

CSE-3024 WEB MINING

LAB ASSIGNMENT 4B

Aim: Use BeautifulSoup or Scrapy to crawl any one of the E-commerce websites of your choice and perform the same. The following information needs to be extracted from the page:

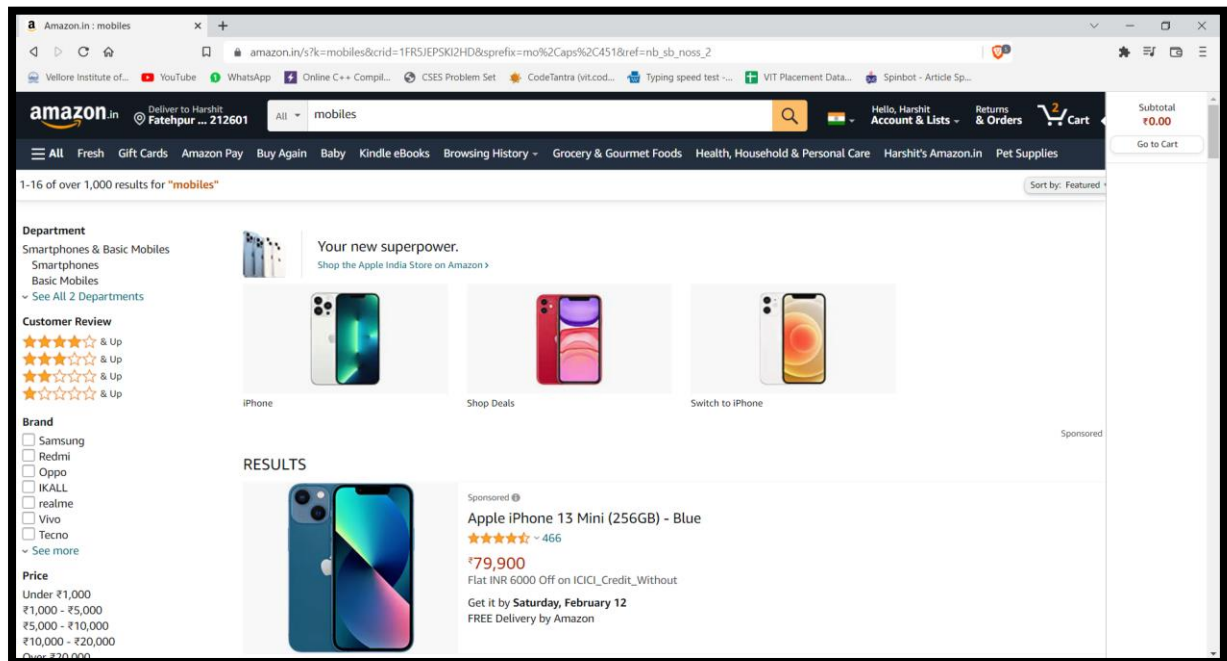
- a) Product Name
- b) Product Price
- c) Product Discount
- d) Product Image

Procedure:

- Firstly, we install scrapy package using 'pip install scrapy' in anaconda prompt shell.
- Then we can start shell by using command 'scrapy shell'.
- To initiate the crawler, we run default commands as in scrapy genspider example example.com
- Create a folder named 'mobile' and change the working directory to current folder.
- Create the python (.py) file inside the mobile directory and initiate the scrapy files using command "scrapy genspider example example.com"
- Here I have scrapped Amazon's website featuring mobile phones.
- The Python code is written in the scrapy file.
- Finally, it is exported as csv file using command "scrapy crawl -o mobile data.csv".

