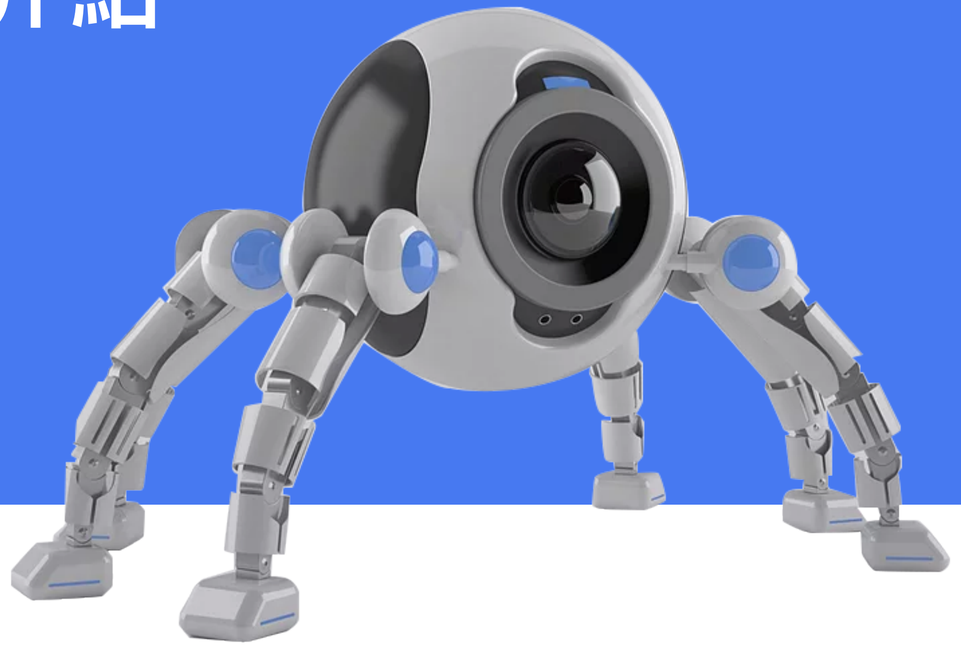
Day 29 SCRAPY 爬蟲框架介紹



Scrapy 多網頁爬蟲

出題教練：楊鎮銘

python

本日知識點目標

- 進一步了解 Scrapy 送請求的流程
- 合併理解多網頁策略在 Scrapy 上的應用

Scrapy 多網頁實作

我們目前的爬蟲功能是對「所有給予的 PTT 文章網址」進行爬蟲

實作 PTT 多網頁爬蟲的實作有兩個方向

- 外部決定網址 + 框架對給予網址進行爬蟲
  - 在外部 (e.g. main.py) 對文章列表進行爬蟲取得所有文章網址
  - 把所有文章網址傳入 scrapy 爬蟲
- 框架爬文章列表 + 文章內容

