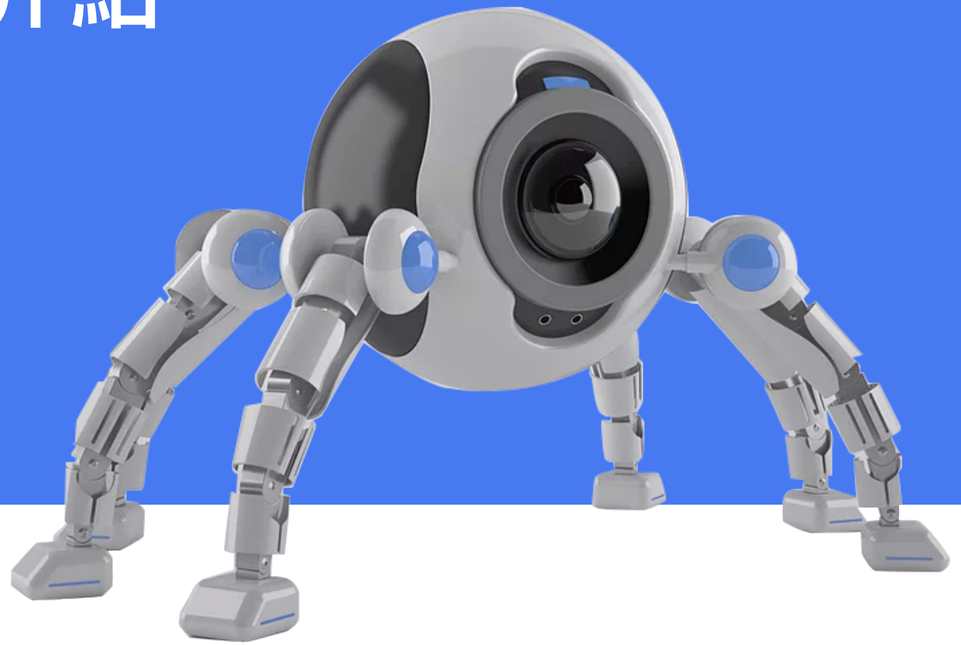
Day 28 SCRAPY 爬蟲框架介紹



載入 Scrapy 爬蟲

出題教練：楊鎮銘

python

本日知識點目標

- 了解如何在 .py 檔裏面操作 scrapy 爬蟲
- 了解如何彈性的透過各種參數控制爬蟲行為

外部執行 Scrapy 爬蟲程序

我們在前面介紹如何執行 Scrapy 的爬蟲都是透過命令列

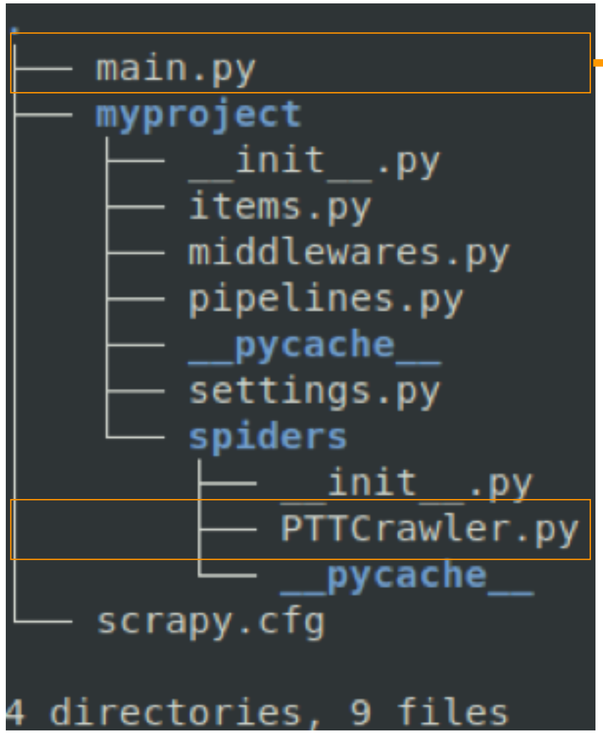
```
$ scrapy crawl PTTCrawler
```

但其實框架本身有提供 API

讓我們可以從外部去呼叫並執行爬蟲甚至是其他元件，這樣可以方便我們串聯其他非框架本身或是沒有提供的功能

外部執行 Scrapy 爬蟲程序

在這邊我們會透過 `scrapy.crawler.CrawlerProcess` 來執行爬蟲



我們會在這裡呼叫 Scrapy 爬蟲程序 (CrawlerProcess)