Harshit Mishra (19BCE0799)

Site we are scrapping:



URL of the above website:

https://www.amazon.in/s?k=mobiles&crid=1FR5JEP5KI2HD&sprefix=mo%2Caps%2C451&ref=nb_sb_noss_2

Installing Scrapy in Anaconda:

```
Anaconda Prompt (anaconda3)

(base) C:\Users\Harshit>pip install scrapy
Collecting scrapy
  Downloading Scrapy-2.5.1-py2.py3-none-any.whl (254 kB)
    |#####| 254 kB 1.6 MB/s
Collecting service-identity<=16.0.0
  Downloading service_identity-21.1.0-py2.py3-none-any.whl (12 kB)
Collecting itemloaders<=1.0.1
  Downloading itemloaders-1.0.4-py3-none-any.whl (11 kB)
Requirement already satisfied: pyOpenSSL<=16.2.0 in c:\users\harshit\anaconda3\lib\site-packages (from scrapy) (19.1.0)
Collecting h2<4.0,>=3.0
  Downloading h2-3.2.0-py2.py3-none-any.whl (65 kB)
    |#####| 65 kB 1.2 MB/s
Collecting w3lib<=1.17.0
  Downloading w3lib-1.22.0-py2.py3-none-any.whl (20 kB)
Requirement already satisfied: zope.interface<=4.1.3 in c:\users\harshit\anaconda3\lib\site-packages (from scrapy) (4.7.1)
Collecting cssselect<=0.9.1
  Downloading cssselect-1.1.0-py2.py3-none-any.whl (16 kB)
Requirement already satisfied: lxml<=3.5.0; platform_python_implementation == "CPython" in c:\users\harshit\anaconda3\lib\site-packages (from scrapy) (4.5.2)
Collecting Twisted[http2]>=17.9.0
  Downloading Twisted-22.1.0-py3-none-any.whl (3.1 MB)
    |#####| 3.1 MB 6.4 MB/s
Requirement already satisfied: cryptography<=2.0 in c:\users\harshit\anaconda3\lib\site-packages (from scrapy) (2.9.2)
Collecting parsel<=1.5.0
  Downloading parsel-1.6.0-py2.py3-none-any.whl (13 kB)
Collecting itemadapter<=0.1.0
  Downloading itemadapter-0.4.0-py3-none-any.whl (10 kB)
Collecting protego<=0.1.15
```

