

591租屋網爬蟲API、模型API和數據儀表板

報告人：廖泓舜 Sam Liao

Aganda

- 01 專案目標與要點
- 02 功能介紹與示範
 - 功能介紹
 - 程式Demo
- 03 問題與處理
- 04 可優化方向



專案目標與要點

591

專案題目：

1. 撰寫爬網程式取得【591房屋交易租屋網】中,位於【臺中市】的所有【租屋物件數據】。
2. 【租屋物件數據】至少須具有下列欄位:
 - 出租者(ex 陳先生)- 出租者身份(ex 屋主) – 聯絡電話(ex 04-25569419)
 - 型態(ex 電梯大樓) – 現況(ex 獨立套房)- 性別要求(ex 男女生皆可)
 - 屋況說明
3. 針對以上數據,請在AWS 或 GCP上實現以下之一任務(擇一或組合表現):
 - 協助數據科學家將其建立的租金價格預測模型,或應用NLP技術提供insight或分類
 - 提供內部人員每日查看的Dashboard
 - 提供內部其他服務存取資料的API

提案目標與要點

Mission 1
spider-591

Mission 2
price predict model

Mission 3
data platform

Github: https://github.com/ba88052/rental_591_analyze

功能介紹

Mission 1
spider-591

Mission 2
price predict model

Mission 3
data platform

包含一個 POST API , 可輸入參數後，回傳爬蟲資料。 <http://35.229.148.113:8081/spider-591>

需輸入參數包含：

- "showMore": "1", (套用更多搜尋方式)默認開啟
- "searchtype": "1", (搜索型態) 1-鄉鎮, 2-商圈, 3-學校, 4-捷運
- "region": "8", (城市) 台中
- "section": "104", (區域) ##詳閱 rental_591_params_detail
- "multiPrice": "", (租金) 0_5000, 5000_10000, 10000_20000, 20000_30000, 30000_40000, 40000_
- "rentprice": "", (自訂租金範圍) ex: 0_10000
- "kind": "", (類型) 0-不限, 1-整層住房, 2-獨立套房, 3-分租套房, 4-雅房, 8-車位, 24-其他,
- "shape": "", (型態) 1-公寓, 2-電梯大樓, 3-透天厝, 4-別墅
- "multiNotice": "", (須知) all_sex, boy, girl, not_cover
- "multiRoom": "", (格局) 1-一房, 2-二房, 3-三房, 4-四房以上
- "multiArea": "", (坪數) 0_10, 10_20, 20_30, 30_40, 40_50, 50_
- "area": "", (自訂坪數範圍) ex: 10_50
- "multiFloor": "", (樓層) 0_1, 2_6, 6_12, 12_
- "option": "" (設備)

功能介紹

Mission 1
spider-591

Mission 2
price predict model

Mission 3
data platform

回傳資料包含：

"title": { "13406220": "正沙田路黃金店面" }, (名稱)
 "kind": { "13406220": "整層住家" }, (現況)
 "community": { "13406220": "" }, (附近設施)
 "area": { "13406220": "55" }, (坪數)
 "section": { "13406220": "大肚區" }, (區域)
 "shape": { "13406220": "透天厝" }, (型態)
 "layout": { "13406220": "9房3廳4衛" }, (格局)
 "address": { "13406220": "台中市大肚區沙田路二段" }, (地址)
 "inName": { "13406220": "徐小姐" }, (出租者)
 "role": { "13406220": "仲介" }, (出租者身份)
 "phone": { "13406220": "0423372000" }, "phone_extension": { "13406220": null },
 "mobile": { "13406220": "0968525915" }, "mobile_extension": { "13406220": null }, (聯絡電話)
 "rule": { "13406220": "此房屋限女生租住" }, (租屋規則)
 "remark": { "13406220": "近大肚市區，熱鬧非凡，門口可停雙車20米大路，車潮量大適辦公室，樓上可住家，可營登\r\n附近有便利商店、傳統市場。" }, (屋況說明)
 "price": { "13406220": 38000 } } (價格)

【租屋物件數據】至少須具有下列欄位：

- 出租者(ex 陳先生)
- 出租者身份(ex 屋主)
- 聯絡電話(ex 04-25569419)
- 型態(ex 電梯大樓)
- 現況(ex 獨立套房)
- 性別要求(ex 男女生皆可)
- 屋況說明

