

东莞理工学院网络空间安全学院

实验报告

课程名称：Python 数据分析与应用

学期：2023 年秋季

实验名称	网络爬虫			实验序号	2
姓 名	皮昊旋	学 号	2021428010127	班 级	21 杨班
实验地点	8B403	实验日期	20231018	指导老师	丁烨
教师评语	-			实验成绩	-
				百分制	100
同组同学	无				

一、 实验目标

- 1.1 了解爬虫机制；
- 1.2 熟悉爬虫工具及代码；
- 1.3 使用爬虫工具，部署爬虫代码。

二、 实验条件

- 2.1 硬件条件：PC 兼容机或 Mac；
- 2.2 软件条件：Python、PyCharm 或 VS Code 等 IDE。

三、 实验内容

- 3.1 安装 Scrapy 或其他爬虫，或开启 SaaS 爬虫服务；
- 3.2 爬取名人名言：<http://quotes.toscrape.com/>；
- 3.3 加入分页识别机制；
- 3.4 爬取名人名言作者信息。

四、 实验作业及分析

- 4.1 实验过程

① 使用 `pip install -U scrapy` 安装相关框架

```

PS E:\python-data-analyse> pip install -U scrapy
Collecting scrapy
  Obtaining dependency information for scrapy from https://files.pythonhosted.org/packages/08/66/22ed9609df4b6d9
etadata
  Downloading Scrapy-2.11.0-py2.py3-none-any.whl.metadata (5.2 kB)
Collecting Twisted<23.8.0,>=18.9.0 (from scrapy)
  Downloading Twisted-22.10.0-py3-none-any.whl (3.1 MB)
    3.1/3.1 MB 2.9 MB/s eta 0:00:00
Collecting cryptography>=36.0.0 (from scrapy)
  Obtaining dependency information for cryptography>=36.0.0 from https://files.pythonhosted.org/packages/d7/78/2
37-abi3-win_amd64.whl.metadata
  Downloading cryptography-41.0.4-cp37-abi3-win_amd64.whl.metadata (5.3 kB)
Collecting cssselect>=0.9.1 (from scrapy)
  Downloading cssselect-1.2.0-py2.py3-none-any.whl (18 kB)
Collecting itemloaders>=1.0.1 (from scrapy)
  Downloading itemloaders-1.1.0-py3-none-any.whl (11 kB)
Collecting parsel>=1.5.0 (from scrapy)
  Downloading parsel-1.8.1-py2.py3-none-any.whl (17 kB)

```

② 编写爬虫代码

```

1 from typing import Any, Iterable
2 import scrapy
3 from scrapy.http import Request, Response
4
5 class QuotesSpider(scrapy.Spider):
6     name = "quotes"
7
8     def start_requests(self):
9         urls = ['http://quotes.toscrape.com/page/1', 'http://quotes.toscrape.com/page/2']
10        for url in urls:
11            yield scrapy.Request(url=url, callback=self.parse)
12
13    def parse_author_info(self, response):
14        yield {
15            'about': response.css('div.author-description::text').get()
16        }
17
18    def parse(self, response):
19        next_page = response.css('li.next a::attr(href)').get()
20        if next_page is not None:
21            yield response.follow(next_page, self.parse)
22
23        for quote in response.css('div.quote'):
24            author_info_page = 'http://quotes.toscrape.com' + quote.css('div.quote a::attr(href)').get()
25            yield Request(author_info_page, callback=self.parse_author_info, meta={'quote': quote})
26
27        page = response.url.split("/")[-2]
28        filename = 'quotes-%s.html' % page
29        with open(filename, "wb") as f:
30            f.write(response.body)
31        self.log("Saved file %s" % filename)
32
33        for quote in response.css('div.quote'):
34            author_info_page = 'http://quotes.toscrape.com' + quote.css('div.quote a::attr(href)').get()
35            yield Request(author_info_page, callback=self.parse_author_info, meta={'quote': quote})
36
37        page = response.url.split("/")[-2]
38        filename = 'quotes-%s.html' % page
39        with open(filename, "wb") as f:
40            f.write(response.body)
41        self.log("Saved file %s" % filename)
42
43    def parse_author_info(self, response):
44        quote = response.meta['quote']
45        yield {
46            'text' : quote.css('span.text::text').get(),
47            'author' : quote.css('small.author::text').get(),
48            'tags' : quote.css('div.tags a.tag::text').getall(),
49            'about' : response.css('div.author-description::text').get(),
50        }
51