Here we use the command `pip install scrapy`

Creating Scrapy Project in mobile directory:

```
Anaconda Prompt (anaconda3)

(base) C:\Users\Harshit>cd "Desktop\Semester 6\G - Wen Mining Lab"

(base) C:\Users\Harshit\Desktop\Semester 6\G - Wen Mining Lab>cd "Lab Assignment 4B"

(base) C:\Users\Harshit\Desktop\Semester 6\G - Wen Mining Lab\Lab Assignment 4B>scrapy startproject mobile
New Scrapy project 'mobile', using template directory 'c:\users\harshit\anaconda3\lib\site-packages\scrapy\templates\project', created in:
C:\Users\Harshit\Desktop\Semester 6\G - Wen Mining Lab\Lab Assignment 4B\mobile

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

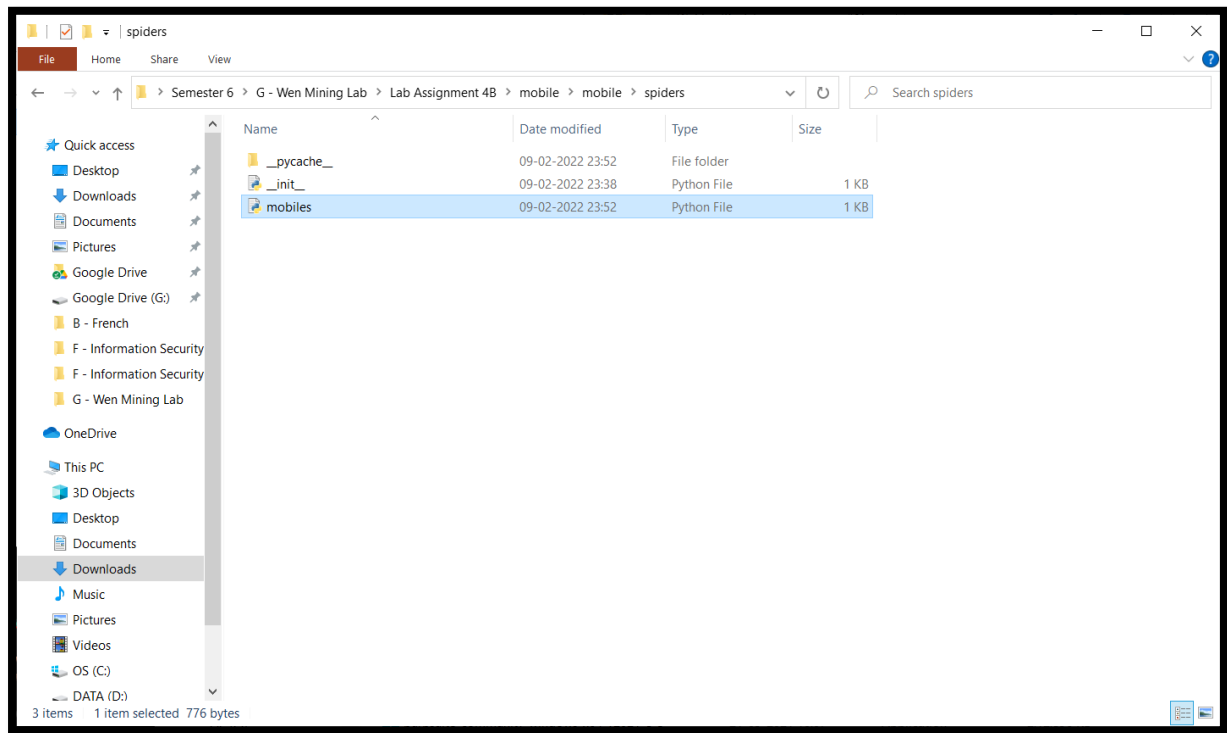
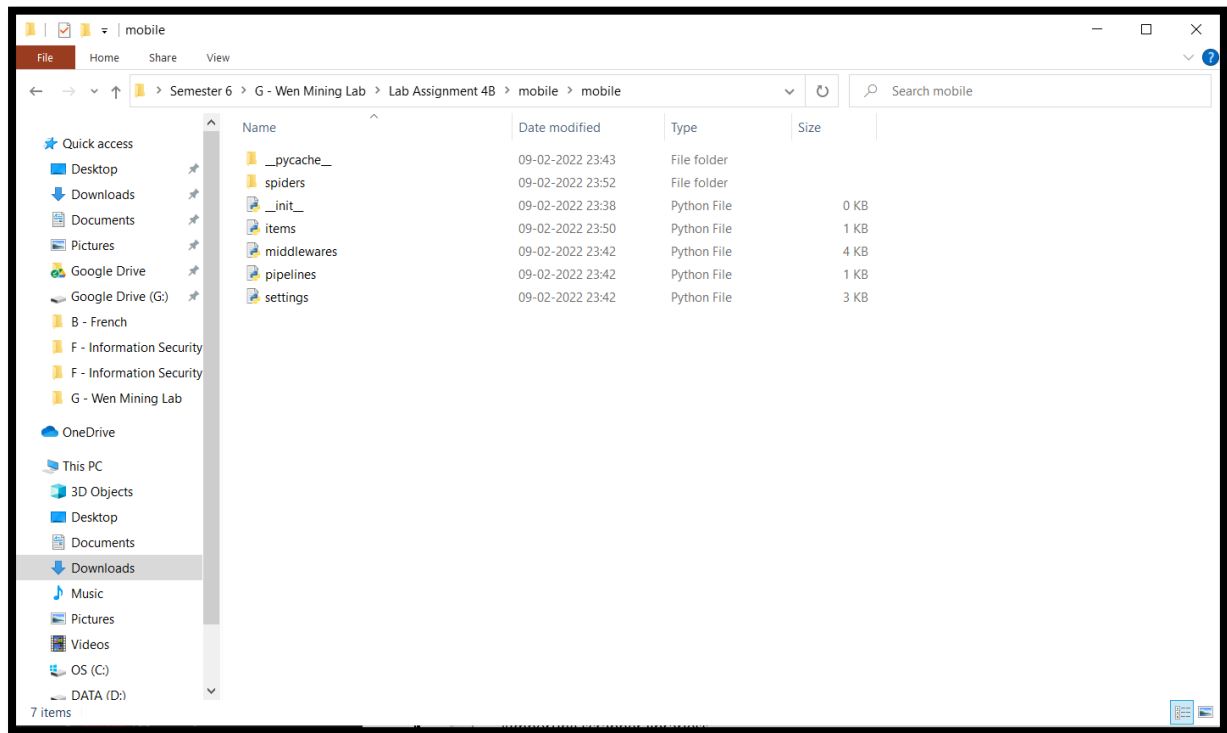
(base) C:\Users\Harshit\Desktop\Semester 6\G - Wen Mining Lab\Lab Assignment 4B>cd mobile

(base) C:\Users\Harshit\Desktop\Semester 6\G - Wen Mining Lab\Lab Assignment 4B\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.

(base) C:\Users\Harshit\Desktop\Semester 6\G - Wen Mining Lab\Lab Assignment 4B\mobile>scrapy crawl mobiles
2022-02-09 23:50:51 [scrapy.utils.log] INFO: Scrapy 2.5.1 started (bot: mobile)
2022-02-09 23:50:51 [scrapy.utils.log] INFO: Versions: lxml 4.5.2.0, libxml2 2.9.10, cssselect 1.1.0, parsel 1.6.0, w3lib 1.22.0, Twisted 22.1.0, Python 3.8.3
(default, Jul 2 2020, 17:30:36) [MSC v.1916 64 bit (AMD64)], pyOpenSSL 19.1.0 (OpenSSL 1.1.1g 21 Apr 2020), cryptography 2.9.2, Platform Windows-10-10.0.19
041-SP0
2022-02-09 23:50:51 [scrapy.utils.log] DEBUG: Using reactor: twisted.internet.selectreactor.SelectReactor
2022-02-09 23:50:51 [scrapy.crawler] INFO: Overridden settings:
{'BOT_NAME': 'mobile',
 'NEWSPIDER_MODULE': 'mobile.spiders'}
```

