D26：Scrapy 爬蟲編寫 (1)

簡報閱讀 範例與作業 問題討論

xpath & 基本爬蟲介紹

本日知識點目標

- 上次我們成功地建立了爬蟲專案以及第一隻爬蟲程式
- 這次我們要使用scrapy將我們要的資料打印出來

在spiders資料夾中你可以看到你的爬蟲程式以scrapy指令自動建立
在程式中你可以看到：

name = 'ettoday'
這是你爬蟲的名字，之後你要執行這隻爬蟲時，你會需要告訴scrapy你要執行的是哪隻爬蟲

allowed_domains = ['www.ettoday.net']

這是你這隻爬蟲被允許請求的domain，他是一個list，也就是說當你需要請求的domain不只一個時，你可以將該domain加入allowed_domains裡

start_urls = ['http://www.ettoday.net/']

你的爬蟲起始網站，他也是一個list，你爬蟲的第一步會先對這個list內的全部網址依序進行請求，你可以依照你的需求添加url進start_urls

function parse

當你對一個url進行請求後，你需要對返回的response進行操作，而parse這個function就是這隻爬蟲默認進行處理response的function，若你沒有特別指定處理response的function，則由parse這個function進行處理

```python
# -*- coding: utf-8 -*-
import scrapy


class EttodaySpider(scrapy.Spider):
    name = 'ettoday'
    allowed_domains = ['www.ettoday.net']
    start_urls = ['http://www.ettoday.net/']

    def parse(self, response):
        pass

```

常見的定位器

- 當我們要進行爬蟲時，常見的定位器有兩種，分別是css selector 和 xpath

- 假設我們要取得回傳的response中title這個標籤裡的文字，以 css selector 及 xpath 分別回傳第一個匹配以己全部匹配資料

	css selector	xpath
第一個匹配的資料	response.css('title::text').get()	response.xpath('//title/text()').get()
全部匹配的資料	response.css('title::text').getall()	response.xpath('//title/text()').getall()

完成一隻簡單的爬蟲

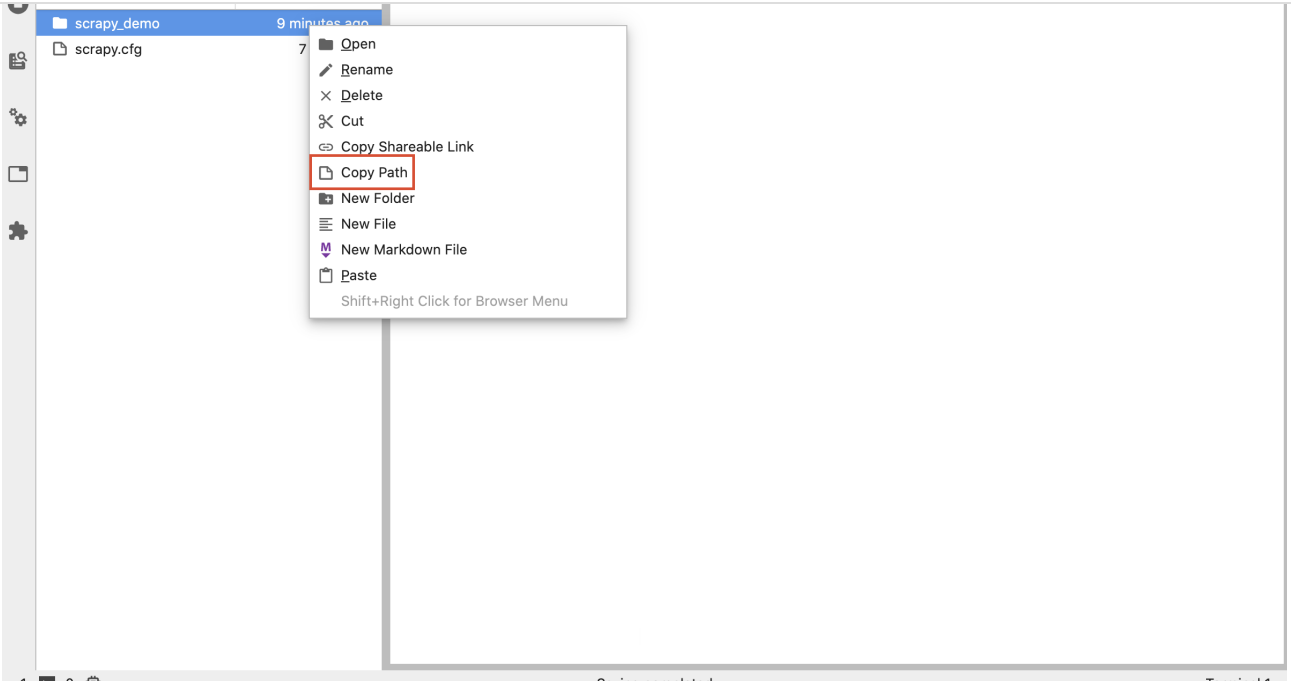
- 那麼，知道如何定位資料後，我們開始試著完成一隻簡單的爬蟲
- 首先將我們要爬取的url 加入start_urls中，並將你要對返回的response要進行的操作放入parse function中，你的爬蟲第一步就算完成了
- 假設我們要爬取某的文章的標題以及內文，並且將爬到的資料print出來，那麼你爬蟲的程式應該會類似這樣

```python
class EttodaySpider(scrapy.Spider):
    name = 'ettoday'
    allowed_domains = ['www.ettoday.net']
    start_urls = ['目標新聞1', '目標新聞2', '目標新聞3']

    def parse(self, response):
        title = response.xpath('/xpath/of/title').get()
        content = response.xpath('/xpath/of/content').getall()
        print(title)
        print(content)
```

