

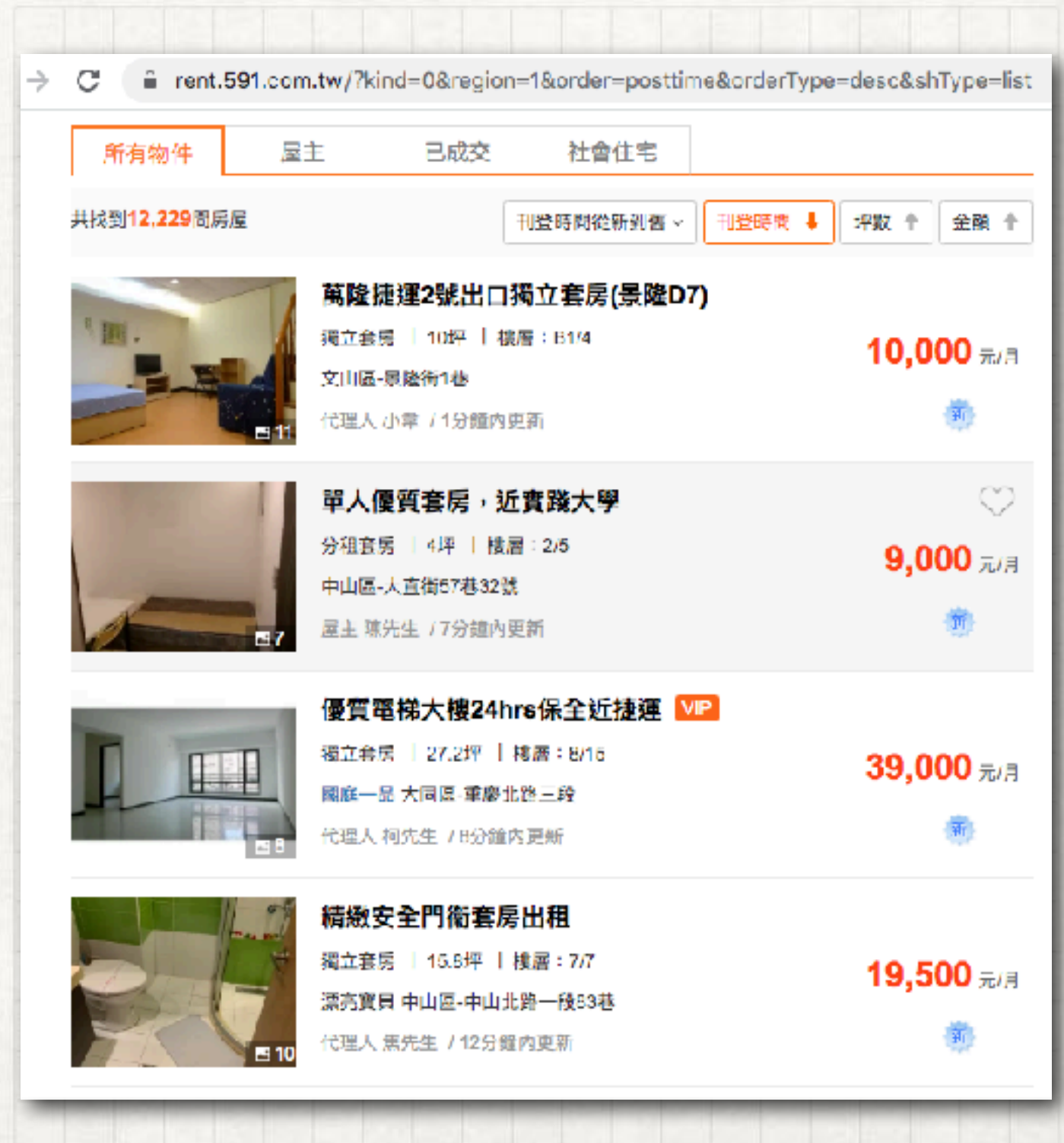
# 591租屋物件爬蟲

- 列表爬蟲

1. 篩選功能需要靠Selenium操作
  2. 節省request及資料完整性
- 故獨立一隻列表爬蟲，先爬回所有連結  
(每個 request 三十筆、台北市約300頁)

