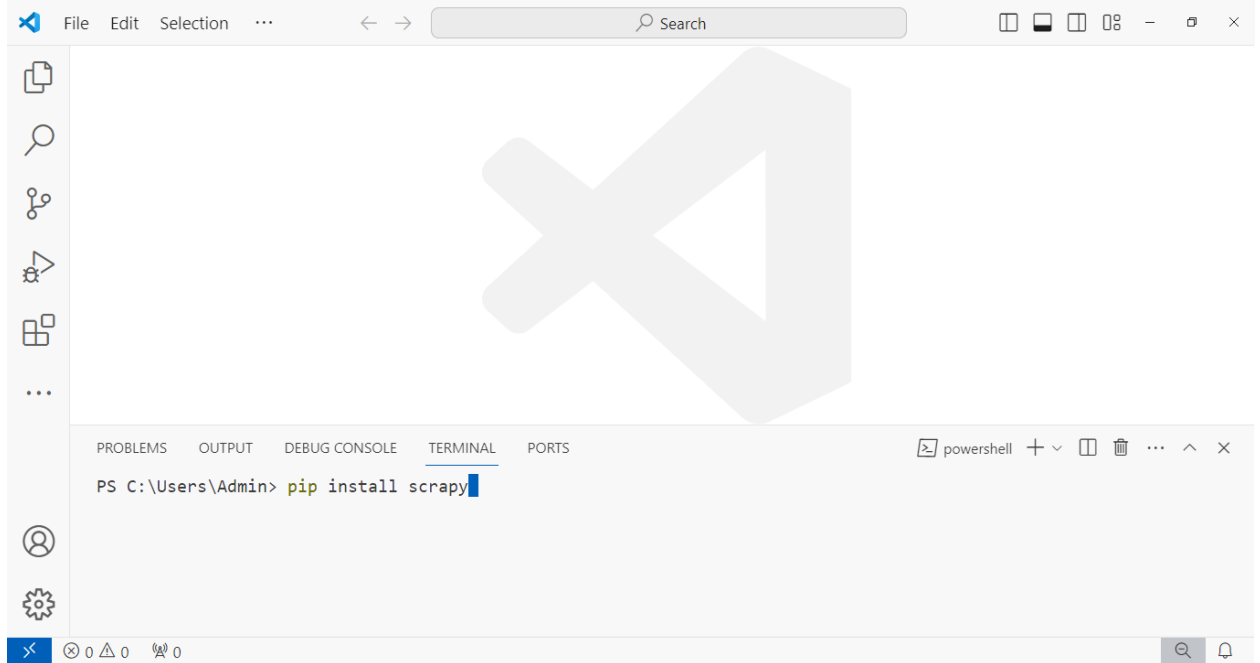


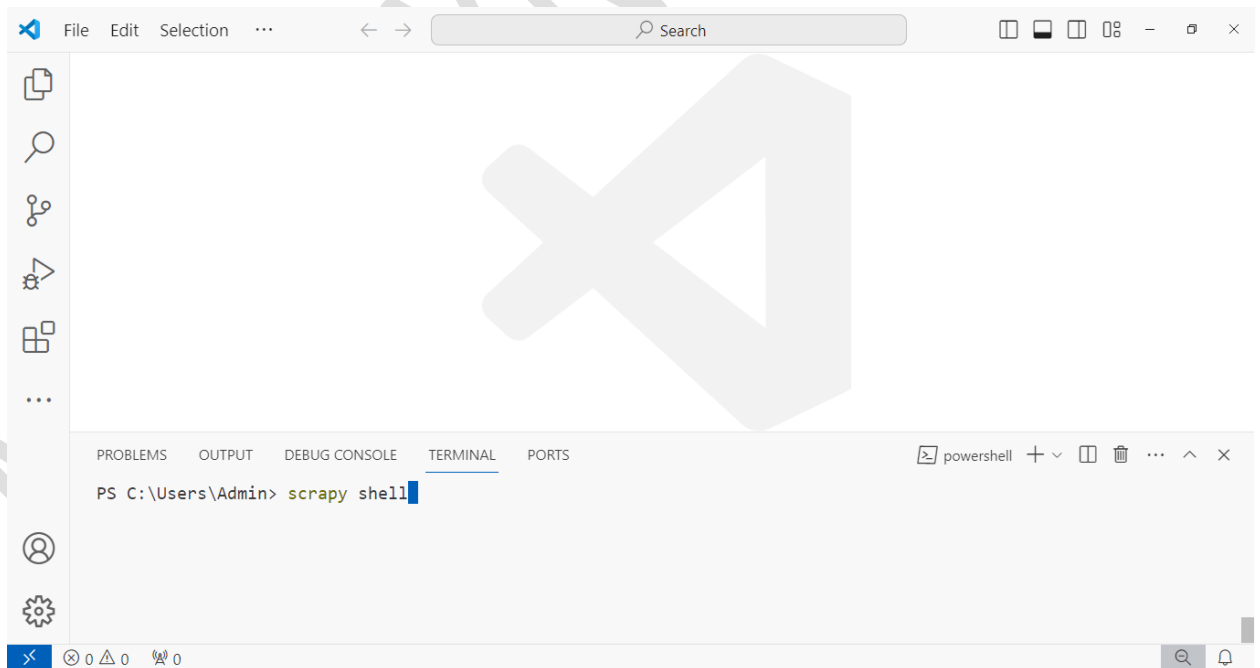
HƯỚNG DẪN SỬ DỤNG FRAMEWORK SCRAPY ĐỂ CRAWL DỮ LIỆU

