extract()
        price = response.css(".a-price-whole::text").extract()
        print("NAME = ", name)
```

```
print("PRICE = ", price)
print("DISCOUNT", discount)
print("image url = ", image)
f = open('img.jpg', 'wb')
f.write(urllib.request.urlopen(image).read())
```

Items.py

```
# Define here the models for your scraped items
# 19BCE2555
# See documentation in:
# https://docs.scrapy.org/en/latest/topics/items.html

import scrapy

class MobileItem(scrapy.Item):
    # define the fields for your item here like:
    # name = scrapy.Field()
    product_name = scrapy.Field()
    product_price = scrapy.Field()
    product_discount = scrapy.Field()
    product_image= scrapy.Field()
    pass
```

Output

```
Anaconda Prompt (anaconda3)
(base) C:\Users\Dell\mobile>scrapy crawl mobiles
2020-07-26 13:51:51 [scrapy.utils.log] INFO: Scrapy 2.2.1 started (bot: mobile)
2020-07-26 13:51:51 [scrapy.utils.log] INFO: Versions: lxml 4.5.0.0, libxml2 2.9.9, cssselect 1.1.0, parsel 1.6.0, w3lib 1.22.0, Twisted 20.3.0, Python 3.7.7 (default, May
6 2020, 11:45:54) [MSC v.1916 64 bit (AMD64)], pyOpenSSL 19.1.0 (OpenSSL 1.1.1g 21 Apr 2020), cryptography 2.9.2, Platform Windows-10-10.0.18362-SP0
2020-07-26 13:51:51 [scrapy.utils.log] DEBUG: Using reactor: twisted.internet.selectreactor.SelectReactor
2020-07-26 13:51:51 [scrapy.crawler] INFO: Overridden settings:
{'BOT_NAME': 'mobile',
 'NEWSPIDER_MODULE': 'mobile.spiders',
 'ROBOTSTXT_OBEY': True,
 'SPIDER_MODULES': ['mobile.spiders']}
2020-07-26 13:51:51 [scrapy.extensions.telnet] INFO: Telnet Password: 472fe2c9971439e2
2020-07-26 13:51:52 [scrapy.middleware] INFO: Enabled extensions:
['scrapy.extensions.corestats.CoreStats',
 'scrapy.extensions.telnet.TelnetConsole',
 'scrapy.extensions.logstats.LogStats']
2020-07-26 13:51:52 [scrapy.middleware] INFO: Enabled downloader middlewares:
['scrapy.downloadermiddlewares.robotstxt.RobotsTxtMiddleware',
 'scrapy.downloadermiddlewares.httpauth.HttpAuthMiddleware',
 'scrapy.downloadermiddlewares.downloadtimeout.DownloadTimeoutMiddleware',
 'scrapy.downloadermiddlewares.defaultheaders.DefaultHeadersMiddleware',
 'scrapy.downloadermiddlewares.useragent.UserAgentMiddleware',
 'scrapy.downloadermiddlewares.retry.RetryMiddleware',
 'scrapy.downloadermiddlewares.redirect.MetaRefreshMiddleware',
 'scrapy.downloadermiddlewares.httpcompression.HttpCompressionMiddleware',
 'scrapy.downloadermiddlewares.redirect.RedirectMiddleware',
 'scrapy.downloadermiddlewares.cookies.CookiesMiddleware',
 'scrapy.downloadermiddlewares.httpproxy.HttpProxyMiddleware',
 'scrapy.downloadermiddlewares.stats.DownloaderStats']
2020-07-26 13:51:52 [scrapy.middleware] INFO: Enabled spider middlewares:
['scrapy.spidermiddlewares.httperror.HttpErrorMiddleware',
 'scrapy.spidermiddlewares.offsite.OffsiteMiddleware',
 'scrapy.spidermiddlewares.referrer.RefererMiddleware',
 'scrapy.spidermiddlewares.urllength.UrlLengthMiddleware',
 'scrapy.spidermiddlewares.depth.DepthMiddleware']
2020-07-26 13:51:52 [scrapy.middleware] INFO: Enabled item pipelines:
[]
2020-07-26 13:51:52 [scrapy.core.engine] INFO: Spider opened
2020-07-26 13:51:52 [scrapy.extensions.logstats] INFO: Crawled 0 pages (at 0 pages/min), scraped 0 items (at 0 items/min)
2020-07-26 13:51:52 [scrapy.extensions.telnet] INFO: Telnet console listening on 127.0.0.1:6023
2020-07-26 13:51:52 [scrapy.downloadermiddlewares.redirect] DEBUG: Redirecting (301) to <GET https://www.amazon.in/robots.txt> from <GET http://www.amazon.in/robots.txt>
```

The basic information of the product is highlighted below