作為練習：我們這邊可以將 PTTCrawler 裡的爬蟲修改為
「給定 PTT 文章網址就會自動把結果存成 JSON 的爬蟲」

然後在 main.py 中整理多個文章網址後再去呼叫爬蟲程式

外部執行 Scrapy 爬蟲程序

CUPOY

在 `main.py` 中我們可以透過 `CrawlerProcess` 建構爬蟲流程

```
import scrapy¬
from scrapy.crawler import CrawlerProcess¬
from scrapy.utils.project import get_project_settings¬
¬
¬
def main():¬
    target_urls = [¬
        'https://www.ptt.cc/bbs/Gossiping/M.1559788476.A.074.html',¬
        'https://www.ptt.cc/bbs/Gossiping/M.1557928779.A.0C1.html'¬
    ]¬
    process = CrawlerProcess(get_project_settings())¬
    process.crawl('PTTCrawler', start_urls=target_urls)¬
    process.start()¬
¬
if __name__ == '__main__':¬
    main()¬
```

呼叫 CrawlerProcess 就會建立爬蟲流程

假如框架中有不同爬蟲

可以在這邊給予不同設定
(e.g. 每個請求之間要 delay 幾秒)

或是透過 get_project_setting
取得 myproject 專案的全局設定 (global setting)

外部執行 Scrapy 爬蟲程序

CUPOY

在 `main.py` 中我們可以透過 `CrawlerProcess` 建構爬蟲流程

```
import scrapy¬
from scrapy.crawler import CrawlerProcess¬
from scrapy.utils.project import get_project_settings¬
¬
¬
def main():¬
    target_urls = [¬
        'https://www.ptt.cc/bbs/Gossiping/M.1559788476.A.074.html',¬
        'https://www.ptt.cc/bbs/Gossiping/M.1557928779.A.0C1.html'¬
    ]¬
    process = CrawlerProcess(get_project_settings())¬
    process.crawl('PTTCrawler', start_urls=target_urls)¬
    process.start()¬
¬
if __name__ == '__main__':¬
    main()¬
```

每個爬蟲都有一個全局唯一的名稱

可以直接透過這個名稱來決定要開始執行哪個爬蟲

後面可以給予爬蟲有定義的參數來改變爬蟲結果
(e.g. start_urls 是目標網頁的參數)

增加爬蟲可控制選項

前面我們介紹如何從外部控制爬蟲，但只能控制框架提供的選項

如果我們想要自己定義爬蟲可控制的參數，就必須要重新修改爬蟲

e.g. PTT 爬蟲希望可以讓外部可以決定存檔的檔名

增加爬蟲可控制選項

CUPOY

我們前面讓外部提供目標網址

所以這邊就不會設定預設目標網址

```
class PttcrawlerSpider(scrapy.Spider):¬
    name = 'PTTCrawler'¬
    allowed_domains = ['www.ptt.cc']¬
    start_urls = []¬
    def __init__(self, filename=None):¬
        self.colkies = {'over18': '1'}¬
        self.filename = filename¬
¬
    def start_requests(self):¬
```

這邊透過 Python class 物件的 __init__ function 來決定爬蟲初始化時可以從外部傳入什麼參數

將外部給予的參數存成 class 內部資訊

增加爬蟲可控制選項

CUPOY

當我們根據外部提供的參數增加了爬蟲內部的資訊

就可以在框架內的其他元件中參考並使用

e.g. 存檔檔名的客製化需要在 Item Pipeline 中做修改

```
class JSONPipeline(onject):
    …
    def close_spider(self, spider):
        print(spider.filename) // 透過這種方式可以取得爬蟲內部的其他資訊
        …                      // 再根據 spider.filename 決定存檔檔名
```

增加爬蟲可控制選項

CUPOY

當爬蟲與其相關對應的元件都修改完成後

就可以直接透過 `CrawlerProcess` 在外部給予其他參數

```
import scrapy¬
from scrapy.crawler import CrawlerProcess¬
from scrapy.utils.project import get_project_settings¬
¬
¬
def main():¬
    target_urls = [¬
        'https://www.ptt.cc/bbs/Gossiping/M.1559788476.A.074.html',¬
        'https://www.ptt.cc/bbs/Gossiping/M.1557928779.A.0C1.html'¬
    ]¬
    process = CrawlerProcess(get_project_settings())¬
    process.crawl('PTTCrawler', start_urls=target_urls, filename='test.json')¬
    process.start()¬
¬
if __name__ == '__main__':¬
    main()¬
```

給予自定義的其他參數

重要知識點複習

CUPOY

- 透過 API CrawlerProcess 我們可以在外部控制專案內的爬蟲
- 框架本身提供很多關於爬蟲細節的控制選項
- 可以根據需求修改爬蟲與相對應的元件從外部給予更多更彈性的控制項

延伸閱讀

- 初始化框架爬蟲
- CrawlerProcess
- 透過命令列給予可控制項參數
  - 我們這邊透過 API 在 main.py 中給予原本框架設定的參數跟我們自己定義的參數，其實也可以透過命令列給予這些參數

解題時間
LET'S CRACK IT

請跳出 PDF 至官網 Sample Code & 作業

開始解題

START