1. Cài đặt Scrapy framework

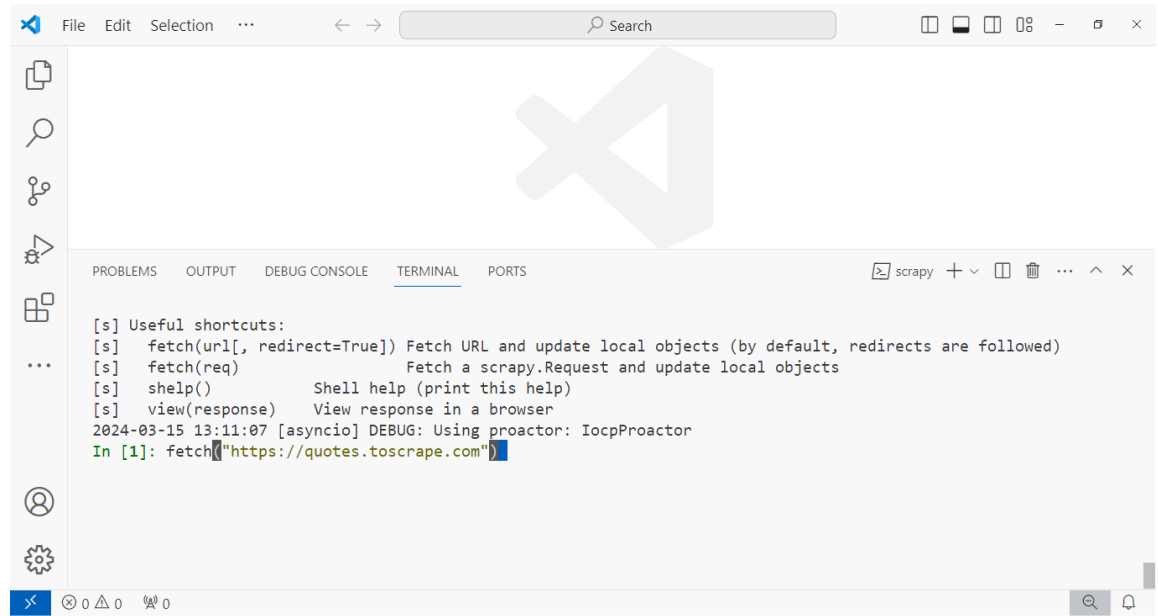


2. Kiểm tra thử với scrapy shell

a. Vào shell



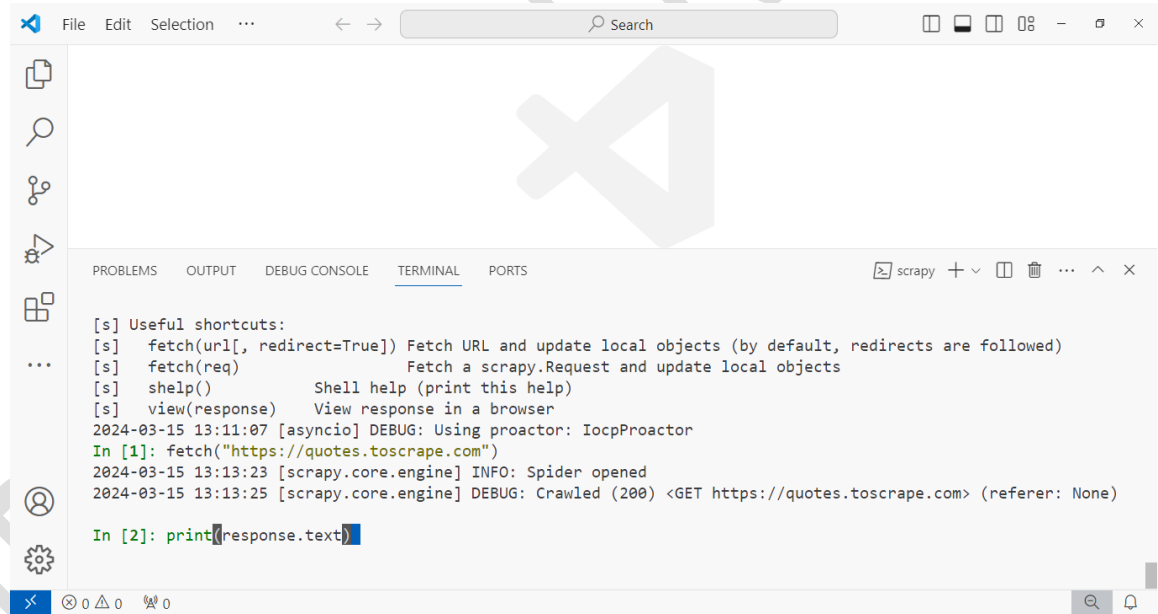
b. Nạp website muốn cào nội dung



```

