Test Automation

Experiment 3

Scrapy scrape quotes

Submitted by :   Submitted to :

Akhand Pratap Singh   Mr. Alind

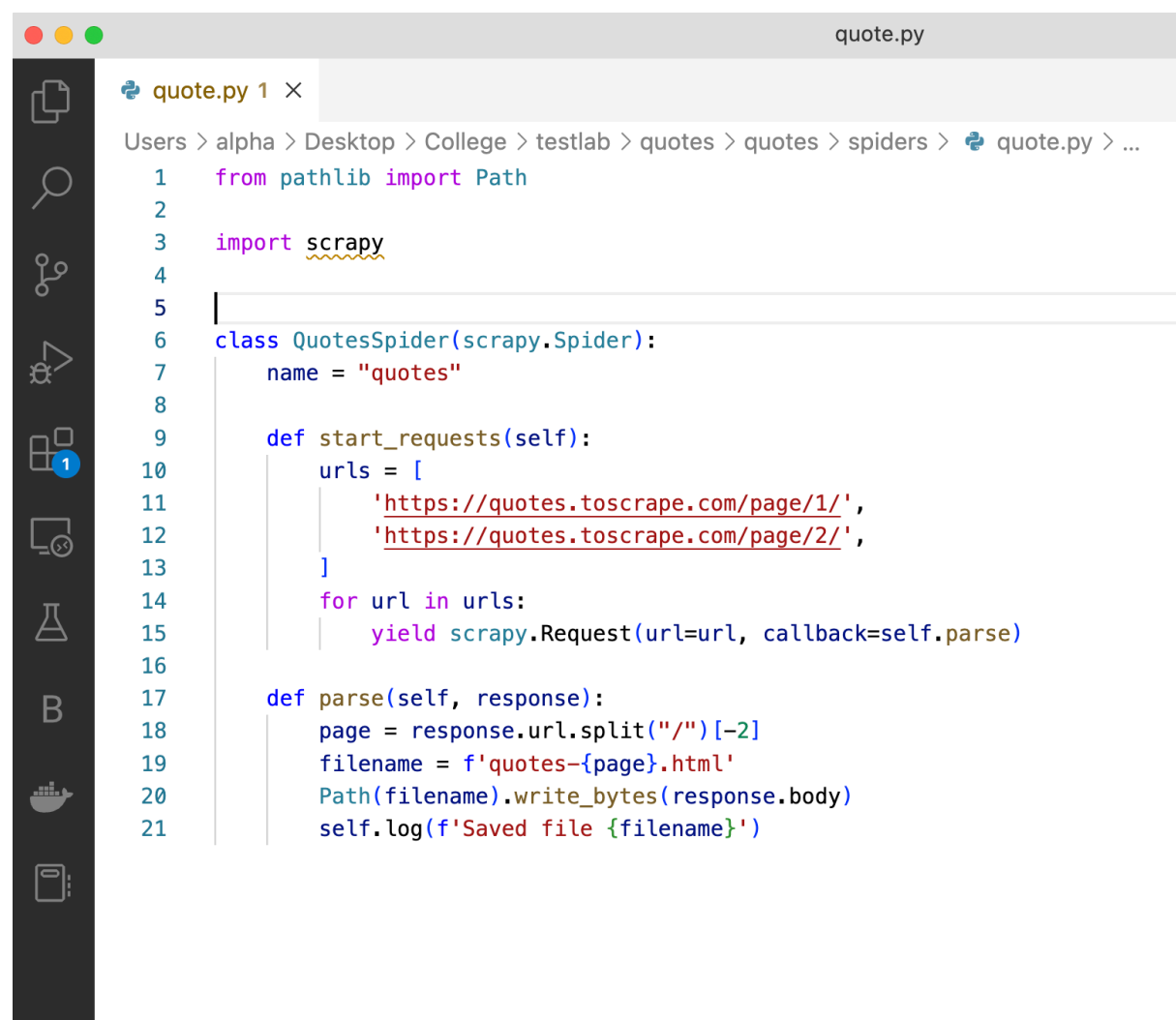**Step 1**: Activate environment:

```
[alpha@Akhands-MacBook-Pro ~ % conda activate base
(base) alpha@Akhands-MacBook-Pro ~ %
```

Step 2: Create a project

Command :- scrapy startproject project_name

```
(primeno) alpha@Akhands-MacBook-Pro testlab % scrapy startproject quotes
New Scrapy project 'quotes', using template directory '/Users/alpha/opt/anaconda3/envs/primeno/lib/python3.8/site-packages/scrapy/templates/project', created
in:
    /Users/alpha/Desktop/College/testlab/quotes

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

Step 3: Create python file named quotes_spider.py in

quote.py

quote.py 1 ×

Users > alpha > Desktop > College > testlab > quotes > quotes > spiders > quote.py > ...

