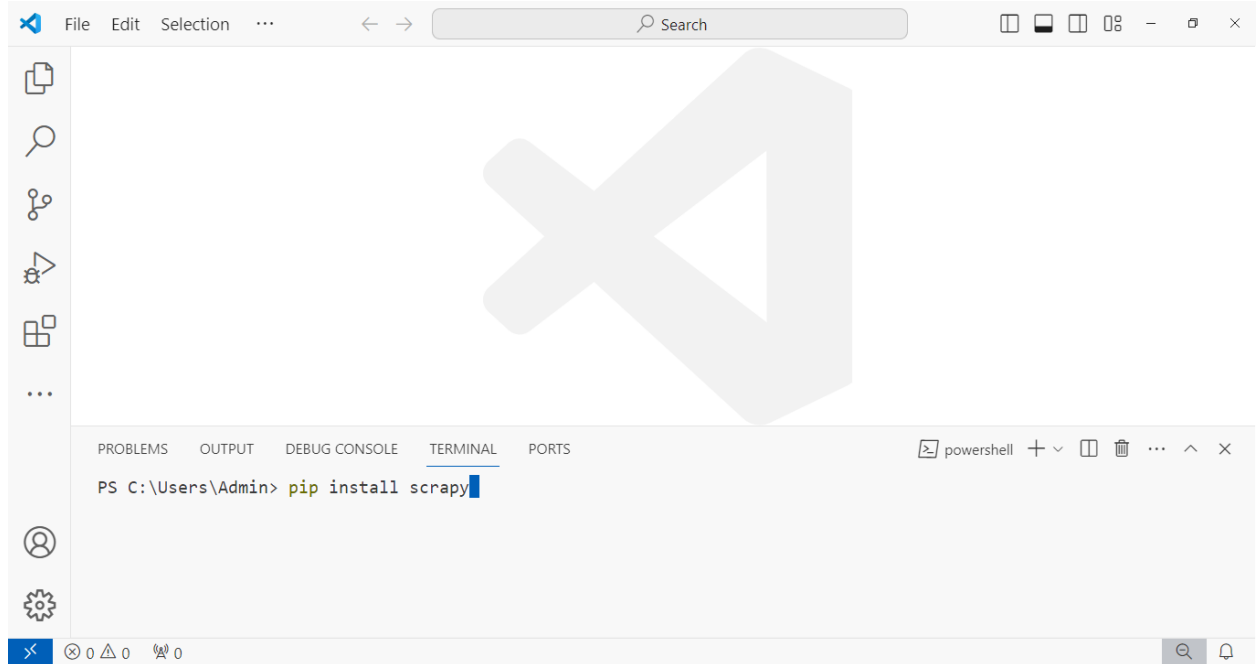


HƯỚNG DẪN SỬ DỤNG FRAMEWORK SCRAPY ĐỂ CRAWL DỮ LIỆU

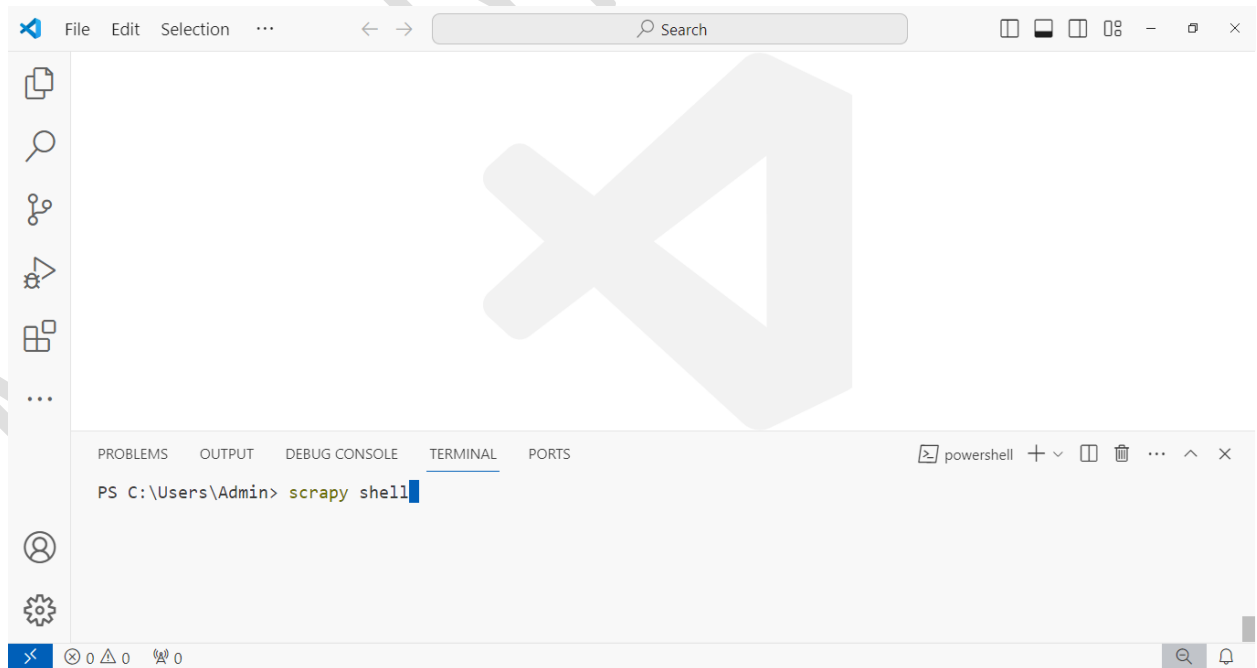
(*) Web Crawler Data: <https://hackthedeveloper.com/python-web-scraping-guide/>