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

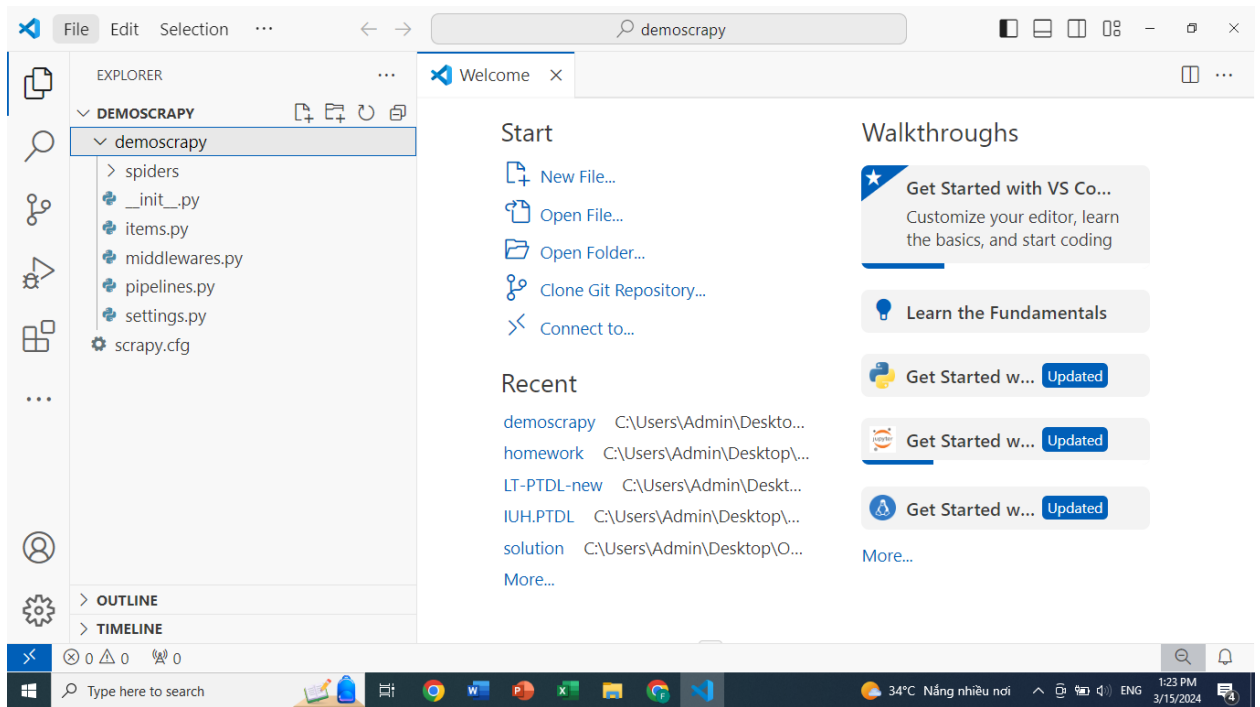
```
In [3]: exit
```

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

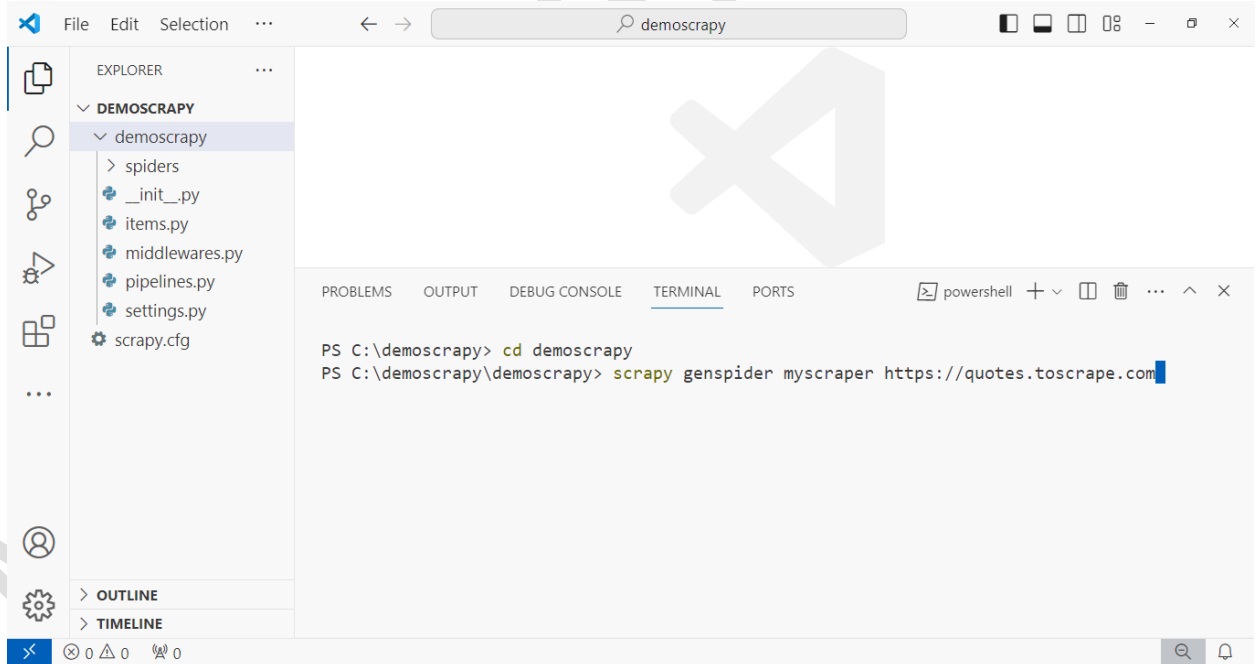
```