- 內容爬蟲

1. 進入各物件頁面爬取資料
2. 內容寫於HTML中，相對容易解析



## ● 資料更新機制

按照刊登時間排序，  
從第一頁開始向後爬取，  
直到id與資料庫中的id重複。

每爬一頁可以更新30個link

## ● 資料庫設計

列表爬蟲資料表：pageIndex


內容爬蟲資料表：pageContent

→ [rent.591.com.tw/?kind=0&region=1&order=posttime&orderType=desc&shType=list](http://rent.591.com.tw/?kind=0&region=1&order=posttime&orderType=desc&shType=list)

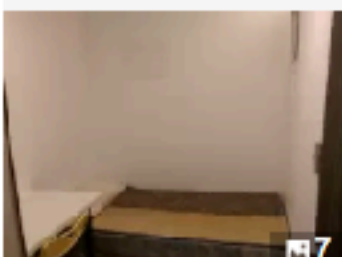
所有物件 屋主 已成交 社會住宅

共找到12,229間房屋

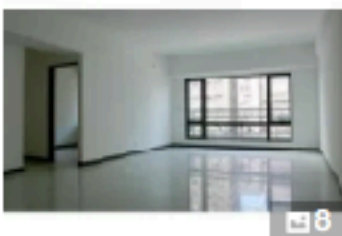
刊登時間從新到舊 ▾ 刊登時間 ↓ 坪數 ↑ 金額 ↑



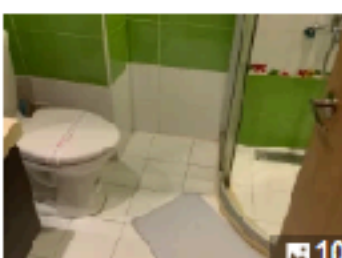
**萬隆捷運2號出口獨立套房(景隆D7)**  
獨立套房 | 10坪 | 樓層：B1/4  
文山區-景隆街1巷  
10,000 元/月  
代理人 小華 / 1分鐘內更新



**單人優質套房，近實踐大學**  
分租套房 | 4坪 | 樓層：2/5  
中山區-大直街57巷32號  
9,000 元/月  
屋主 陳先生 / 7分鐘內更新



**優質電梯大樓24hrs保全近捷運** VIP  
獨立套房 | 27.2坪 | 樓層：8/15  
國庭一品 大同區-重慶北路三段  
39,000 元/月  
代理人 柯先生 / 8分鐘內更新



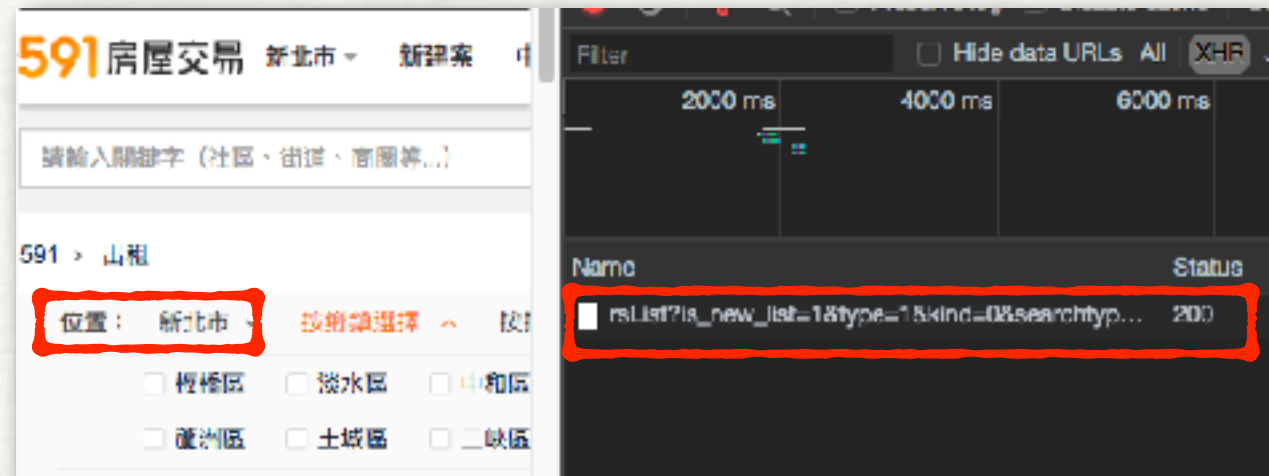
**精緻安全門衛套房出租**  
獨立套房 | 15.8坪 | 樓層：7/7  
漂亮寶貝 中山區-中山北路一段83巷  
19,500 元/月  
代理人 焦先生 / 12分鐘內更新