Saving completed Ln 14, Col 22 Spaces: 4 ettoday.py

- 那麼爬蟲編輯完之後要怎麼執行呢？
- scrapy要執行爬蟲的時候需要使用terminal，先開啟terminal 並移動到爬蟲專案的資料夾中
- 如果不知道路徑是什麼，可以對該資料夾點擊右鍵並選擇copy path



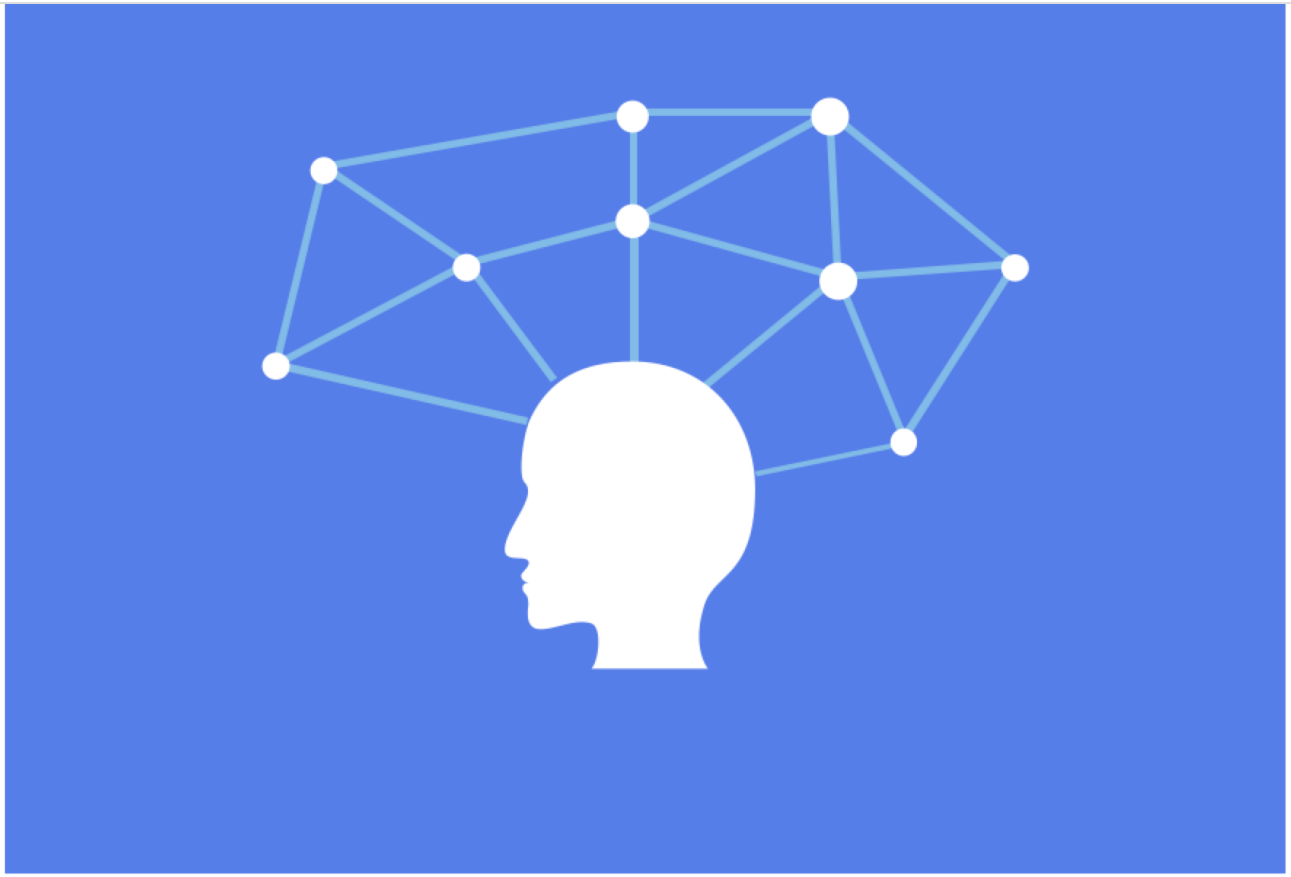
- 到該資料夾底下之後，執行指令 scrapy crawl spider_name
- 這個spider_name是你要執行的爬蟲的名稱，範例爬蟲名稱是ettoday，因此執行指令scrapy crawl ettoday