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

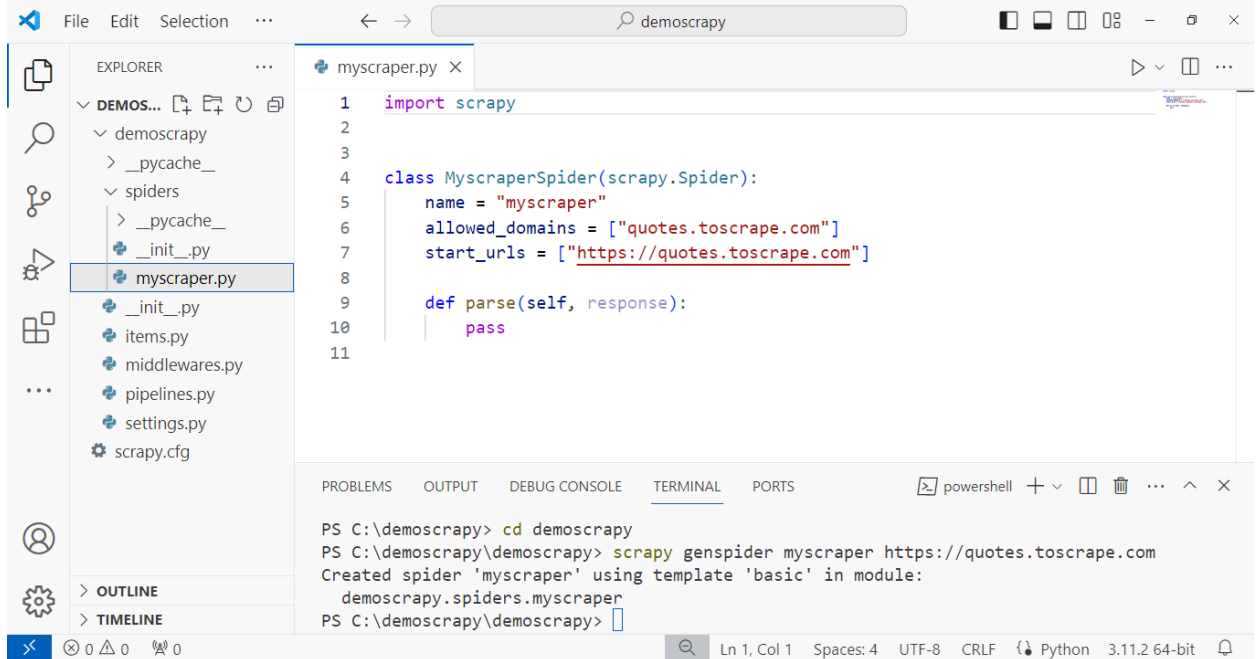
4. Xem cấu trúc project đã được tạo bằng VS Code