1. Cài đặt Scrapy framework

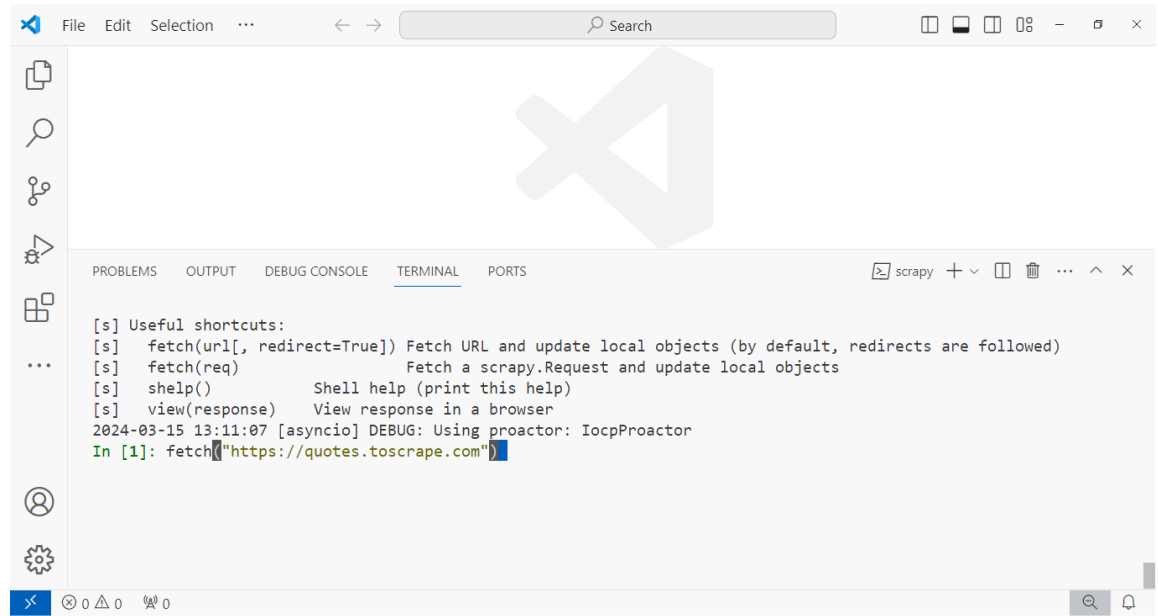


2. Kiểm tra thử với scrapy shell

a. Vào shell



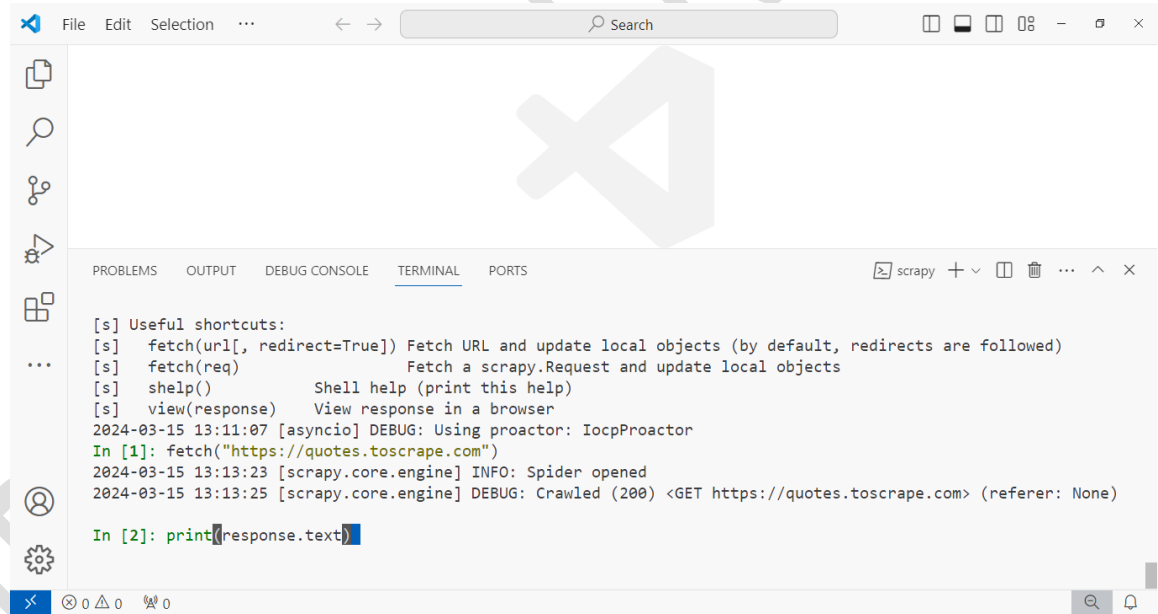
b. Nạp website muốn cào nội dung



```

File Edit Selection ... Search
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS
[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
2024-03-15 13:11:07 [asyncio] DEBUG: Using proactor: IocpProactor
In [1]: fetch("https://quotes.toscrape.com")
  
```

c. Xem thử dữ liệu được cào về



```

File Edit Selection ... Search
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS
[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
2024-03-15 13:11:07 [asyncio] DEBUG: Using proactor: IocpProactor
In [1]: fetch("https://quotes.toscrape.com")
2024-03-15 13:13:23 [scrapy.core.engine] INFO: Spider opened
2024-03-15 13:13:25 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://quotes.toscrape.com> (referer: None)
In [2]: print(response.text)
  
