訊息理解與 Web 智慧

[Homework 1]Web crawler and Open source IR system

學號 110522053　　姓名 王瑋

1 Abstract

撰寫一個Web Scraper程式，爬取採用動態Ajax傳輸網頁內容的KKday官網，蒐集全世界的當前熱門活動，濾掉顯示非中文的活動後將這4940筆資料匯入Elasticsearch，結合Kibana介面進行資料搜尋，藉由KQL快速篩選出想參加的活動。

2 Introduction

KKday是一個專門提供旅遊體驗與行程的線上平台，蒐集超過三萬種旅遊行程與體驗，本次作業目標希望能透過自己建立的IR系統，更有效地完成活動搜尋。一開始直接以requests套件爬取活動內容，發現KKday官網是使用Ajax的方式傳遞網頁內容，所以改從開發者模式取得伺服器回應的資料。

2.1 Ajax 動態網頁爬蟲

首先從開發者模式的Network下點開XHR(XML Http Request)，找到存有活動內容的請求回應，從header找到實際發送request的URL，接著使用requests套件發出需求，並使用fake-useragent套件偽裝成真實使用者，爬取時只保留每筆活動的重要資訊，也會在爬完一百筆後暫停十秒避免過多次的爬蟲造成IP封鎖，而KKday也會將已售罄的活動留在網頁的後段部分，所以我只保留可購買狀態的活動，最終網頁顯示的14396筆爬取後剩下6390筆。

3 Related work

使用requests發出請求後得到的資料類似Json格式，包括總共活動筆數、頁數，也可以將rows數量當作參數送出請求，不會有例外處理的問題，因為已售罄的活動都會擺在最後面，所以一爬到已售罄的活動就停止爬蟲，接著將爬到的資料輸出成JSON、CSV檔。

4 Method

4.1 爬蟲過程講解

KKday的活動分類及城市名稱都是以代碼呈現，活動分類的部分透過搜尋篩選欄位對照代碼轉換成文字，而城市代碼會搭配名稱一起出現所以只要抓取名稱欄位即可，再來先使用requests套件完成第一頁的網頁請求，從第一頁中抓取總筆數及頁數便可以透過迴圈完成全部資料抓取。

4.2 輸出：JSON, CSV格式

爬取後分別輸出成json及csv檔案格式，發現有非中文格式的活動，第一輪先使用langid套件篩選出非中文的活動，第二輪再以人工還原誤判成外文的活動，最後產出剩下4940筆有效活動。

4.3 資料清洗 & 了解各國活動數量

透過langid套件及人工方式刪除非中文格式的活動後，檢查資料內的空值，發現可預訂日期總共有767個Nan值，以爬完蟲當下的時間做補值，並將型態轉成datetime，最後也有將每個國家的活動數量做統計，繪製成圓餅圖。

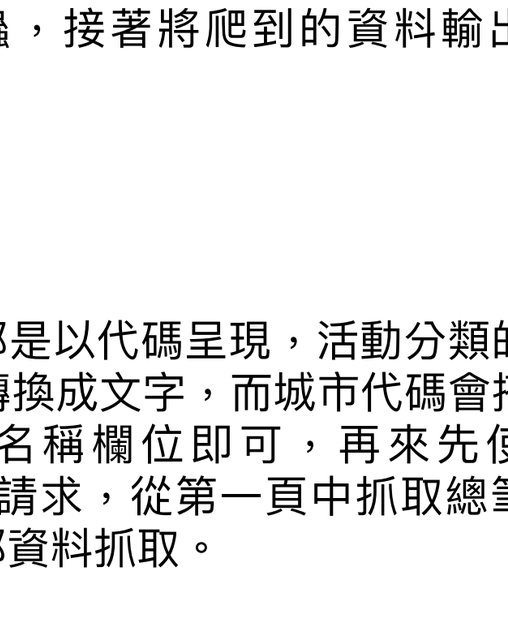

Figure 1:各國活動佔比圓餅圖

4.4 Data Schema (Mapping type)

直接將上一步處理好的json檔以批次的方式將檔案直接匯入KKday的index裡，透過Elasticsearch自動產生mapping type。

欄位名稱	欄位定義	欄位說明
title	text, keyword	活動標題
introduction	text, keyword	活動介紹
link	text, keyword	活動連結
cat_key	text, keyword	活動類別分類
cities	text, keyword	適用程式
country	text, keyword	國家
booking_date	date	最早可預定日
price	long	活動價格
rating_count	long	評比次數
rating_star	float	評比星級

Table 1:各欄位代表意義

以下稍微說明選擇這些欄位的原因，title及introduction都可以透過城市、地名及節慶等等詞找到想要的活動，而price、rating_count、rating_star都是查詢做篩選時的參考依據，還有cat_key、cities都是一個多元素的列表，可以透過活動類別及城市做進一步搜尋，最後如果真的有興趣也有link可以直接連到KKday的該活動網頁。

5 Experiment

KQL查詢	第一筆輸出	總hits數
title : ("門票" or "農場" or "遊樂園") and country.keyword : "台灣" and introduction : "優惠"	title:優惠94折｜屏東海生館門票 introduction:即訂即用屏東國立海洋生物博物館門票，透過 KKday 線上購票立即擁有 94 折門票優惠…… country:台灣	57
price <= 1000 and rating_star >= 4.2 and country.keyword : "泰國"	price:851 rating_star:4.85 country:泰國	43
cities.keyword : "台南" and (introduction: "美食" or cat_key : "當地美食")	introduction:立即預訂台南火車站前機車租借，來趟台南最在地的玩法，前往安平景點品嚐美食、觀夕平台…… cities:台南 cat_key:一日遊, 租車自駕	22
(rating_count >= 2000 or rating_star >= 4.8) and country: "日本" and cat_key: "戶外休閒"	country:日本 cat_key:一日遊, 特色行程, 戶外休閒 rating_star:4.64 rating_count:3,747	16
title: ("住宿" or "飯店") and introduction: "旅遊" and cities.keyword : "澎湖"	title:【KKday 獨家畢旅專區 8 折起】澎湖三天兩夜自由行｜機票＋住宿＋機車使用三天兩夜……cities:澎湖	4

Table 2:KQL查詢測試

在Elasticsearch上建立index後，於Kibana建立新的index pattern，Table 2的前三筆沒有將booking_date設定成搜尋範圍，後兩筆則設定時間範圍在2021/12/31~2022/12/31之間。

6 Conclusion

Ajax動態傳輸資料內容的網站相比於其他爬蟲，不能直接使用套件（beautiful soup直接爬取網頁標籤內容、selenium動態執行網頁上的操作），需要先透過人工的方式找出XHR，不過找到後再取得資料部分就會簡單許多，而在過濾非中文字部分以套件做第一輪篩選後還需人工做第二次確認，在未來還是會繼續找尋更好的解法改善此部分。Elasticsearch使用上可以批次且動態決定檔案內容型態是十分方便的，搭配Kibana簡化Elasticsearch查詢上的不便，以KQL語法做簡單的資料過濾與查詢。

7 Reference

Python packages:

1. langid https://github.com/saffsd/langid.py
2. requests https://docs.python-requests.org/en/latest/
3. fake_useragent https://github.com/hellysmile/fake-useragent

Tools:

1. Online API testing tool https://reqbin.com/
2. Online Json formatter https://jsonformatter.curiousconcept.com/

Source code: https://github.com/kazuyahooo/web_crawler_kkday