```
Anaconda Prompt (anaconda3)
2020-07-26 13:51:53 [scrapy.downloadermiddlewares.redirect] DEBUG: Redirecting (301) to <GET https://www.amazon.in/s?k=mobile/> from <GET http://www.amazon.in/s?k=mobile/>
2020-07-26 13:51:53 [scrapy.downloadermiddlewares.redirect] DEBUG: Redirecting (301) to <GET https://www.amazon.in/mobile/s?k=mobile%2F> from <GET https://www.amazon.in/s?k=mobile/>
2020-07-26 13:51:54 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.amazon.in/mobile/s?k=mobile%2F> (referrer: None)
NAME = OPPO A5 2020 (Dazzling White, 4GB RAM, 64GB Storage) with No Cost EMI/Additional Exchange Offers
PRICE = 12,490
DISCOUNT Save ₹3,500 (22%)
image url = https://m.media-amazon.com/images/I/71wPwmxo2NL._AC_UY218_.jpg
2020-07-26 13:51:54 [scrapy.core.scraper] ERROR: Spider error processing <GET https://www.amazon.in/mobile/s?k=mobile%2F> (referrer: None)
Traceback (most recent call last):
  File "c:\users\dell\anaconda3\lib\site-packages\twisted\internet\defer.py", line 654, in _runCallbacks
    current.result = callback(current.result, *args, **kw)
  File "C:\Users\Dell\mobile\mobile\spiders\mobiles.py", line 20, in parse
    f.write(urllib.request.urlopen(image).read())
NameError: name 'urllib' is not defined
2020-07-26 13:51:54 [scrapy.core.engine] INFO: Closing spider (finished)
2020-07-26 13:51:54 [scrapy.statscollectors] INFO: Dumping Scrapy stats:
{'downloader/request_bytes': 1127,
 'downloader/request_count': 5,
 'downloader/request_method_count/GET': 5,
 'downloader/response_bytes': 74529,
 'downloader/response_count': 5,
 'downloader/response_status_count/200': 2,
 'downloader/response_status_count/301': 3,
 'elapsed_time_seconds': 2.280784,
 'finish_reason': 'finished',
 'finish_time': datetime.datetime(2020, 7, 26, 8, 21, 54, 584324),
 'log_count/DEBUG': 5,
 'log_count/ERROR': 1,
 'log_count/INFO': 10,
 'response_received_count': 2,
 'robotstxt/request_count': 1,
 'robotstxt/response_count': 1,
 'robotstxt/response_status_count/200': 1,
 'scheduler/dequeued': 3,
 'scheduler/dequeued/memory': 3,
 'scheduler/enqueued': 3,
 'scheduler/enqueued/memory': 3,
 'spider_exceptions/NameError': 1,
 'start_time': datetime.datetime(2020, 7, 26, 8, 21, 52, 303540)}
2020-07-26 13:51:54 [scrapy.core.engine] INFO: Spider closed (finished)
```

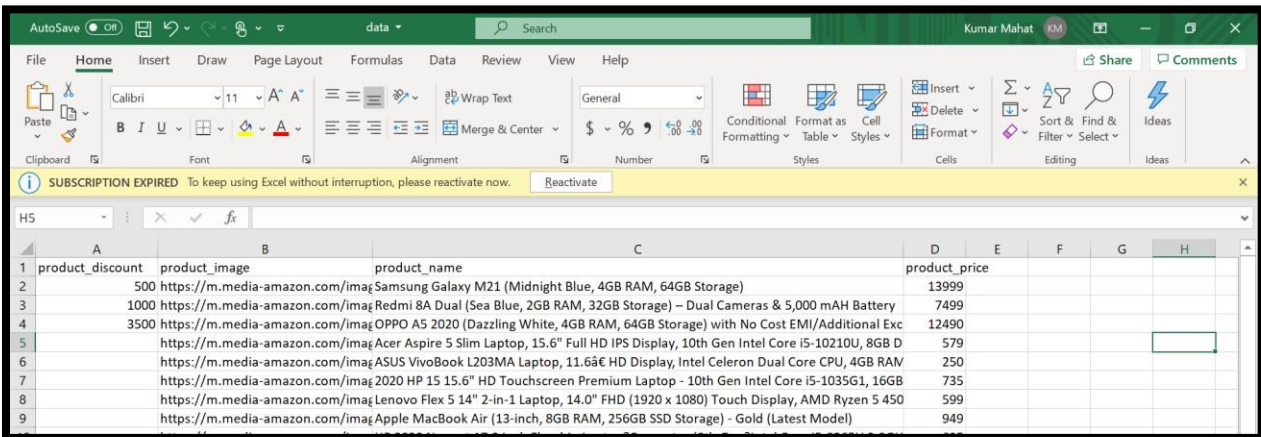
By clicking on the link extracted from the webpage we get the following image: Link: https://m.media-amazon.com/images/I/71wPwmxo2NL.AC_UY218.jpg



Results and Output

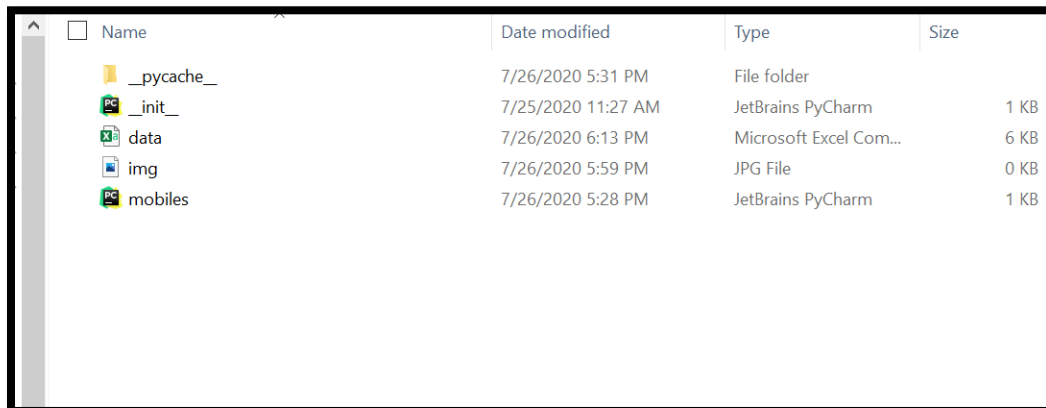
#Exporting scrapped data as csv