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



6. Kết quả sau khi tạo xong



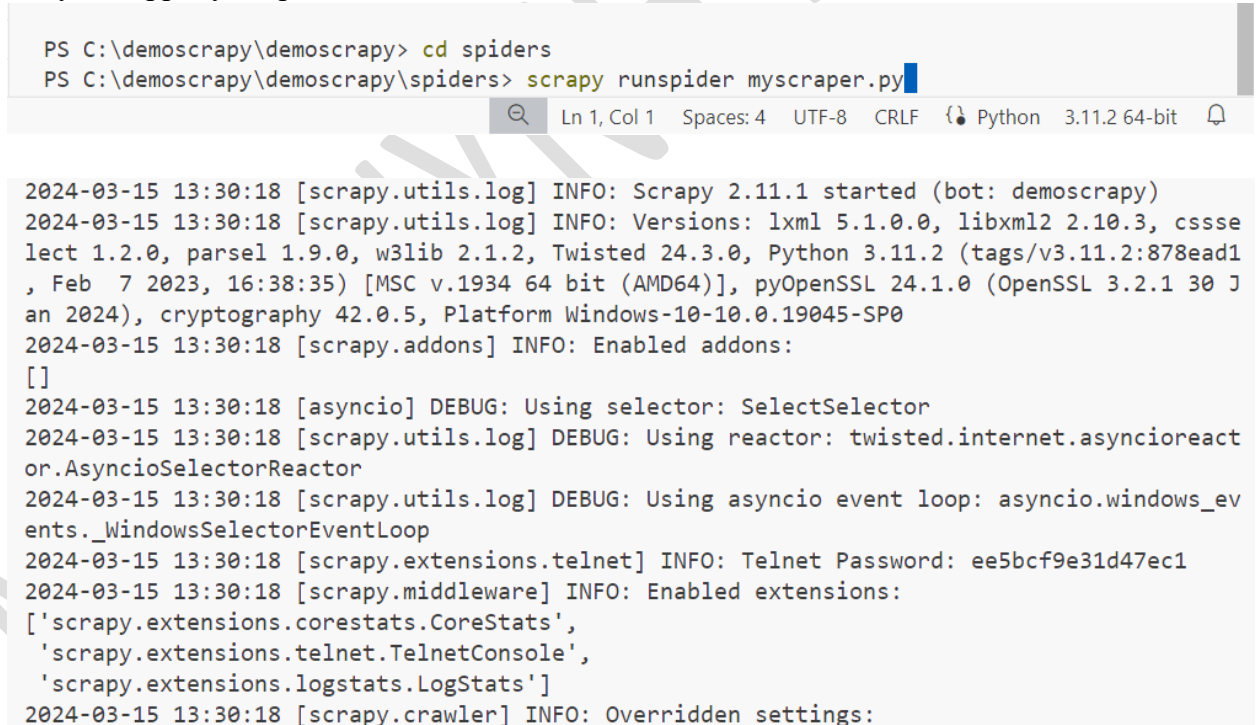
The screenshot shows the VS Code editor with the file explorer on the left displaying the project structure. The file `myscraper.py` is selected. The editor window shows the following Python code:

```
1 import scrapy
2
3
4 class MyscraperSpider(scrapy.Spider):
5     name = "myscraper"
6     allowed_domains = ["quotes.toscrape.com"]
7     start_urls = ["https://quotes.toscrape.com"]
8
9     def parse(self, response):
10         pass
11
```

The terminal at the bottom shows the execution of the spider:

```
PS C:\democrap> cd democrap
PS C:\democrap\democrap> scrapy genspider myscraper https://quotes.toscrape.com
Created spider 'myscraper' using template 'basic' in module:
democrap.spiders.myscraper
PS C:\democrap\democrap>
```

7. Chạy thử app myscraper



The screenshot shows a PowerShell terminal window with the following commands and output:

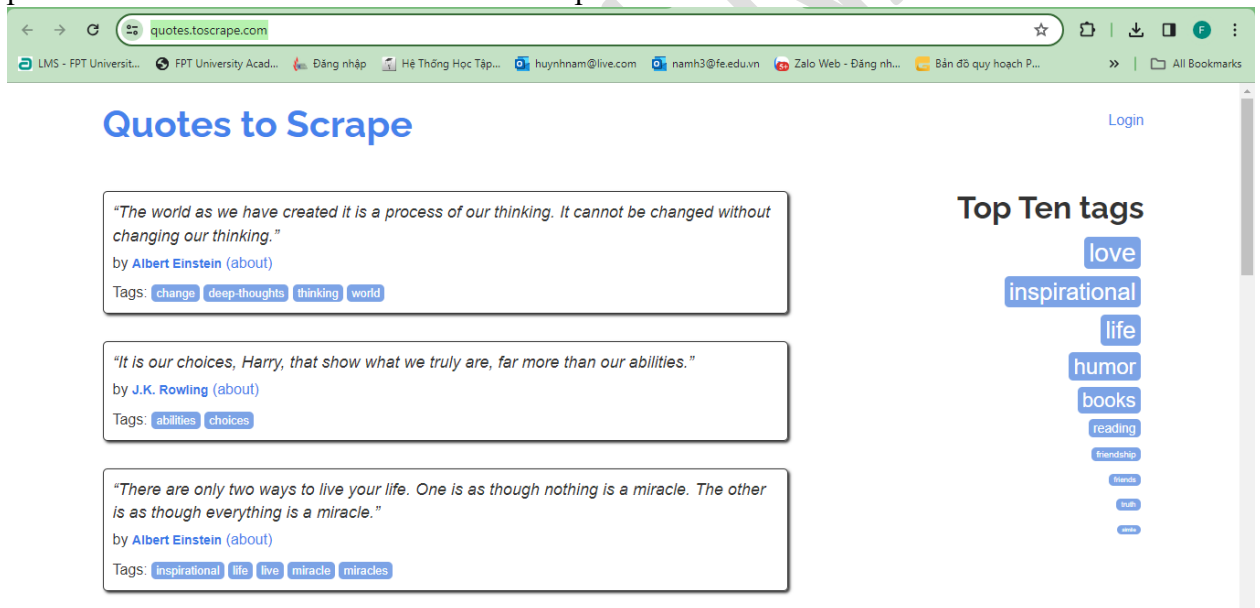
```
PS C:\democrap\democrap> cd spiders
PS C:\democrap\democrap\spiders> scrapy runspider myscraper.py
```

The output of the command is as follows:

```
2024-03-15 13:30:18 [scrapy.utils.log] INFO: Scrapy 2.11.1 started (bot: democrap)
2024-03-15 13:30:18 [scrapy.utils.log] INFO: Versions: lxml 5.1.0.0, libxml2 2.10.3, cssselect 1.2.0, parsel 1.9.0, w3lib 2.1.2, Twisted 24.3.0, Python 3.11.2 (tags/v3.11.2:878ead1, Feb 7 2023, 16:38:35) [MSC v.1934 64 bit (AMD64)], pyOpenSSL 24.1.0 (OpenSSL 3.2.1 30 Jan 2024), cryptography 42.0.5, Platform Windows-10-10.0.19045-SP0
2024-03-15 13:30:18 [scrapy.addons] INFO: Enabled addons:
[]
2024-03-15 13:30:18 [asyncio] DEBUG: Using selector: SelectSelector
2024-03-15 13:30:18 [scrapy.utils.log] DEBUG: Using reactor: twisted.internet.asyncioreactor.AsyncioSelectorReactor
2024-03-15 13:30:18 [scrapy.utils.log] DEBUG: Using asyncio event loop: asyncio.windows_events._WindowsSelectorEventLoop
2024-03-15 13:30:18 [scrapy.extensions.telnet] INFO: Telnet Password: ee5bcf9e31d47ec1
2024-03-15 13:30:18 [scrapy.middleware] INFO: Enabled extensions:
['scrapy.extensions.corestats.CoreStats',
'scrapy.extensions.telnet.TelnetConsole',
'scrapy.extensions.logstats.LogStats']
2024-03-15 13:30:18 [scrapy.crawler] INFO: Overridden settings:
```

```
'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

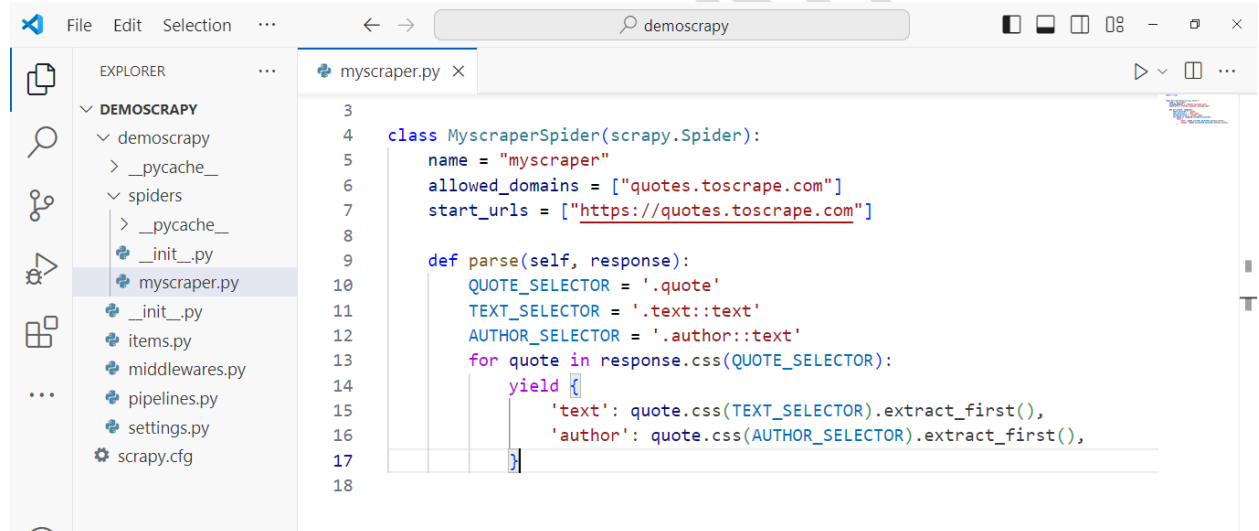