功能介紹

Mission 1

spider-591

Mission 2

price predict model

Mission 3

data platform

回傳資料轉 DataFrame

post_id	title	kind	community	area	section	shape	layout	address	inName	role	phone	phone_extension	mobile	mobile_extension	rule	remark	price
13550926	13550926 大富路/近交流道/全新電梯2房	整層住家		21	神岡區	透天厝	2房2廳1衛	台中市神岡區大富路35巷	吳志麒	仲介	None		None 0982655605	None	不可養寵物，不可開伙	包水/電5公共電費300元/月獨立陽台/獨洗獨曬/24小時監視系統/大門密碼管制/戶數少/住...	19000
13495596	13495596 國1麥當勞全新一廳一廳乾濕分離★獨洗	獨立套房		12	神岡區	透天厝	--	台中市神岡區社南街		仲介	None		None 0916507977	None	此房屋男女皆可租住，不可養寵物，不可開伙；適合上班族及家庭	1.交通位於社南街社口國小旁。2.全新設備，冷氣、冰箱、電視、網路Wife、第四台、床、衣櫃...	7800
13469594	13469594 神岡三民南路!可住家可店面搶先看	整層住家		52	神岡區	透天厝	4房3廳4衛	台中市神岡區三民南路19號	陳建興	仲介	None		None 0902146688	None	 ----- ----- -----	----- 35000 ----- -----	35000
13531712	13531712 免管理費全新全配備電梯套房	分租套房		8	神岡區	透天厝		台中市神岡區社南街5巷	陳先生	仲介	None		None 0933552280	None	此房屋男女皆可租住，不可養寵物，不可開伙	社口麥當勞旁全新套房 電梯 獨洗獨曬精裝全配備 :....	8600
13486977	13486977 優質寧靜、精美裝潢、交通便利、整潔舒適	分租套房		6	神岡區	透天厝		台中市神岡區大裡街29巷16號	陳啟邦	代理人	None		None 0911769500	None	此房屋男女皆可租住，不可養寵物，不可開伙；適合上班族	*****乾淨優質、環境好、房客單純、寧靜*****1.自租免仲介費，免管理費.清潔費.公用...	6800

功能介紹

Mission 1
spider-591

Mission 2
price predict model

Mission 3
data platform

GET API, <http://35.229.148.113:8081/spider-591/daily>。

每天爬取指定參數的資料，並將資料上傳到 BigQuery 中。

Explorer + 新增資料

2022-11-13_RENTAL

查詢 共用 複製 快照 刪除 匯出

	post_id	title	kind	community	area	section
1	13538597	美村南路露天平面~限量車位	車位	寶雅美村南店	3	南區
2	13396474	嘟嘟房 多場站室內停車位	車位		10	西區
3	13540138	74潭子重機車位24小時進出保...	車位		1.5	潭子區
4	13480723	正西屯路近何厝街停車場	車位		29	西屯區

每頁結果數: 50 | 1 - 50 / 12891 | < > >>

功能介紹

Mission 1

spider-591

Mission 2

price predict model

Mission 3

data platform

運用crontab將每日爬蟲部署在GCP上

```
# Edit this file to introduce tasks to be run by cron.  
#  
# Each task to run has to be defined through a single line  
# indicating with different fields when the task will be run  
# and what command to run for the task  
#  
# To define the time you can provide concrete values for  
# minute (m), hour (h), day of month (dom), month (mon),  
# and day of week (dow) or use '*' in these fields (for 'any').#  
# Notice that tasks will be started based on the cron's system  
# daemon's notion of time and timezones.  
#  
# Output of the crontab jobs (including errors) is sent through  
# email to the user the crontab file belongs to (unless redirected).  
#  
# For example, you can run a backup of all your user accounts  
# at 5 a.m every week with:  
# 0 5 * * 1 tar -zcf /var/backups/home.tgz /home/  
#  
# For more information see the manual pages of crontab(5) and cron(8)  
#  
# m h dom mon dow command  
0 16 * * * /usr/bin/python3 /home/ba88052/rental_591_analyze/app/spider_591/spider_591.py
```