```

d. Thoát scrapy shell

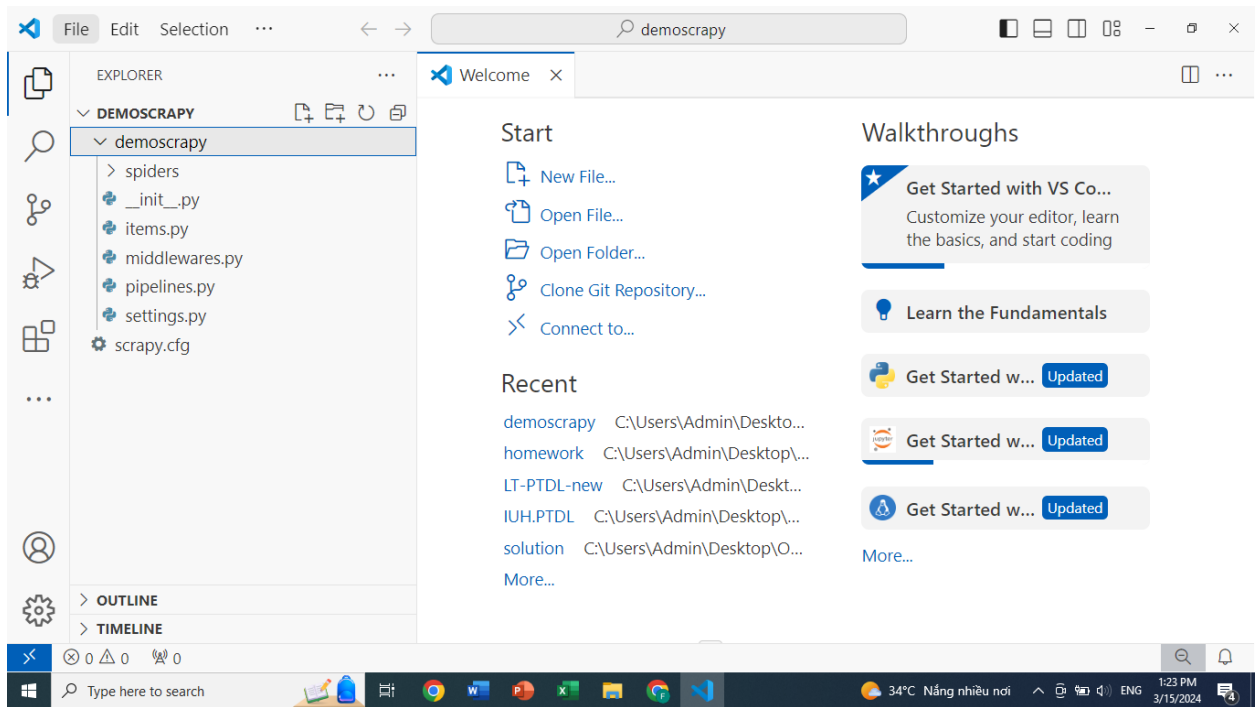
```
In [3]: exit
```

3. Tạo scrapy project mang tên demoscrapy trong ổ đĩa C

```

PS C:\Users\Admin> cd\
PS C:\> scrapy startproject demoscrapy
  
```

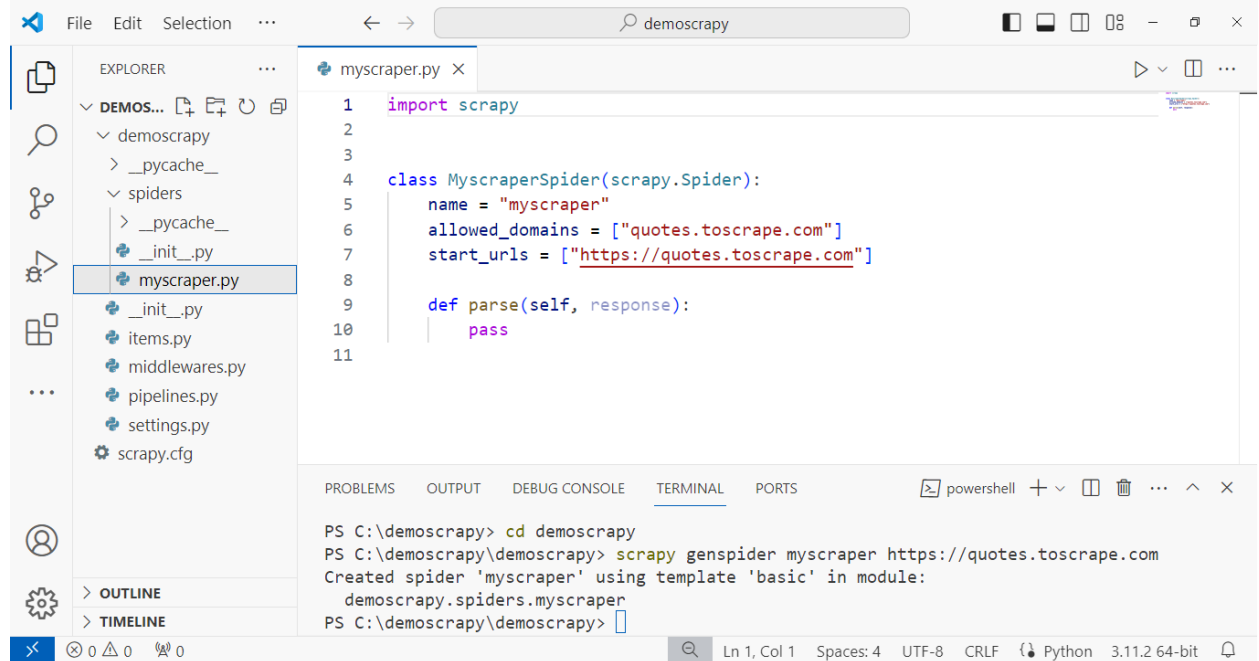
4. Xem cấu trúc project đã được tạo bằng VS Code