```

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  <div class="row">
25    <div class="col-md-8">
26      <div class="quote" itemscope itemtype="http://schema.org/CreativeWork">
27        <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>
28        <span by <small class="author" itemprop="author">Albert Einstein</small>
29          <a href="/author/Albert-Einstein">(about)</a>
30        </span>
31        <div class="tags">
32          <meta class="keywords" itemprop="keywords" content="change,deep-thoughts,thinking,world" />
33        </div>
34      </div>
35    </div>
36  </div>
37  </div>
38  </div>

```

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



```

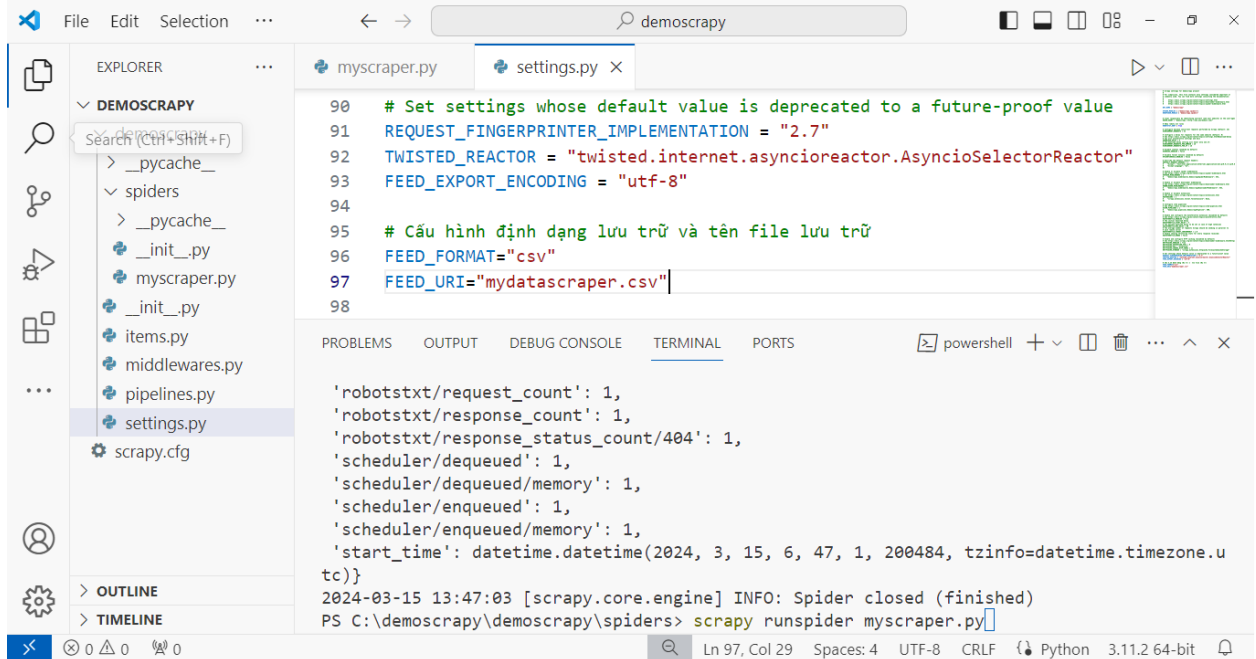
File Edit Selection ...
democracy
myscraper.py
3
4 class MyscraperSpider(scrapy.Spider):
5     name = "myscraper"
6     allowed_domains = ["quotes.toscrape.com"]
7     start_urls = ["https://quotes.toscrape.com"]
8
9     def parse(self, response):
10        QUOTE_SELECTOR = '.quote'
11        TEXT_SELECTOR = '.text::text'
12        AUTHOR_SELECTOR = '.author::text'
13        for quote in response.css(QUOTE_SELECTOR):
14            yield {