```

为了爬取作者详细信息，加入新函数

```
1 def parse_author_info(self, response):
2     yield {
3         'about': response.css('div.author-description::text').get()
4     }
```

专门获取作者页的信息

因为获取作者页的信息也需要 请求一次，所以在组织信息前用 meta 来记录请求间的关系。

③ 测试并完善爬虫代码

```
'finish_time': datetime.datetime(2023, 10, 18, 1, 38, 26, 144025, tzinfo=datetime.timezone.utc),
'item_scraped_count': 20,
'log_count/DEBUG': 30,
'log_count/INFO': 11,
'response_received_count': 3,
'robotstxt/request_count': 1,
'robotstxt/response_count': 1,
'robotstxt/response_status_count/404': 1,
'scheduler/dequeued': 4,
'scheduler/dequeued/memory': 4,
'scheduler/enqueued': 4,
'scheduler/enqueued/memory': 4,
'start_time': datetime.datetime(2023, 10, 18, 1, 38, 23, 944890, tzinfo=datetime.timezone.utc)}
2023-10-18 09:38:26 [scrapy.core.engine] INFO: Spider closed (finished)
PS E:\python-data-analyse\lab2\tutorial> |
```

4.2 实验结果

```
{
  "text": "\"A day without sunshine is like, you know, night.\"\"", "author": "Steve Martin", "tags": ["humor", "obvious", "simile"], "about": "\n\n Stephen Glenn \"S
  "text": "\"The world as we have created it is a process of our thinking. It cannot be changed without changing our thinking.\"\"", "author": "Albert Einstein", "tags": [
  "text": "\"It is our choices, Harry, that show what we truly are, far more than our abilities.\"\"", "author": "J.K. Rowling", "tags": ["abilities", "choices"], "about":
  "text": "\"Imperfection is beauty, madness is genius and it's better to be absolutely ridiculous than absolutely boring.\"\"", "author": "Marilyn Monroe", "tags": ["be-y
  "text": "\"The person, be it gentleman or lady, who has not pleasure in a good novel, must be intolerably stupid.\"\"", "author": "Jane Austen", "tags": ["aliteracy", "b
  "text": "\"I have not failed. I've just found 10,000 ways that won't work.\"\"", "author": "Thomas A. Edison", "tags": ["edison", "failure", "inspirational"], "paraphrase
  "text": "\"It is better to be hated for what you are than to be loved for what you are not.\"\"", "author": "Andr  Gide", "tags": ["life", "love"], "about": "\n\n A
  "text": "\"A woman is like a tea bag; you never know how strong it is until it's in hot water.\"\"", "author": "Eleanor Roosevelt", "tags": ["misattributed-eleanor-roose
  "text": "\"Life is what happens to us while we are making other plans.\"\"", "author": "Allen Saunders", "tags": ["fate", "life", "misattributed-john-lennon", "planning"
  "text": "\"Good friends, good books, and a sleepy conscience: this is the ideal life.\"\"", "author": "Mark Twain", "tags": ["books", "contentment", "friends", "friendsh
  "text": "\"It is not a lack of love, but a lack of friendship that makes unhappy marriages.\"\"", "author": "Friedrich Nietzsche", "tags": ["friendship", "lack-of-friend
  "text": "\"The opposite of love is not hate, it's indifference. The opposite of art is not ugliness, it's indifference. The opposite of faith is not heresy, it's indi
  "text": "\"I may not have gone where I intended to go, but I think I have ended up where I needed to be.\"\"", "author": "Douglas Adams", "tags": ["life", "navigation"],
  "text": "\"I like nonsense, it wakes up the brain cells. Fantasy is a necessary ingredient in living.\"\"", "author": "Dr. Seuss", "tags": ["fantasy"], "about": "\n\n
  "text": "\"You may not be her first, her last, or her only. She loved before she may love again. But if she loves you now, what else matters? She's not perfect-you ar
```