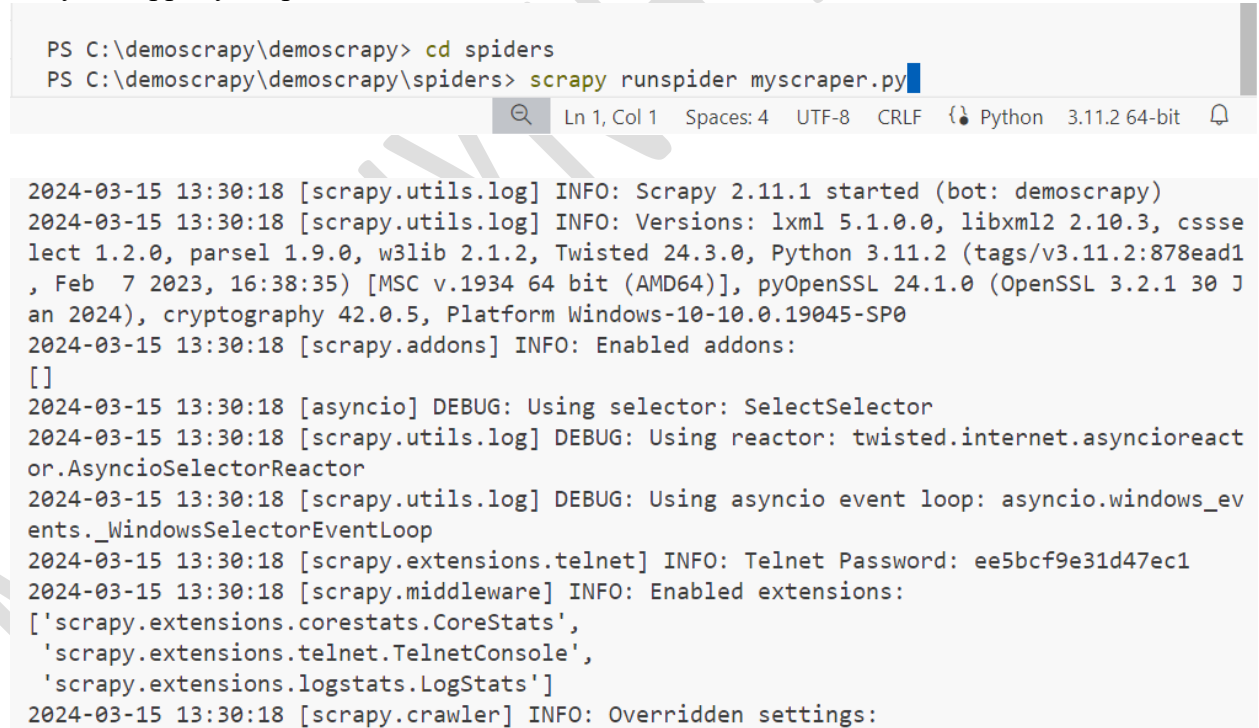
5. Tạo class để cào dữ liệu từ trang <https://quotes.toscrape.com>