這兩種方式都可以，但是先從外部取得網址的方式會比較慢

這邊我們可以更深入了解框架送請求的過程為什麼會比較快

Scrapy 多網頁實作

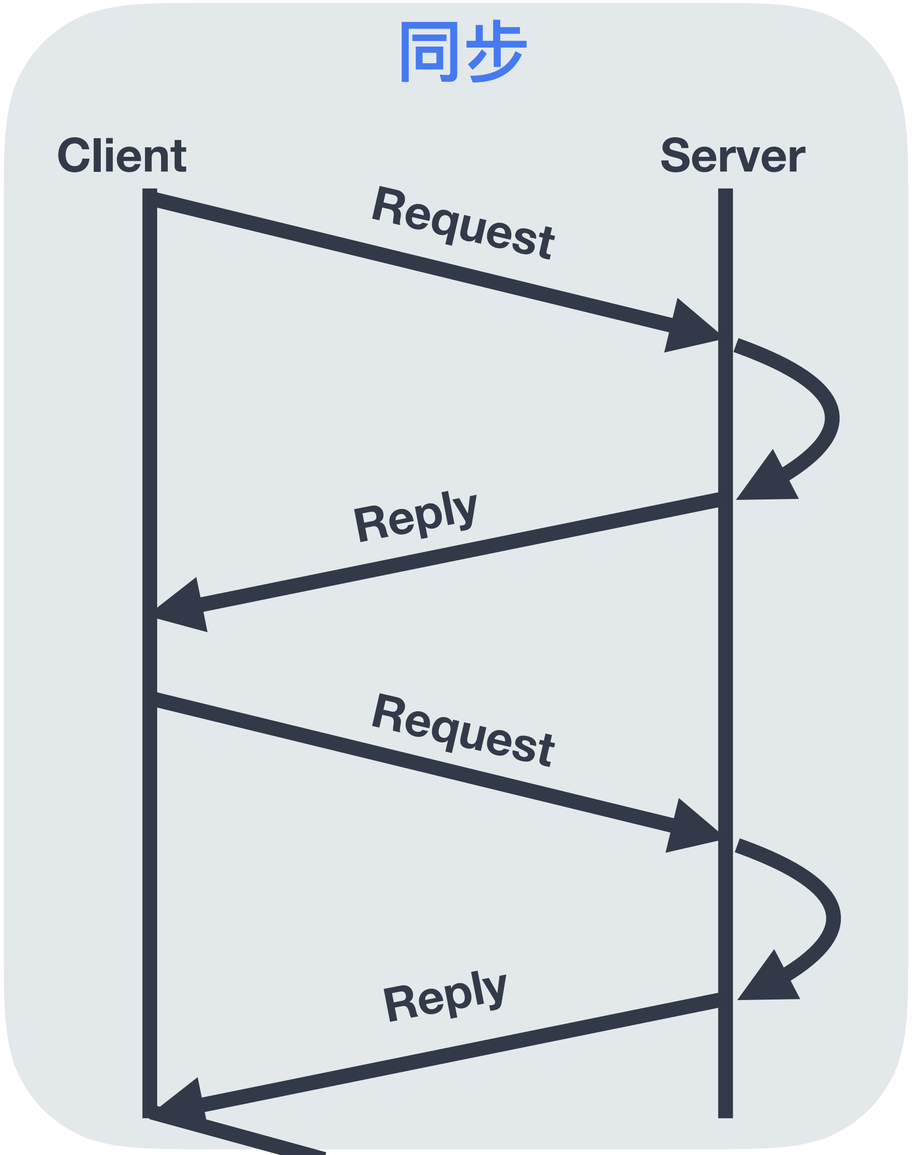
原本 requests 的方式，程式會送出第一個請求後會等到第一個 response 傳回來才會送第二個請求

