



---

# Python 網路爬蟲

丘祐瑋 – David Chiu

EMAIL: [david@largitdata.com](mailto:david@largitdata.com)

網站: [www.largitdata.com](http://www.largitdata.com)

電話: +886929094381

# 關於我

---



- 大數軟體有限公司創辦人

- 前趨勢科技工程師

- 大數學堂

<https://www.largitdata.com/>

- 粉絲頁

<https://www.facebook.com/largitdata>

- R for Data Science Cookbook

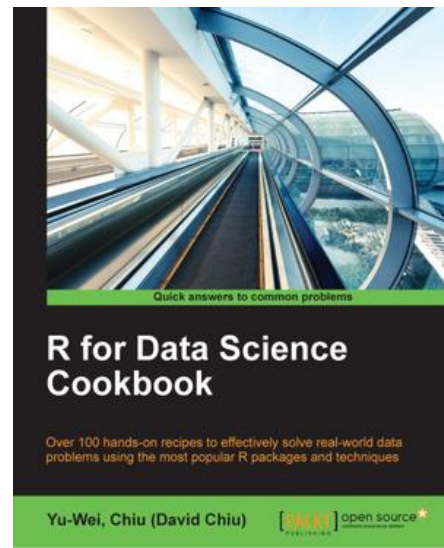
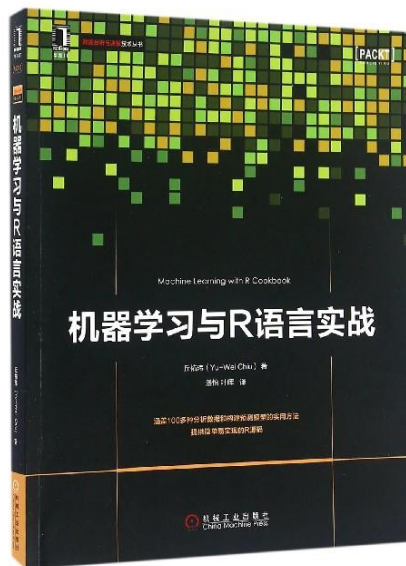
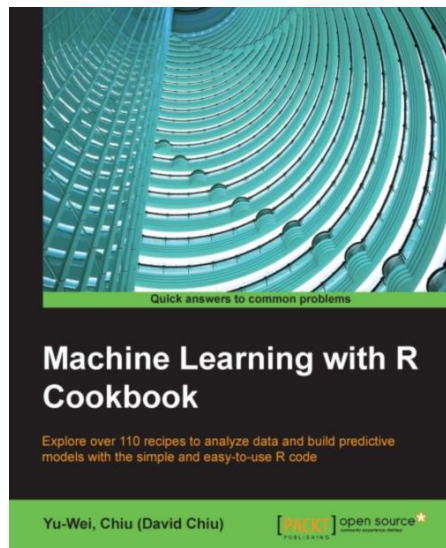
<https://www.packtpub.com/big-data-and-business-intelligence/r-data-science-cookbook>

- Machine Learning With R Cookbook

<https://www.packtpub.com/big-data-and-business-intelligence/machine-learning-r-cookbook>