```
(base) C:\Users\Dell\mob\mob\spiders>scrapy crawl mob -o data.csv
```



product_discount	product_image	product_name	product_price
500	https://m.media-amazon.com/ima	Samsung Galaxy M21 (Midnight Blue, 4GB RAM, 64GB Storage)	13999
1000	https://m.media-amazon.com/ima	Redmi 8A Dual (Sea Blue, 2GB RAM, 32GB Storage) – Dual Cameras & 5,000 mAh Battery	7499
3500	https://m.media-amazon.com/ima	OPPO A5 2020 (Dazzling White, 4GB RAM, 64GB Storage) with No Cost EMI/Additional Exc	12490
	https://m.media-amazon.com/ima	Acer Aspire 5 Slim Laptop, 15.6" Full HD IPS Display, 10th Gen Intel Core i5-10210U, 8GB D	579
	https://m.media-amazon.com/ima	ASUS VivoBook L203MA Laptop, 11.6" HD Display, Intel Celeron Dual Core CPU, 4GB RAM	250
	https://m.media-amazon.com/ima	2020 HP 15 15.6" HD Touchscreen Premium Laptop - 10th Gen Intel Core i5-1035G1, 16GB	735
	https://m.media-amazon.com/ima	Lenovo Flex 5 14" 2-in-1 Laptop, 14.0" FHD (1920 x 1080) Touch Display, AMD Ryzen 5 450	599
	https://m.media-amazon.com/ima	Apple MacBook Air (13-inch, 8GB RAM, 256GB SSD Storage) - Gold (Latest Model)	949

We can see that the data is scrapped and it is dumped in excel sheet



Name	Date modified	Type	Size
__pycache__	7/26/2020 5:31 PM	File folder	
__init__	7/25/2020 11:27 AM	JetBrains PyCharm	1 KB
data	7/26/2020 6:13 PM	Microsoft Excel Com...	6 KB
img	7/26/2020 5:59 PM	JPG File	0 KB
mobiles	7/26/2020 5:28 PM	JetBrains PyCharm	1 KB