Here we firstly move the directory mobile that we have created.
Then we run the command `scrapy genspyder mobiles amazon.com`

Directory Structure:



Code in mobiles.py:

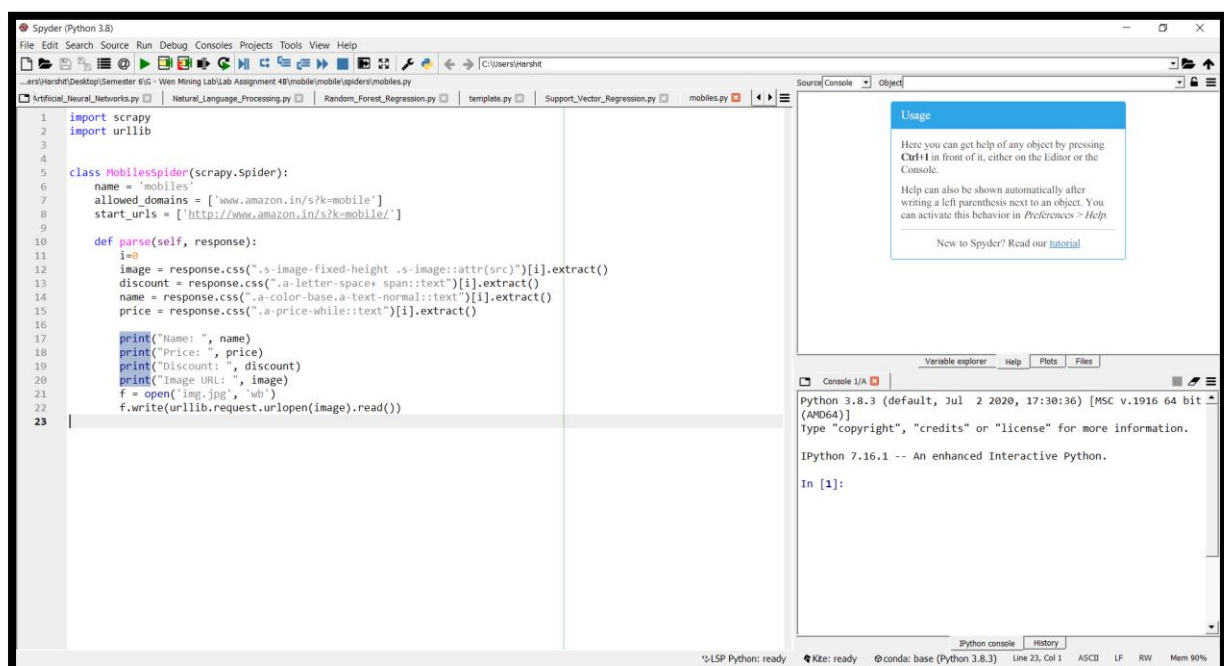
```
import scrapy
import urllib

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)")[i].extract()
        discount = response.css(".a-letter-space+ span::text")[i].extract()
        name = response.css(".a-color-base.a-text-normal::text")[i].extract()
        price = response.css(".a-price-while::text")[i].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())data!! Results in " + Fore.GREEN+"result.csv")
```