# Machine Learning With R Cookbook (机器学习与R语言实战) & R for Data Science Cookbook (数据科学：R语言实现)

---



---

Author:  
David (YU-WEI CHIU) Chiu

# 課程補充資料

---

投影片、程式碼放置於：

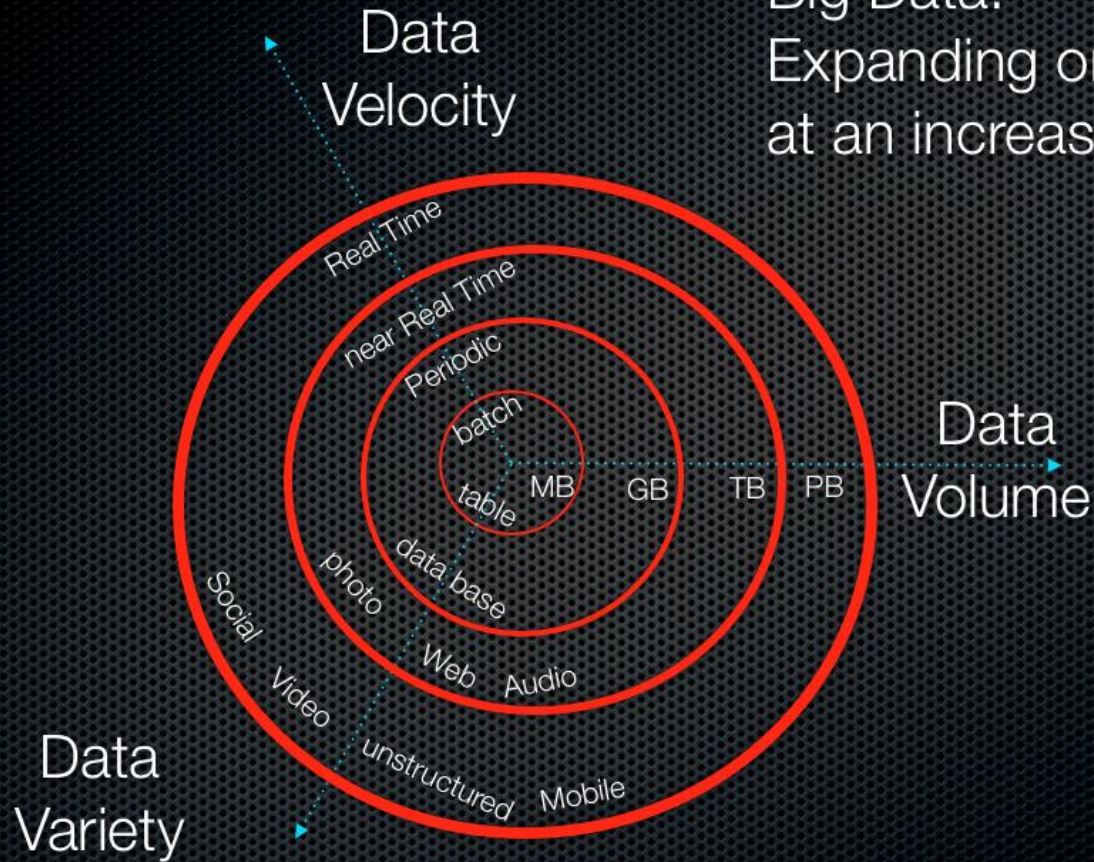
[https://github.com/ywchiu/nccu\\_crawler](https://github.com/ywchiu/nccu_crawler)



# 什麼是網路爬蟲？

---

Big Data:  
Expanding on 3 fronts  
at an increasing rate.







Infer, Predict,  
Recommend & Visualize

Contextualize,  
Model & Reason

De-complexify,  
Transform,  
Analyze &  
Network

Store

Collect

Hadoop

R, Hive, Pig, python, Java, Mahout

D3.js, Dashboards, web apps

ETL, Storm,  
Scribe, Flume,...