```python
from pathlib import Path

import scrapy


class QuotesSpider(scrapy.Spider):
    name = "quotes"

    def start_requests(self):
        urls = [
            'https://quotes.toscrape.com/page/1/',
            'https://quotes.toscrape.com/page/2/',
        ]
        for url in urls:
            yield scrapy.Request(url=url, callback=self.parse)

    def parse(self, response):
        page = response.url.split("/")[-2]
        filename = f'quotes-{page}.html'
        Path(filename).write_bytes(response.body)
        self.log(f'Saved file {filename}')
```

Step 3: Run the script

```
(primeno) alpha@Akhands-MacBook-Pro quotes % scrapy crawl quotes
2023-02-13 22:41:04 [scrapy.utils.log] INFO: Scrapy 2.6.2 started (bot: quotes)
2023-02-13 22:41:04 [scrapy.utils.log] INFO: Versions: lxml 4.9.1.0, libxml2 2.9.14, cssselect 1.1.0, parsel 1.6
| packaged by conda-forge | (default, Mar  5 2020, 16:54:44) - [Clang 9.0.1 ], pyOpenSSL 19.1.0 (OpenSSL 1.1.1s
S-10.16-x86_64-i386-64bit
2023-02-13 22:41:04 [scrapy.crawler] INFO: Overridden settings:
{'BOT_NAME': 'quotes',
 'NEWSPIDER_MODULE': 'quotes.spiders',
 'ROBOTSTXT_OBEY': True,
 'SPIDER_MODULES': ['quotes.spiders']}
2023-02-13 22:41:04 [scrapy.utils.log] DEBUG: Using reactor: twisted.internet.selectreactor.SelectReactor
2023-02-13 22:41:04 [scrapy.extensions.telnet] INFO: Telnet Password: 618b0a9d097645df
2023-02-13 22:41:04 [scrapy.middleware] INFO: Enabled extensions:
['scrapy.extensions.corestats.CoreStats',
 'scrapy.extensions.telnet.TelnetConsole',
 'scrapy.extensions.memusage.MemoryUsage',
 'scrapy.extensions.logstats.LogStats']
2023-02-13 22:41:04 [py.warnings] WARNING: /Users/alpha/opt/anaconda3/envs/primeno/lib/python3.8/site-packages/r
ing: urllib3 (1.26.14) or chardet (3.0.4) doesn't match a supported version!
  warnings.warn("urllib3 ({}) or chardet ({}) doesn't match a supported "

2023-02-13 22:41:04 [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']
2023-02-13 22:41:04 [scrapy.middleware] INFO: Enabled spider middlewares:
['scrapy.spidermiddlewares.httperror.HttpErrorMiddleware',
 'scrapy.spidermiddlewares.offsite.OffsiteMiddleware',
 'scrapy.spidermiddlewares.referer.RefererMiddleware',
 'scrapy.spidermiddlewares.urllength.UrlLengthMiddleware',
 'scrapy.spidermiddlewares.depth.DepthMiddleware']
2023-02-13 22:41:04 [scrapy.middleware] INFO: Enabled item pipelines:
[]
2023-02-13 22:41:04 [scrapy.core.engine] INFO: Spider opened
2023-02-13 22:41:04 [scrapy.extensions.logstats] INFO: Crawled 0 pages (at 0 pages/min), scraped 0 items (at 0 i
2023-02-13 22:41:04 [scrapy.extensions.telnet] INFO: Telnet console listening on 127.0.0.1:6024
2023-02-13 22:41:10 [scrapy.core.engine] DEBUG: Crawled (404) <GET https://quotes.toscrape.com/robots.txt> (refe
2023-02-13 22:41:16 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://quotes.toscrape.com/page/1/> (referer
2023-02-13 22:41:04 [scrapy.core.engine] INFO: Spider opened
2023-02-13 22:41:04 [scrapy.extensions.logstats] INFO: Crawled 0 pages (at 0 pages/min), scraped 0 items (at 0 items/min)
2023-02-13 22:41:04 [scrapy.extensions.telnet] INFO: Telnet console listening on 127.0.0.1:6024
2023-02-13 22:41:10 [scrapy.core.engine] DEBUG: Crawled (404) <GET https://quotes.toscrape.com/robots.txt> (referer: None)
2023-02-13 22:41:16 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://quotes.toscrape.com/page/1/> (referer: None)
2023-02-13 22:41:16 [quotes] DEBUG: Saved file quotes-1.html
2023-02-13 22:41:20 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://quotes.toscrape.com/page/2/> (referer: None)
2023-02-13 22:41:21 [quotes] DEBUG: Saved file quotes-2.html
2023-02-13 22:41:21 [scrapy.core.engine] INFO: Closing spider (finished)
2023-02-13 22:41:21 [scrapy.statscollectors] INFO: Dumping Scrapy stats:
{'downloader/request_bytes': 681,
 'downloader/request_count': 3,
 'downloader/request_method_count/GET': 3,
 'downloader/response_bytes': 25579,
 'downloader/response_count': 3,
 'downloader/response_status_count/200': 2,
 'downloader/response_status_count/404': 1,
 'elapsed_time_seconds': 16.16399,
 'finish_reason': 'finished',
 'finish_time': datetime.datetime(2023, 2, 13, 17, 11, 21, 31001),
 'log_count/DEBUG': 6,
 'log_count/INFO': 10,
 'log_count/WARNING': 1,
 'memusage/max': 59076608,
 'memusage/startup': 59068416,
 'response_received_count': 3,
 'robotstxt/request_count': 1,
 'robotstxt/response_count': 1,
 'robotstxt/response_status_count/404': 1,
 'scheduler/dequeued': 2,
 'scheduler/dequeued/memory': 2,
 'scheduler/enqueued': 2,
 'scheduler/enqueued/memory': 2,
 'start_time': datetime.datetime(2023, 2, 13, 17, 11, 4, 867011)}
2023-02-13 22:41:21 [scrapy.core.engine] INFO: Spider closed (finished)
(primeno) alpha@Akhands-MacBook-Pro quotes % cd
```