Tripresso 面試專案

本次考試以1.請求行為分析 以及 2.爬蟲專案結構 兩個章節進行專案說明

1. 請求行為分析

本次面試題目為爬取新魅力旅遊網以及世邦旅行社的產品資訊。在經過幾次觀察後，發現兩家旅行社的商品清單皆透過” EW/Services/SearchPageList.asp”的服務傳送商品資料。

新魅力旅遊網由https://www.newamazing.com.tw/EW/Services/SearchListData.asp 以POST方法傳送json格式商品資料。

世邦旅行社則由https://www.4p.com.tw/EW/Services/SearchListData.asp 以POST方法傳送json格式商品資料。

根據需求，需要抓取旅遊天數、行程名稱、出發日期、價錢、可售位、總團位，其抓取回來的json資料對應的欄位如下，另外我個人額外增加出團型態。

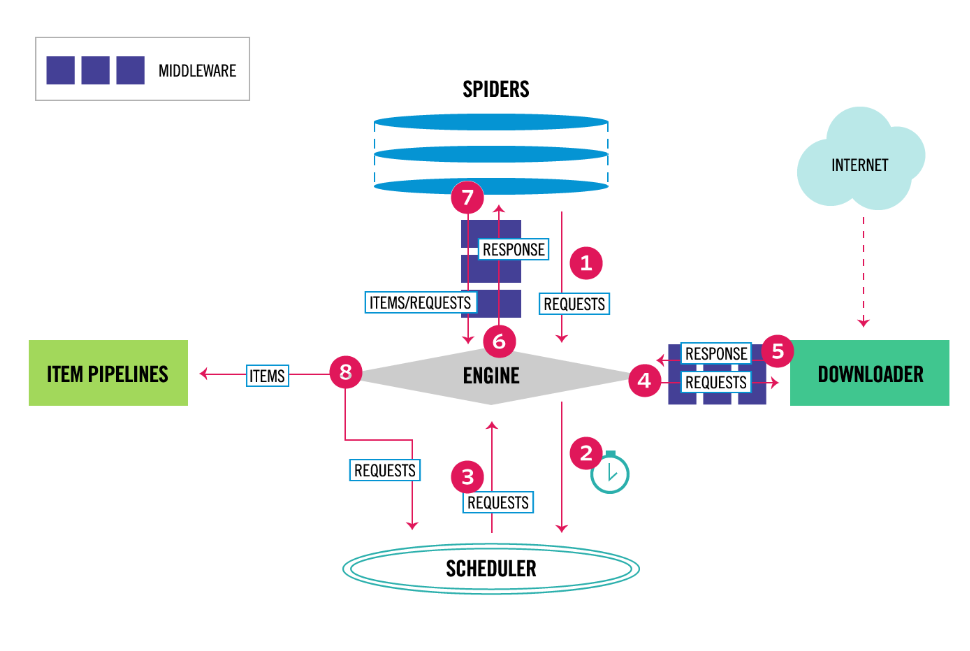
|  |  |
| --- | --- |
| 名稱 | 欄位 |
| 產品id | GrupCd |
| 出發日期 | LeavDt |
| 售價 | SaleAm |
| 旅遊天數 | GrupLn |
| 產品名稱 | GrupSnm |
| 總團位 | EstmTotqt |
| 可售位 | SaleYqt |
| 出團型態 | SubCdAnm |

爬取出發時間為2020/07/06至2021/07/06

1. 爬蟲專案結構

本次爬蟲我選擇Scarpy框架作為爬蟲工具。以下為Scrapy流程圖，取自官網

<https://docs.scrapy.org/en/latest/_images/scrapy_architecture_02.png>



專案啟動流程為

1.爬蟲檔會將資料傳送給engine

2.engine會將爬蟲工作放置於scheduler

3. scheduler將工作批次的交給engine

4. engine透過middleware送請求給下載器(downloader)向網站進行請求

5.資料請求後傳送回給engine再傳給spider內的方法進行解析

6.於spider解析資料完畢後整理成item，再透過itempipeline進行資料儲存。

專案結構如下

tripresso\_inter

spider

\_\_init\_\_.py

newamazing.py → 此為新魅力旅遊網爬蟲檔案

shipang.py → 此為世邦旅行社爬蟲檔案

\_\_init\_\_.py

items.py

middlewares.py

pipelines.py

setting.py

於spider資料夾內的檔案為爬蟲檔，其中newamazing.py是爬取新魅力旅遊網的資料，而shipang.py是爬取世邦旅行社的資料。

以下檔案內容分別為

1. tems.py是宣告儲存資料時的物件，spider將資料解析完畢後可由items.py中建立好的物件來進行傳送。
2. Middlewares.py為中間件，用途為發送請求前更改配置的地方，比如更換代理伺服器位址或更換瀏覽器的user-agent皆在這個檔案中進行設置。然而這兩個網站並未有反爬蟲機制，故Middlewares.py並不需要進行設置。
3. Pipelines.py為資料儲存配置文件，資料儲存方式為寫入MySql資料庫。

本次爬取範圍為旅遊網清單，以及商品的詳細內頁，內容包含詳細內頁的旅遊天數，每人訂金，包含項目以及不包含項目。

商品清單的資料表名稱為GroupList

商品詳細資料資料表名稱為Detail