執行指令後

執行指令後，你會看到scrapy 開始執行該爬蟲，並印出log
此時你會看到你要print出來的資料
我這邊是以https://www.ettoday.net/news/20201004/1824032.htm做示範

```
ferer: None)
2020-10-04 18:29:59 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.ettoday.net/news/20201004/1824032.htm> (referer: None)
雙十國慶節目三大亮點曝光　內行攻略看這裡 | ETtoday政治新聞 | ETtoday新聞雲
['記者蕭羽耘／台北報導', '這個周末將迎來國慶大會，2020年籌備委員會規劃三大亮點，一早空軍「雷虎小組」軍機分列式表演，將開放500民眾觀禮，晚間還有台南漁光島長達33分鐘的煙火秀，以及總統府建築投影光雕，系列活動屆時將讓民眾「大飽眼福」。', '國慶日系列活動吸睛，凱道一早舉行大典，將由空軍「雷虎小組」進行軍機分列式表演，陸軍的CH-47直升機和空軍的UH-60M直升機，也會先後飛越總統府上空，現場開放500名民眾觀禮。', '由於近期共機頻擾台，導致24日國慶預演時，戰機轟轟作響讓民眾恐慌，當時總統蔡英文甚至發文一解誤會，「今早台北市區的鬧鐘有點不一樣，這是國慶預演戰機飛行的聲音。」', '國慶晚間重頭戲還有煙火與總統府光雕，國慶煙火選在台南漁光島施放，總計長達33分鐘、共1萬7000發，為讓民眾有最佳的觀賞煙火體驗，台南市政府規劃多個場域，包括主舞台的安平戀愛廣場，副舞台的觀夕平台、1661台灣船園區、四鯤鯓（龍崗天橋停車場）等4個場地。', '至於光雕部分，由文化總會參與設計，以總統府為背景投放「口罩」、「豬肉」等圖樣，欲傳達這些成功的「台灣模式」，且象徵在動盪不安的國際情勢中，台灣仍能在黑暗中發出明亮的光，為國人指引方向，屹立風雨中守護國人，相當吸睛。']
2020-10-04 18:29:59 [scrapy.core.engine] INFO: Closing spider (finished)
2020-10-04 18:29:59 [scrapy.statscollectors] INFO: Dumping Scrapy stats:
{'downloader/request_bytes': 463,
 'downloader/request_count': 2,
 'downloader/request_method_count/GET': 2,
 'downloader/response_bytes': 42297,
 'downloader/response_count': 2,
 'downloader/response_status_count/200': 2,
 'finish_reason': 'finished',
 'finish_time': datetime.datetime(2020, 10, 4, 10, 29, 59, 280416),
 'log_count/DEBUG': 2,
 'log_count/INFO': 9,
 'memusage/max': 48828416,
 'memusage/startup': 48828416,
 'response_received_count': 2,
 'robotstxt/request_count': 1,
 'robotstxt/response_count': 1,
 'robotstxt/response_status_count/200': 1,
 'scheduler/dequeued': 1,
 'scheduler/dequeued/memory': 1,
 'scheduler/enqueued': 1,
 'scheduler/enqueued/memory': 1,
 'start_time': datetime.datetime(2020, 10, 4, 10, 29, 59, 19128)}
2020-10-04 18:29:59 [scrapy.core.engine] INFO: Spider closed (finished)
```

重要知識點複習



1. 瞭解如何在scrapy中使用 xpath 以及 css selector 對資料進行定位
2. 了解默認產生的爬蟲程式架構以及簡易編寫流程
3. 如何執行scrapy的爬蟲程式

解題時間

下一步：閱讀範例與完成作業