Mission 1
spider-591

Mission 2
price predict model

Mission 3
data platform

程式Demo

功能介紹

Mission 1
spider-591

Mission 2
price predict model

Mission 3
data platform

包含一個 POST API , 可輸入參數後, 回傳模型預測租金。 <http://35.229.148.113:8081/model>

需輸入參數包含:

“model”: “XGB”,	(模型) XGB 、 keras
“area”: “35.7”,	(坪數)
“bedroom”: “4”,	(臥室數量)
“livingroom”: “2”,	(客廳數量)
“bathroom”: “2”,	(浴廁數量)
“section”: “霧峰區”,	(區域)
“kind”: “整層住家”,	(現況) 整層住房, 獨立套房, 分租套房, 雅房, 車位
“shape”: “電梯大樓”,	(型態) 公寓, 電梯大樓, 透天厝, 別墅
“role”: “仲介”,	(出租者身份) 仲介、屋主、代理人
“girl_cant_live”: “0”,	(不租給女性) 0- Flase 1-True
“boy_cant_live”: “1”,	(不租給男性) 0- Flase 1-True
“pet_cant_live”: “1”,	(不可養寵物) 0- Flase 1-True
“cant_cooking”: “0”	(不可開伙) 0- Flase 1-True

功能介紹

Mission 1
spider-591

Mission 2
price predict model

Mission 3
data platform

回傳預測租金

```
▶ #price_predict_model
#API網址
url = "http://35.229.148.113:8081/model"

#標頭
headers = {"Content-Type": "application/json"}

#參數
data = {
    "model": "XGB",
    "area": "35.7", "bedroom": "4",
    "livingroom": "2", "bathroom": "2",
    "section": "霧峰區", "kind": "整層住家",
    "shape": "電梯大樓", "role": "仲介",
    "girl_cant_live": 0, "boy_cant_live": 1,
    "pet_cant_live": 1, "cant_cooking": 0
}

access_token = requests.post(url, headers=headers, json=data)
print(access_token.text)
```

⇒ 28903.145

功能介紹

GET API

<http://35.229.148.113:8081/spider-591/model/web>

將模型以網頁方式呈現。

Mission 1
spider-591

Mission 2
price predict model

Mission 3
data platform

歡迎使用 台中租屋租金預測模型

數據來源：591租屋網

要使用的模型

XGBoost TensorFlow Keras

坪數

臥室數量(沒有填0)

1

客廳數量(沒有填0)

0

浴廁數量(沒有填0)

0

區域

西屯區 ▼

房間類型

電梯大樓 ▼

房間類型

獨立套房 ▼

由誰販售

仲介 ▼

特殊規定

- 禁止女生住宿
- 禁止男生住宿
- 禁養寵物
- 不可開伙

submit

預測價格：

功能介紹

Mission 1
spider-591

Mission 2
price predict model

Mission 3
data platform

GET API,

<http://35.229.148.113:8081/model/weekly-training>

根據7天內 BigQuery 上資料重新訓練模型。

連接 GCP

```
<google.cloud.bigquery.client.Client object at 0x7fdc74611160>
```

取得 Bigquery 資料

訓練 XGB 模型

XGB 模型 儲存 完成

```
XGB_mean_absolute_error:373284621.5840538
```

```
XGB_mean_squared_error:3298.534207027765
```

```
XGB_mean_absolute_percentage_error:0.18511086907636443
```

訓練 keras 模型

keras 模型 儲存 完成

```
keras_mean_absolute_error:134320008.40733144
```

```
keras_mean_squared_error:7270.707537023553
```

```
keras_mean_absolute_percentage_error:0.36983542826821075
```

每週訓練完成

```
1.174.27.248 - - [17/Nov/2022 06:26:54] "GET /model/weekly-training HTTP/1.1" 200 -
```