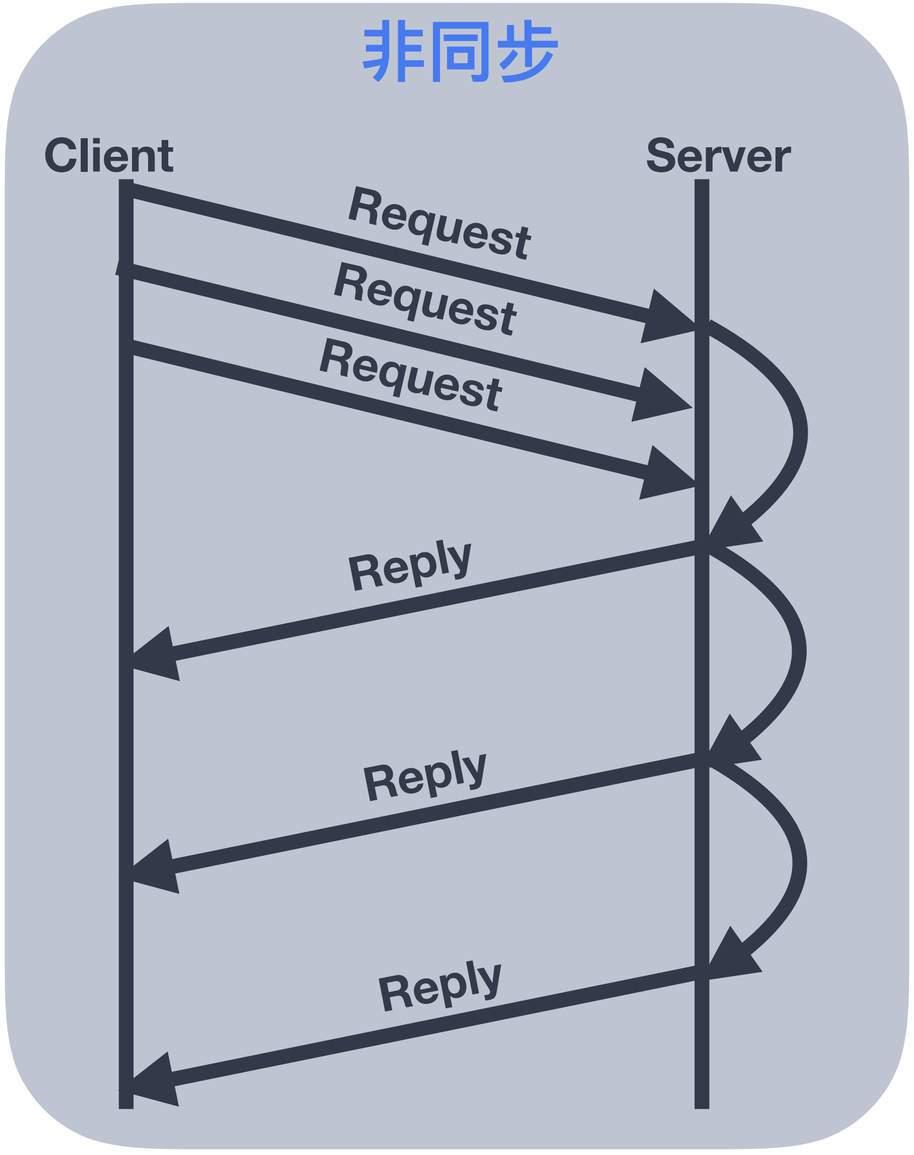
而框架內的請求方式 `yield scrapy.Request`
在送出第一個請求後會直接送第二個請求，並不會卡著等第一個 response，而是等第一個 response 送回來的時候再處理

這種方式可以縮短因為網路延遲造成的等待，加速整個爬蟲過程

Scrapy 多網頁實作

CUPOY





Scrapy 多網頁實作

CUPOY

```
class PttcrawlSpider(scrapy.spider):
```

`def __init__(self, …):` `...`	整理出要開始爬的最初目標網址 e.g. https://www.ptt.cc/bbs/Gossiping/index.html
`def start_requests(self):` `…`	對最初的目標網址送出請求
`def parse(self,response):` `…`	resposne 取得文章列表，整理後對文章網址逐一以框架內的方式送出請求
`def parse_article(` `self,response): ...`	resposne 取得文章內容，整理取得目標資料，送入 Item Pipeline

從外部給予參數

上一個小節我們介紹從 CrawlerProcess 給予爬蟲參數來取得不一樣的結果，這邊我們補充介紹從命令列給予參數的方式

```
class PttcrawlSpider(scrapy.spider):
    name = 'PTTCrawler'
    def __init__(self, board='Gossiping'):
        …
```

我們可以透過 -a 的參數來傳參數 e.g. `scarpy crawl PTTCrawler -a board=Stock`

重要知識點複習

CUPOY

- 這邊主要是把前面教的內容全部都以框架的方式實作，加深對框架的理解與融會貫通，主要包含
  - Scrapy 送請求的方式
  - Item 與 Item Pipeline 的實作
  - 增加可控制的選項與外部呼叫爬蟲的方式

參考資料

- 【知乎】Scrapy中的scrapy.Spider.parse()如何被調用?
  - 參考 Scrapy 的架構圖，再透過該篇文章可以更加了解 parse 的時候 yield request 跟 yield item 的差別

解題時間
LET'S CRACK IT

請跳出 PDF 至官網 Sample Code & 作業

開始解題

START