6. Kết quả sau khi tạo xong

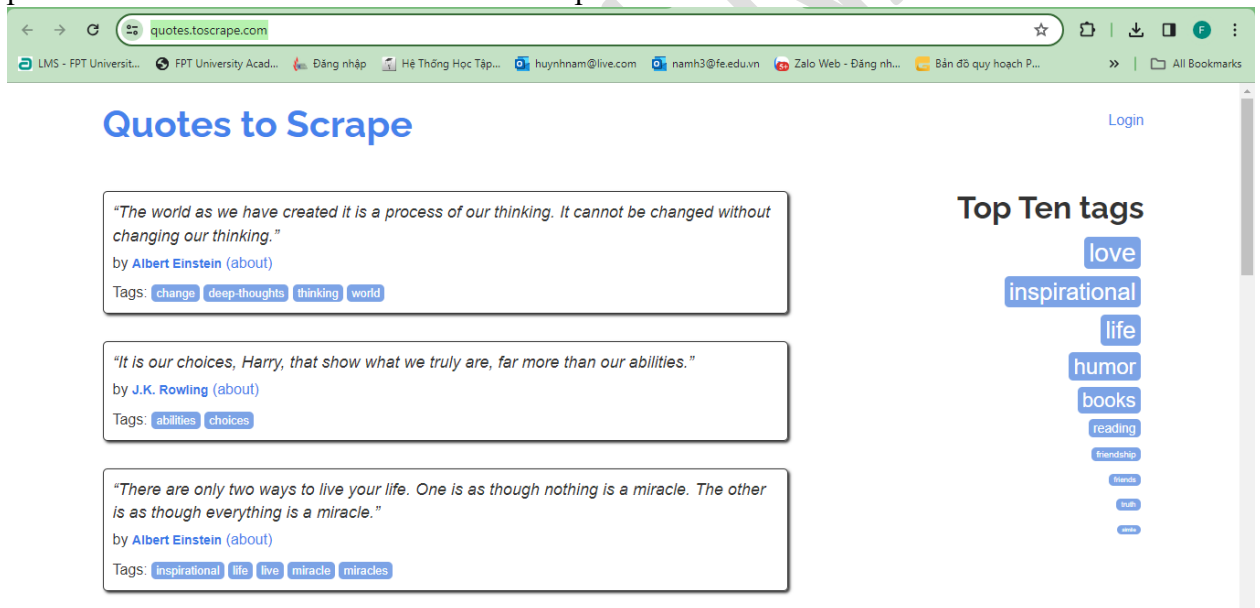


7. Chạy thử app myscraper



```
'elapsed_time_seconds': 1.443571,
'finish_reason': 'finished',
'finish_time': datetime.datetime(2024, 3, 15, 6, 30, 20, 36080, tzinfo=datetime.timezone.
utc),
'log_count/DEBUG': 5,
'log_count/INFO': 10,
'response_received_count': 2,
'robotstxt/request_count': 1,
'robotstxt/response_count': 1,
'robotstxt/response_status_count/404': 1,
'scheduler/dequeued': 1,
'scheduler/dequeued/memory': 1,
'scheduler/enqueued': 1,
'scheduler/enqueued/memory': 1,
'start_time': datetime.datetime(2024, 3, 15, 6, 30, 18, 592509, tzinfo=datetime.timezone.
utc)}}
2024-03-15 13:30:20 [scrapy.core.engine] INFO: Spider closed (finished)
```

8. Mở thử website <https://quotes.toscrape.com/> bằng trình duyệt Chrome, sau đó click chuột phải lên website bấm View Source để khám phá cấu trúc website

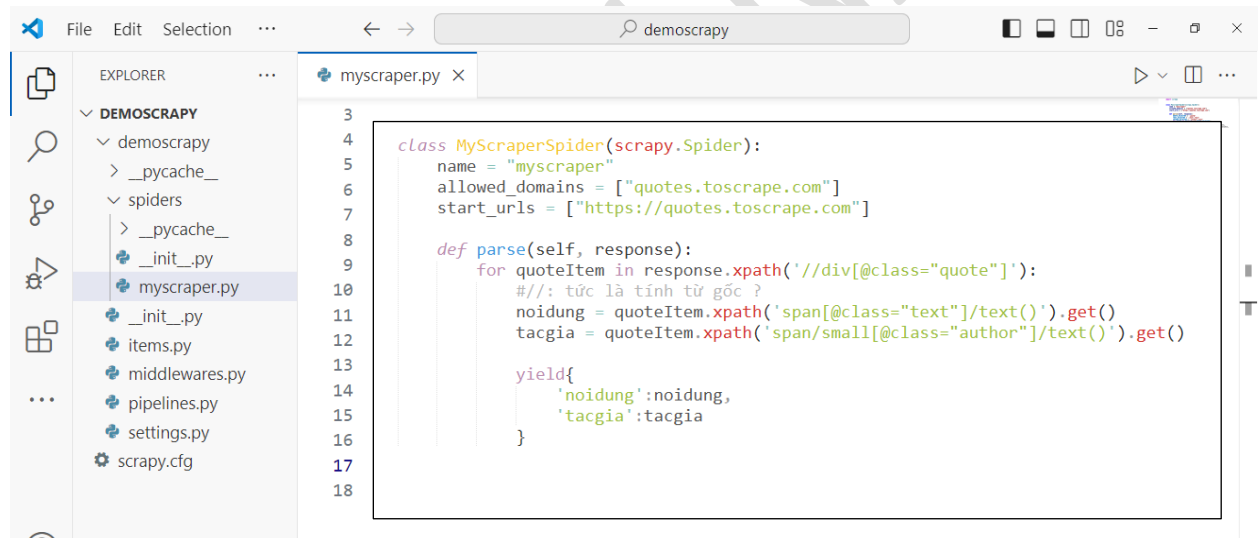


```

Line wrap
1 <!DOCTYPE html>
2 <html lang="en">
3 <head>
4   <meta charset="UTF-8">
5   <title>Quotes to Scrape</title>
6   <link rel="stylesheet" href="/static/bootstrap.min.css">
7   <link rel="stylesheet" href="/static/main.css">
8 </head>
9 <body>
10  <div class="container">
11    <div class="row header-box">
12      <div class="col-md-8">
13        <h1>
14          <a href="/" style="text-decoration: none;">Quotes to Scrape</a>
15        </h1>
16      </div>
17      <div class="col-md-4">
18        <p>
19          <a href="/login">Login</a>
20        </p>
21      </div>
22    </div>
23  </div>
24
25  <div class="row">
26    <div class="col-md-8">
27
28      <div class="quote" itemscope itemtype="http://schema.org/CreativeWork">
29        <span class="text" itemprop="text">"The world as we have created it is a process of our thinking. It cannot be changed without changing our thinking."</span>
30        <span by <small class="author" itemprop="author">Albert Einstein</small>
31          <a href="/author/Albert-Einstein">(about)</a>
32        </span>
33      </div>
34      <div class="tags">
35        <meta class="keywords" itemprop="keywords" content="change,deep-thoughts,thinking,world" />
36      </div>
37    </div>
38  </div>