About 内容是作者页的详细内容

```

1 actor, comedian, writer, playwright, producer, musician, and composer. He was raised in Southern California in a Baptist family, where his early influences were wa
2 thinking", "world"]], "about": "\n In 1879, Albert Einstein was born in Ulm, Germany. He completed his Ph.D. at the University of Zurich by 1909. His 1905 pap
3 gatoratchatrough she writes under the pen name J.K. Rowling, pronounced like rolling, her name when her first Harry Potter book was published was simply Joanne
4 "about": "\n Marilyn Monroe (born Norma Jeane Mortenson; June 1, 1926 - August 5, 1962) was an American actress, model, and singer, who became a major sex s
5 "about": "\n Jane Austen was an English novelist whose works of romantic fiction, set among the landed gentry, earned her a place as one of the most widely
6 mas Alva Edison was an American inventor, scientist and businessman who developed many devices that greatly influenced life around the world, including the phonogr
7
8
9 Anna Eleanor Roosevelt was an American political leader who used her influence as an active First Lady from 1933 to 1945 to promote the New Deal policies of her hus
10 Allen Saunders was an American writer, journalist and cartoonist who wrote the comic strips Steve Roper and Mike Nomad, Mary Worth and Kerry Drake. His full nam
11 Samuel Langhorne Clemens, better known by his pen name Mark Twain, was an American author and humorist. He is noted for his novels Adventures of Huckleberry F
12 ", "marriage", "unhappy-marriage"]], "about": "\n Friedrich Wilhelm Nietzsche (1844-1900) is a German philosopher of the late 19th century who challenged the
13 f life is not death, it's indifference."], "author": "Elie Wiesel", "tags": ["activism", "apathy", "hate", "indifference", "inspirational", "love", "opposite", "phi
14 Noël Adams was an English author, comic radio dramatist, and musician. He is best known as the author of the Hitchhiker's Guide to the Galaxy series. Hitchhiker's
15 born 2 March 1904 in Springfield, MA. He graduated Dartmouth College in 1925, and proceeded on to Oxford University with the intent of acquiring a doctorate in lite
16 you may never be perfect together but if she can make you laugh, cause you to think twice, and admit to being human and making mistakes, hold onto her and give her
17 ut": "\n James Maury \"Jim\" Henson was the most widely known puppeteer in American television history. He was the creator of The Muppets and the leading for
18 ], "about": "\n Garrison Keillor (born Gary Edward Keillor on August 7, 1942 in Anoka, Minnesota) is an American author, storyteller, humorist, columnist, mu
19 orn Agnes Gonxha Bojaxhiu was an Albanian Roman Catholic nun who founded the Missionaries of Charity in Kolkata (Calcutta), India in 1950. For over forty years she
20 which there is no I or you, so intimate that your hand upon my chest is my hand, so intimate that when I fall asleep your eyes close."], "author": "Pablo Neruda",
21 Acevedo (Spanish pronunciation: [ˈxoɾxe ˈlwis ˈboɾxɐs], Russian: Хорхе Луис Борхес) was an Argentine writer and poet born in Buenos Aires. In 1914, his family moved to
22 as born at a farmstead in Nuneaton, Warwickshire, England, where her father was estate manager. Mary Ann, the youngest child and a favorite of her father's, receive
23 in the GoodReads database with this name. See this thread for more information. William Nicholson was born in 1948, and grew up in Sussex and Gloucestershire. His
24 hulz was an American cartoonist, whose comic strip Peanuts proved one of the most popular and influential in the history of the medium, and is still widely reprinte
25 n in Boston. Educated at Harvard and the Cambridge Divinity School, he became a Unitarian minister in 1826 at the Second Church Unitarian. The congregation, with Ch
26 E STAPLES LEWIS (1898-1963) was one of the intellectual giants of the twentieth century and arguably one of the most influential writers of his day. He was a Fellow

```

五、实验总结

本次实验使用 Scrapy 框架编写了一个爬虫程序，从 <http://quotes.toscrape.com> 上爬取名言和作者信息，并实现了将名、作者、tags 和作者简介进行保存的功能。

在实验过程中，首先定义了一个 QuotesSpider 的 Spider 类。在 start_requests 方法中，设置了要爬取的页面 URL 列表，并通过循环为每个 URL 发起请求，指定回调函数为 parse 方法。

在 parse 方法中，先获取下一页的链接，如果存在下一页，则继续发起请求。然后，利用 CSS 选择器提取出每个名言的文本、作者、标签等信息，并通过 yield 语句返回一个包含这些信息的字典。然后对每个名言提取到的链接发起请求，以获取作者信息页面的内容。

在新的回调函数 parse_author_info 中，通过 response.meta 获取到之前存入的名言对象，然后从响应中提取作者简介信息，并将其添加到之前的名言字典中。最后，再次通过 yield 语句返回完整的名言信息。

本次实验通过 Scrapy 框架实现了对名言网站的爬取，提取了所需的名言和作者信息，并将作者简介作为名言的一部分进行保存。通过这次实验，我对 Scrapy 的使用有了更深入的了解，并掌握了如何在爬虫程序中处理多个请求的数据关系，以及如何将爬取到的数据进行保存和后续处理。