功能介紹

Mission

1

spider-591

Mission

2

price predict model

Mission

3

data platform

運用crontab將每週模型訓練部署在GCP上

```
# Edit this file to introduce tasks to be run by cron.  
#  
# Each task to run has to be defined through a single line  
# indicating with different fields when the task will be run  
# and what command to run for the task  
#  
# To define the time you can provide concrete values for  
# minute (m), hour (h), day of month (dom), month (mon),  
# and day of week (dow) or use '*' in these fields (for 'any').#  
# Notice that tasks will be started based on the cron's system  
# daemon's notion of time and timezones.  
#  
# Output of the crontab jobs (including errors) is sent through  
# email to the user the crontab file belongs to (unless redirected).  
#  
# For example, you can run a backup of all your user accounts  
# at 5 a.m every week with:  
# 0 5 * * 1 tar -zcf /var/backups/home.tgz /home/  
#  
# For more information see the manual pages of crontab(5) and cron(8)  
#  
# m h dom mon dow   command  
0 6 * * * /usr/bin/python3 /home/ba88052/rental_591_analize/app/spider_591/spider_591.py  
0 0 * * 1 /usr/bin/python3 /home/ba88052/rental_591_analize/app/rental_price_model/rental_price_model.py
```

功能介紹

Mission 1
spider-591

Mission 2
price predict model

Mission 3
data platform

程式Demo

功能介紹

Mission 1
spider-591

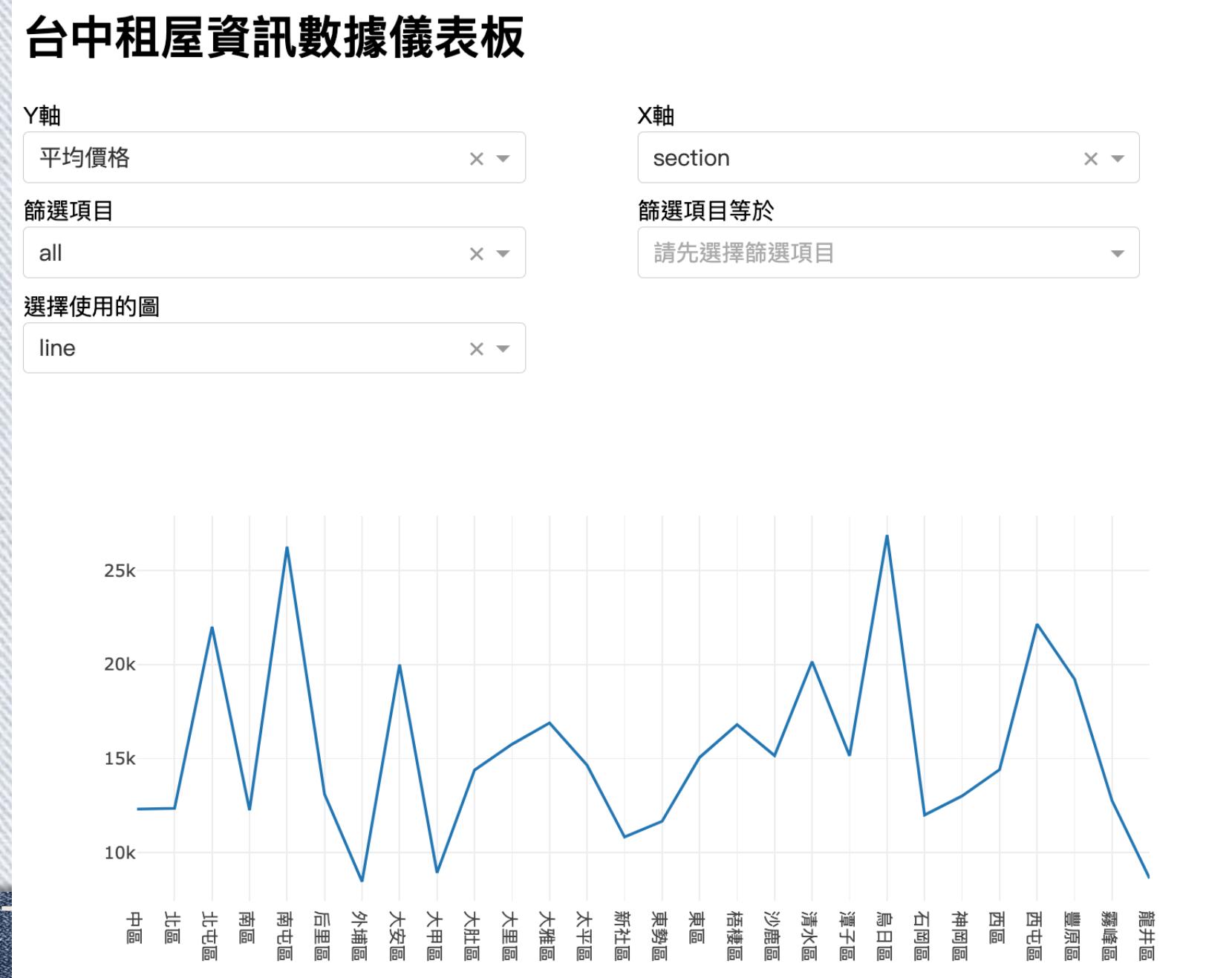
Mission 2
price predict model

Mission 3
data platform

GET API

<http://35.229.148.113:8081/spider-591/platform>

抓取 BigQuery 的資料，使用Dash框架，以網頁方式呈現動態儀表板。



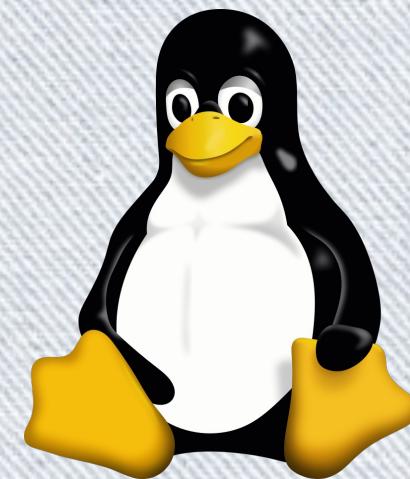
Mission 1
spider-591

Mission 2
price predict model

Mission 3
data platform

程式Demo

問題與處理

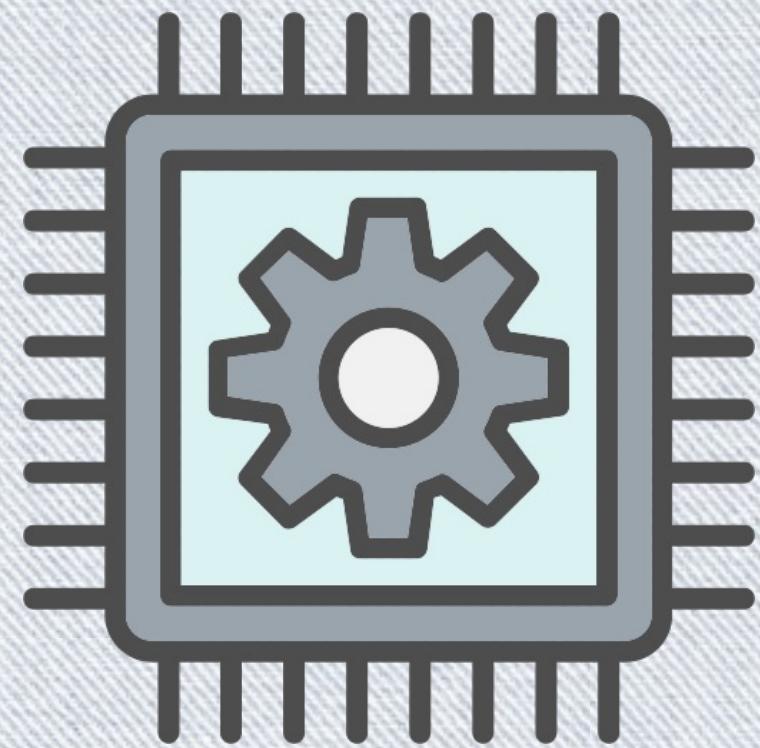


問題	闡述	解決方法
前端與Dash的撰寫	對於JS較為不熟，但想做一個動態圖表。	使用新學的套件Dash
GCP的權限設置	在IAM設置上，不確定如何讓使用者有使用python連接做 SELECT 和創建 table 的權限。	試錯法，最後找到需要給予 BigQuery使用者 和 BigQuery資料編輯者 的權限
Crontab突然消失	發現有一天設置的定時功能並沒有進行爬蟲。	發現在設置每週模型訓練時，因沒有進行儲存的情況下意外關閉頁面，導致過去設置的任務消失。

可優化方向



Dash與前端介面優化



模型準確率優化