# 物件列表爬蟲-解法

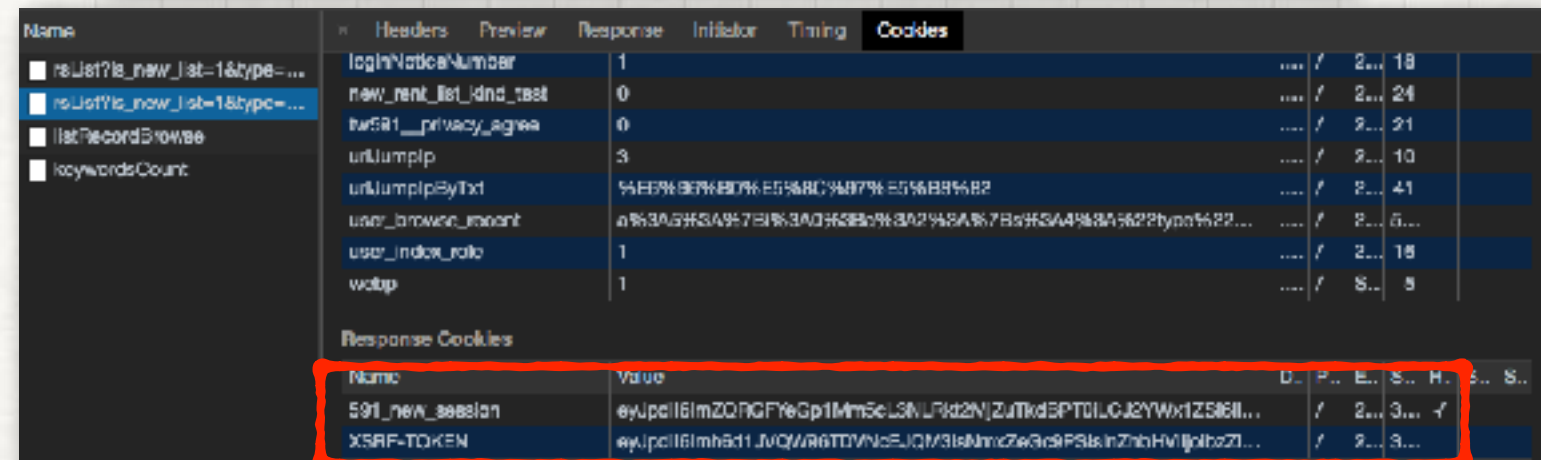
## • 切換縣市

需點選**按鈕**切換後**重新整理**  
網址的Region參數  
無法直接用request.url切換



## • XSRF TOKEN

無法透過HTTP取得後模擬  
故採用Selenium爬取



## • 翻頁功能

翻到最後一頁時，  
下一頁按鈕會無法點選，  
此時切換至下一個Region



## • 儲存結果

將如右圖之格式結果  
存進pageIndex

```
{
  "_id" : ObjectId("5e6d0374fd92196d935c7a55"),
  "id" : "8969800",
  "title" : "台北租屋,士林租屋,獨立套房出租,士林節能套房(限女性)",
  "link" : "https://rent.591.com.tw/rent-detail-8969800.html",
  "region" : "台北市"
}
> db.pageIndex.findOne()
```

# 物件內容爬蟲

- 取出列表

從DB中取出所有link

```
{
  "_id" : ObjectId("5e6d0374fd92196d935c7a55"),
  "id" : "8969800",
  "title" : "台北租屋,士林租屋,獨立套房出租,士林節能套房(限女性)",
  "link" : "https://rent.591.com.tw/rent-detail-8969800.html",
  "region" : "台北市"
}
> db.pageIndex.findOne()
```

- 取得頁面資訊

網頁以單純HTML呈現

經測試可不設定Download Delay

 小章 (代理人)