```

9. Tiến hành sửa code dùng XPath để trích xuất dữ liệu text và author



```

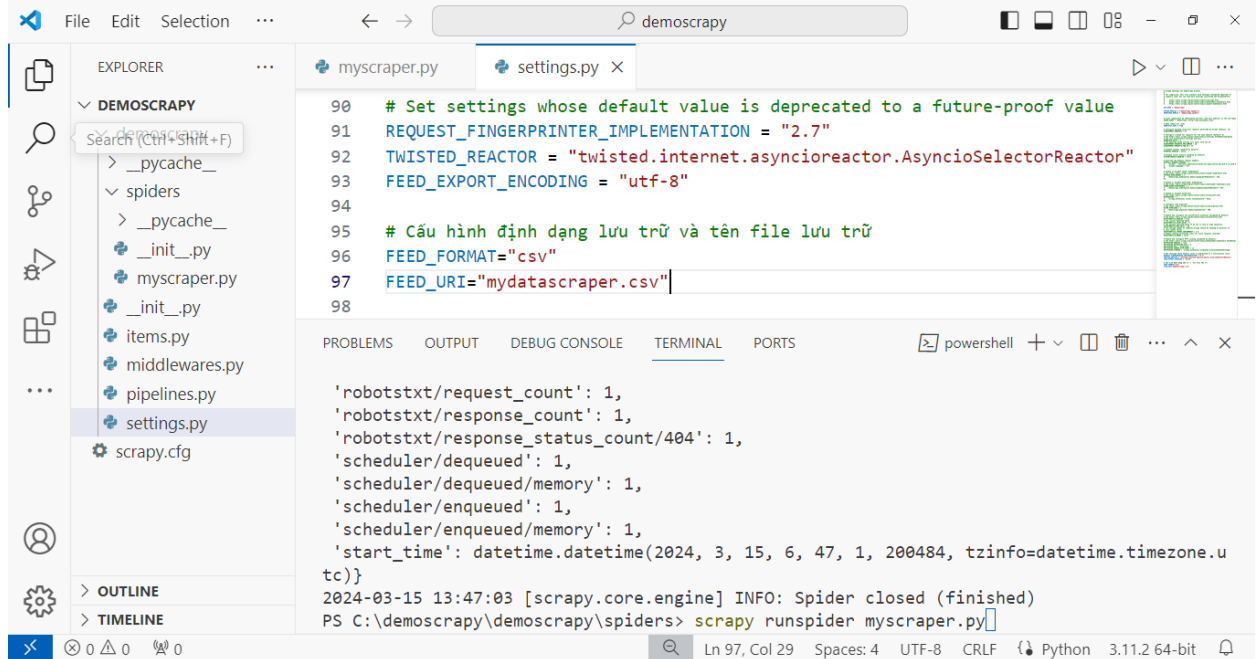
3
4
5 class MyScrapperSpider(scrapy.Spider):
6     name = "myscraper"
7     allowed_domains = ["quotes.toscrape.com"]
8     start_urls = ["https://quotes.toscrape.com"]
9
10    def parse(self, response):
11        for quoteItem in response.xpath('//div[@class="quote"]'):
12            #/: tức là tính từ gốc ?
13            noidung = quoteItem.xpath('span[@class="text"]/text()').get()
14            tacgia = quoteItem.xpath('span/small[@class="author"]/text()').get()
15
16            yield{
17                'noidung':noidung,
18                'tacgia':tacgia
19            }
20