Code Snippet:



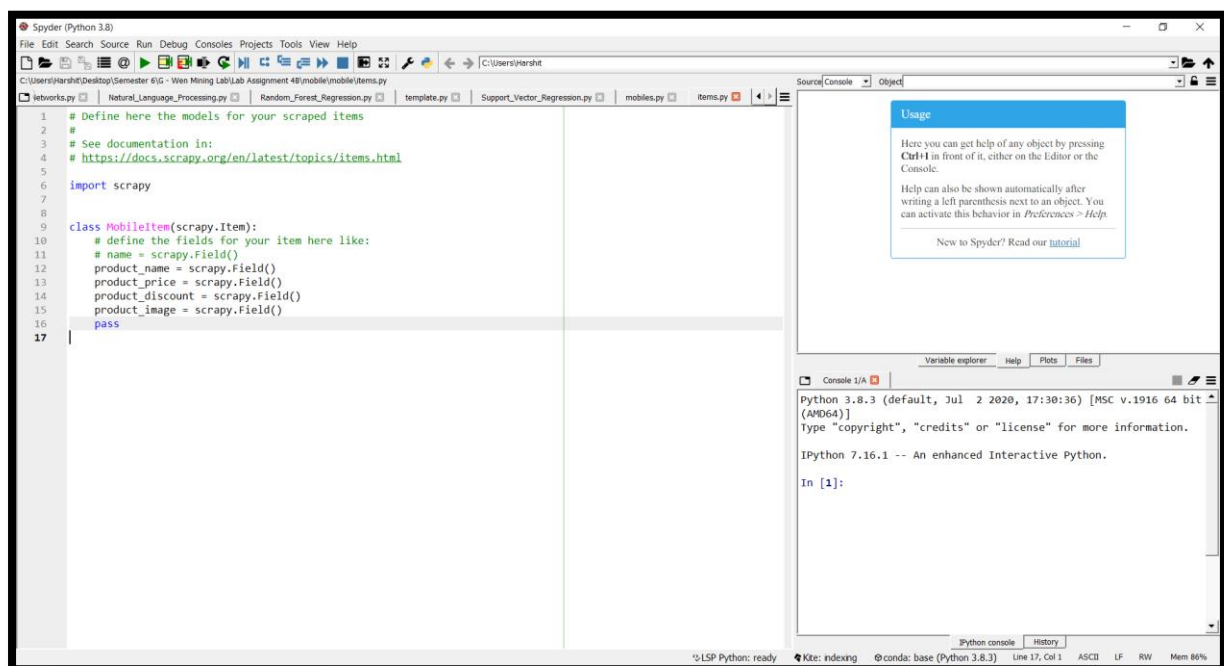
Code in Items.py:

```
# Define here the models for your scraped items
#
# 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()
```

Code Snippet:



Running the code in prompt shell:

```
Anaconda Prompt (anaconda3)

(base) C:\Users\Harshit\Desktop\Semester 6\G - Wen Mining Lab\Lab Assignment 4B\mobile>scrapy crawl mobiles
2022-02-09 23:52:55 [scrapy.utils.log] INFO: Scrapy 2.5.1 started (bot: mobile)
2022-02-09 23:52:55 [scrapy.utils.log] INFO: Versions: lxml 4.5.2.0, libxml2 2.9.10, cssselect 1.1.0, parsel 1.6.0, w3lib 1.22.0, Twisted 22.1.0, Python 3.8.3
(default, Jul 2 2020, 17:30:36) [MSC v.1916 64 bit (AMD64)], pyOpenSSL 19.1.0 (OpenSSL 1.1.1g 21 Apr 2020), cryptography 2.9.2, Platform Windows-10-10.0.19
041-SP0
2022-02-09 23:52:55 [scrapy.utils.log] DEBUG: Using reactor: twisted.internet.selectreactor.SelectReactor
2022-02-09 23:52:55 [scrapy.crawler] INFO: Overridden settings:
{'BOT_NAME': 'mobile',
 'NEWSPIDER_MODULE': 'mobile.spiders',
 'ROBOTSTXT_OBEY': True,
 'SPIDER_MODULES': ['mobile.spiders']}
2022-02-09 23:52:55 [scrapy.extensions.telnet] INFO: Telnet Password: 7a213215c3ebc497
2022-02-09 23:52:55 [scrapy.middleware] INFO: Enabled extensions:
['scrapy.extensions.corestats.CoreStats',
 'scrapy.extensions.telnet.TelnetConsole',
 'scrapy.extensions.logstats.LogStats']
2022-02-09 23:52:56 [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']
```

Here we use the command ‘scrapy crawl mobiles’ to crawl the web page.

Storing the results of scrapper in a csv file:

```
Anaconda Prompt (anaconda3)

(base) C:\Users\Harshit\Desktop\Semester 6\G - Wen Mining Lab\Lab Assignment 4B\mobile>scrapy crawl mob -o data.csv
2022-02-09 23:53:31 [scrapy.utils.log] INFO: Scrapy 2.5.1 started (bot: mobile)
2022-02-09 23:53:31 [scrapy.utils.log] INFO: Versions: lxml 4.5.2.0, libxml2 2.9.10, cssselect 1.1.0, parsel 1.6.0, w3lib 1.22.0, Twisted 22.1.0, Python 3.8.3
(default, Jul 2 2020, 17:30:36) [MSC v.1916 64 bit (AMD64)], pyOpenSSL 19.1.0 (OpenSSL 1.1.1g 21 Apr 2020), cryptography 2.9.2, Platform Windows-10-10.0.19
041-SP0
2022-02-09 23:53:31 [scrapy.utils.log] DEBUG: Using reactor: twisted.internet.selectreactor.SelectReactor
Traceback (most recent call last):
  File "c:\users\harshit\anaconda3\lib\site-packages\scrapy\spiderloader.py", line 75, in load
    return self._spiders[spider_name]
KeyError: 'mob'

During handling of the above exception, another exception occurred:

Traceback (most recent call last):
  File "c:\users\harshit\anaconda3\lib\runpy.py", line 194, in _run_module_as_main
    return _run_code(code, main_globals, None,
  File "c:\users\harshit\anaconda3\lib\runpy.py", line 87, in _run_code
    exec(code, run_globals)
  File "C:\Users\Harshit\anaconda3\Scripts\scrapy.exe\_main_.py", line 7, in <module>
  File "c:\users\harshit\anaconda3\lib\site-packages\scrapy\cmdline.py", line 145, in execute
    _run_print_help(parser, _run_command, cmd, args, opts)
  File "c:\users\harshit\anaconda3\lib\site-packages\scrapy\cmdline.py", line 100, in _run_print_help
    func(*a, **kw)
  File "c:\users\harshit\anaconda3\lib\site-packages\scrapy\cmdline.py", line 153, in _run_command
    cmd.run(args, opts)
  File "c:\users\harshit\anaconda3\lib\site-packages\scrapy\commands\crawl.py", line 22, in run
    crawl_defer = self.crawler_process.crawl(spname, **opts.spargs)
  File "c:\users\harshit\anaconda3\lib\site-packages\scrapy\crawler.py", line 191, in crawl
    crawler = self.create_crawler(crawler or spidercls)
```

Results:

result.csv file:

Directory:

We can see that the result.csv file is dumped in the directory.