 0910-301-023 

坪數: 10坪

樓層: B1/4F

型態: 公寓

現況: 獨立套房

更新於: 36分鐘內 有效期: 2020-04-14

押 金: 二個月

最短租期: 一年

身份要求: 學生、上班族、家庭

法定用途: 住家用

車 位: 無

開 伙: 不可以

性別要求: 男女生皆可

建物面積: 34.92坪 (不含公設)

管 理 費: 無

養 寵 物: 不可以

可遷入日: 隨時

產權登記: 已辦

```
{
  "_id" : ObjectId("5e6d99ac2b35b13d122d9c2a"),
  "row" : {
    "身分" : "(屋主聲明: 仲介勿擾)",
    "稱謂" : "游先生",
    "電話" : "0986-851-077 轉 1360391",
    "id" : "1515548",
    "title" : "獨立套房出租,走路80秒到捷運站,獨立套房-新北三軍區房屋出租-591租屋網",
    "Status_code" : 200,
    "有效期" : "2020-04-13",
    "坪數" : "8坪",
    "樓層" : "4F/4F",
    "型態" : "公寓",
    "現況" : "獨立套房",
    "押金" : "二個月",
    "車 位" : "無",
    "管理費" : "--",
    "最短租期" : "一年",
    "開伙" : "不可以",
    "養寵物" : "不可以",
    "身份要求" : "學生、上班族、家庭",
    "性別要求" : "男女生皆可",
    "可遷入日" : "隨時",
    "非於政府免付費公開資料可查詢法定用途" : "未核實",
    "非於政府免付費公開資料可查詢建物面積" : "未核實",
    "產權登記" : "已辦"
  }
}
```



# 爬蟲結果

- PAGEINDEX

總筆數：21733 (隨時間變動)

台北市：12350

新北市：9383

- PAGECONTENT

總筆數：21534 (扣除[404]NotFound)

- 補充

房屋列表有大部分的所需資訊，  
但缺少聯絡電話及性別需求，  
故仍須進入頁面內容進行爬取

```
> db.pageIndex.find().count()
21733
> db.pageIndex.find({'region':{'regex':'台北'}}).count()
12350
> db.pageIndex.find({'region':{'regex':'新北'}}).count()
9383
```

```
> db.pageContent.find().count()
21534
> db.pageContent.distinct('row.身分')
[ "(仲介，不須服務費)", "(仲介，收取服務費)", "(代理人)", "(屋主聲明：仲介勿擾)", "(屋主)" ]
> db.pageContent.distinct('row.型態')
[ "住宅大樓", "倉庫", "公寓", "別墅", "店面（店鋪）", "華廈", "透天厝", "電梯大樓" ]
```



## 中正橋頭水岸景觀全新久泰逸品頂溪站9分鐘

整層住家 | 2房1廳1衛 | 28.9坪 | 樓層：3/15

久泰逸品 永和區-環河東路一段

屋主 趙先生 / 6分鐘內更新 / 63人瀏覽

# 待改善項目

## 【資料清整】

電話格式清整（刪除“-”及轉接）：

暫解：用regex比對，效率較差

## 【爬蟲程式】

原狀態 Retry：

因為電腦網路速度不一

JavaScript未載入完全時

會造成selenium 定位的exception

應利用try catch實作

隔一段時間retry的機制

暫解：手動time.sleep(2)

```
"開伙": ": 不可以",  
"電話": "0986-851-077 轉 1360391",  
"非於政府免付費公開資料可查詢建物面積": ": 未核實",  
"非於政府免付費公開資料可查詢法定用途": ": 未核實",  
"養寵物": ": 不可以"
```

GET

→

http://localhost:8080/rent/phone/phone=0989-924-818

Send

Params

Authorization

Headers (7)

Body

Pre-request Script

Tests

Settings

Query Params

	KEY	VALUE	DESCRIPTION
<input checked="" type="checkbox"/>	phone	0989-924-818	
	Key	Value	Description

StaleElementReferenceException: The element reference of e7

```
def next_page(self, bool_gonext):  
    time.sleep(1)  
    btn_next_page = self.driver.find_element_by_css_selector('div.pageBar a.  
    href_next_page = self.driver.find_element_by_css_selector('div.pageBar a  
    bool_next_page = (href_next_page != None)  
    if (bool_next_page & bool_gonext):  
        self.driver.execute_script("arguments[0].click();", btn_next_page)  
        time.sleep(5)  
        print('Loading Next Page ... ')
```