```

10. Tiến hành chạy app cào dữ liệu

PS C:\demoscrapy\demoscrapy\spiders> scrapy runspider myscraper.py

11. Tiến hành cấu hình định dạng lưu trữ dữ liệu là csv và tên file lưu trữ trong settings.py



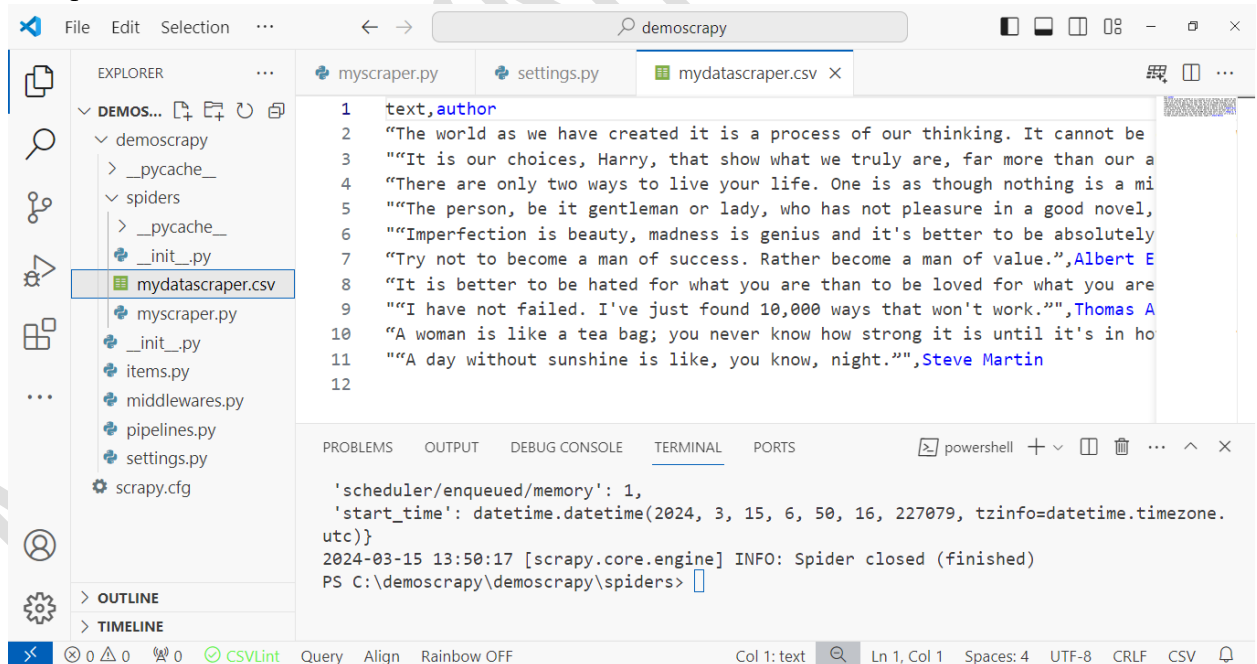
```
90 # Set settings whose default value is deprecated to a future-proof value
91 REQUEST_FINGERPRINTER_IMPLEMENTATION = "2.7"
92 TWISTED_REACTOR = "twisted.internet.asyncioreactor.AsyncioSelectorReactor"
93 FEED_EXPORT_ENCODING = "utf-8"
94
95 # Cấu hình định dạng lưu trữ và tên file lưu trữ
96 FEED_FORMAT="csv"
97 FEED_URI="mydatascraper.csv"
98
```

```
'robotstxt/request_count': 1,
'robotstxt/response_count': 1,
'robotstxt/response_status_count/404': 1,
'scheduler/dequeued': 1,
'scheduler/dequeued/memory': 1,
'scheduler/enqueued': 1,
'scheduler/enqueued/memory': 1,
'start_time': datetime.datetime(2024, 3, 15, 6, 47, 1, tzinfo=datetime.timezone.u
tc)}
2024-03-15 13:47:03 [scrapy.core.engine] INFO: Spider closed (finished)
PS C:\demoscrapy\demoscrapy\spiders> scrapy runspider myscrapper.py
```

12. Chạy lại app cào dữ liệu và xem kết quả

```
PS C:\demoscrapy\demoscrapy\spiders> scrapy runspider myscrapper.py
```

Kết quả cào dữ liệu



```
1 text,author
2 "The world as we have created it is a process of our thinking. It cannot be
3 "It is our choices, Harry, that show what we truly are, far more than our a
4 "There are only two ways to live your life. One is as though nothing is a mi
5 "The person, be it gentleman or lady, who has not pleasure in a good novel,
6 "Imperfection is beauty, madness is genius and it's better to be absolutely
7 "Try not to become a man of success. Rather become a man of value.",Albert E
8 "It is better to be hated for what you are than to be loved for what you are
9 "I have not failed. I've just found 10,000 ways that won't work.",Thomas A
10 "A woman is like a tea bag; you never know how strong it is until it's in ho
11 "A day without sunshine is like, you know, night.",Steve Martin
12
```

```
'scheduler/enqueued/memory': 1,
'start_time': datetime.datetime(2024, 3, 15, 6, 50, 16, 227079, tzinfo=datetime.timezone.
utc)}
2024-03-15 13:50:17 [scrapy.core.engine] INFO: Spider closed (finished)
PS C:\demoscrapy\demoscrapy\spiders>
```

THAM KHẢO

1. <https://www.digitalocean.com/community/tutorials/how-to-crawl-a-web-page-with-scrapy-and-python-3>
2. <https://www.datacamp.com/tutorial/making-web-crawlers-scrapy-python>