15                'text': quote.css(TEXT_SELECTOR).extract_first(),
16                'author': quote.css(AUTHOR_SELECTOR).extract_first(),
17            }
18

```

10. Tiến hành chạy app cào dữ liệu

PS C:\democrapy\democrapy\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

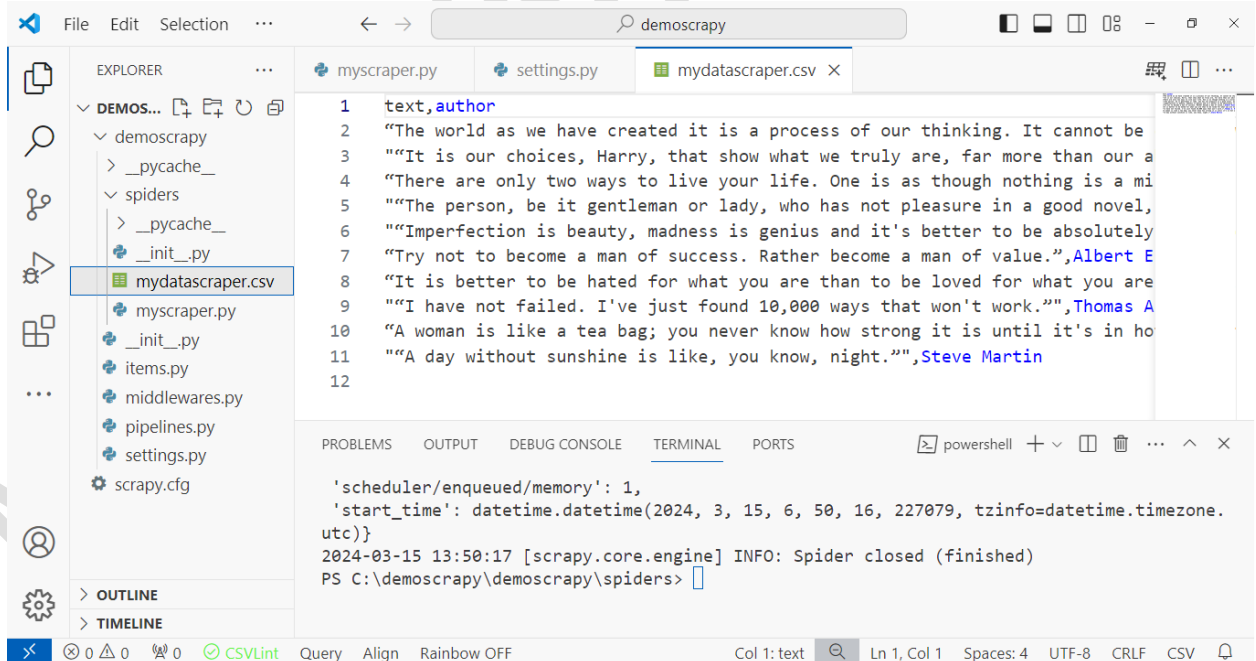


The screenshot shows the VS Code interface with the `settings.py` file open. The file contains settings for a scrapy spider, including `REQUEST_FINGERPRINTER_IMPLEMENTATION = "2.7"`, `TWISTED_REACTOR = "twisted.internet.asyncioreactor.AsyncioSelectorReactor"`, and `FEED_URI = "mydatascraper.csv"`. The terminal shows the output of the command `scrapy runspider myscraper.py`, indicating that the spider is closed (finished).

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

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

Kết quả cào dữ liệu



The screenshot shows the VS Code interface with the `mydatascraper.csv` file open. The file contains the scraped data, including the text and author of the page. The terminal shows the output of the command `scrapy runspider myscraper.py`, indicating that the spider is closed (finished).

THAM KHẢO

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