Machine Learning

Analytics

Data Architecture / Management

**Data Scientist**

Infer-ability

Model

Context

Connectedness

Variety

Variability

Velocity

Volume

Technologies

Roles

# 結構化vs半結構化vs非結構化數據

---

## 結構化資料

每筆資料都有固定的欄位、固定的格式，方便程式進行後續取用與分析  
例如：資料庫

## 半結構化資料

資料介於結構化資料與非結構化資料之間  
資料具有欄位，也可以依據欄位來進行查找，使用方便，但每筆資料的欄位可能不一致  
例如：XML, JSON

## 非結構化資料

沒有固定的格式，必須整理以後才能存取  
沒有格式的文字、網頁數據



# 非結構化資料

沒有固定的資料格式  
例如網頁數據

必須透過ETL  
(Extract, Transformation,  
Loading) 工具將資料轉換  
為結構化資料才能取用

共找到671間房屋

默认排序

刊登時間

地點

金額



大安捷運300公尺全新飯店裝潢2房1廳 黃金地段

整層住宅 | 2房1廳1衛 | 22坪 | 樓層: 4/5

大安區-基隆路52巷

屋主 David / 2小時內更新 / 244人瀏覽

34,000 元/月



復興南建國科技大樓大安森公園 黃金地段

整層住宅 | 2房1廳1衛 | 16坪 | 樓層: 2/4

大安區-和平東路二段

屋主 林小姐 / 2小時內更新 / 8條問詢 / 332人瀏覽

23,909 元/月



大安區六張犁捷運站採光佳, 3房2廳 黃金地段

整層住宅 | 3房2廳2衛 | 38坪 | 樓層: 4/7

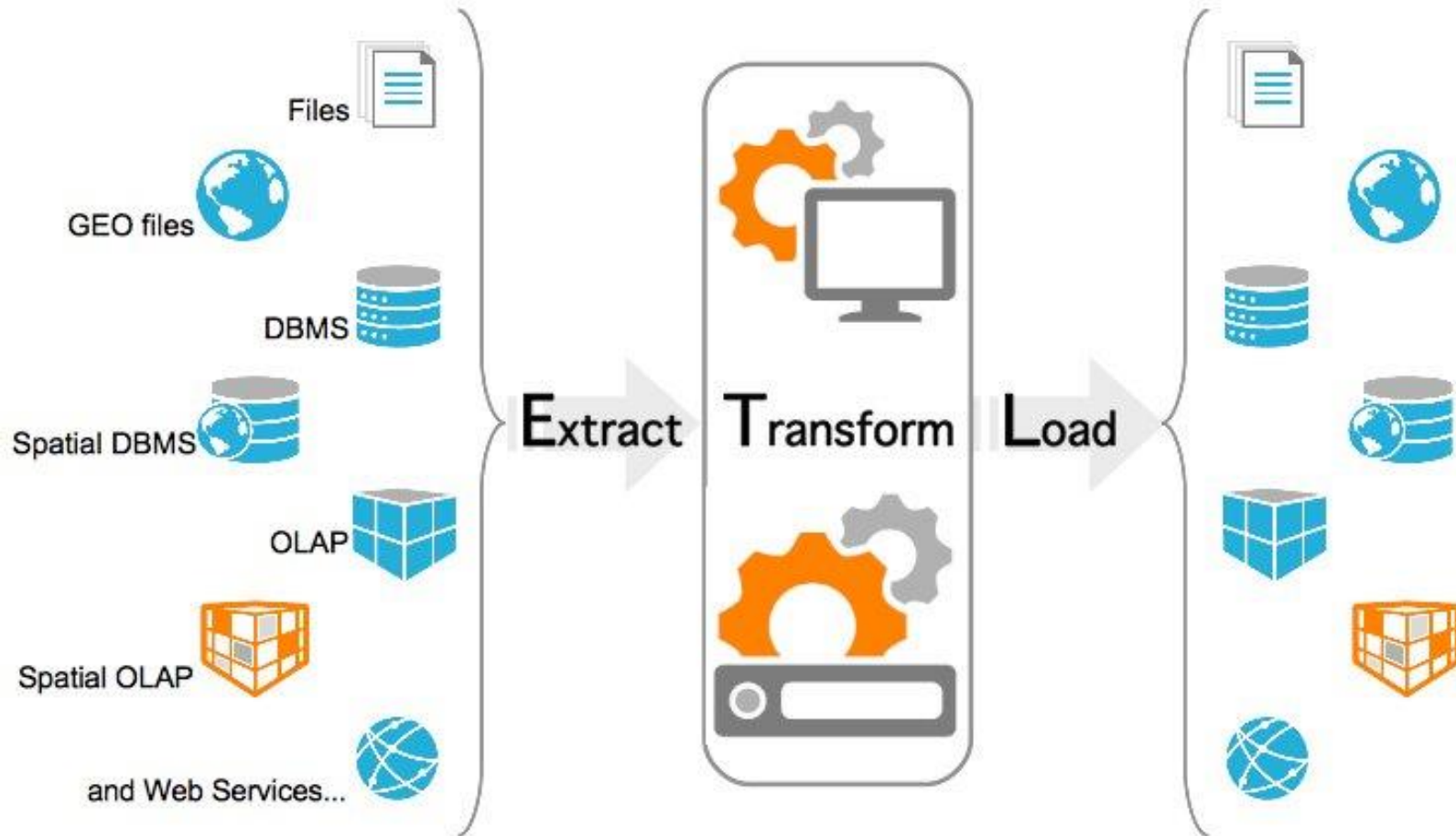
大安區-樂利路

屋主 謝太太 / 3小時內更新 / 99人瀏覽

43,800 元/月

如何從非結構化資料挖  
出價值資料是一大挑戰

# Extract, Transformation, Loading



# 資料抽取、轉換、儲存 (Data ETL)



原始資料

Raw Data



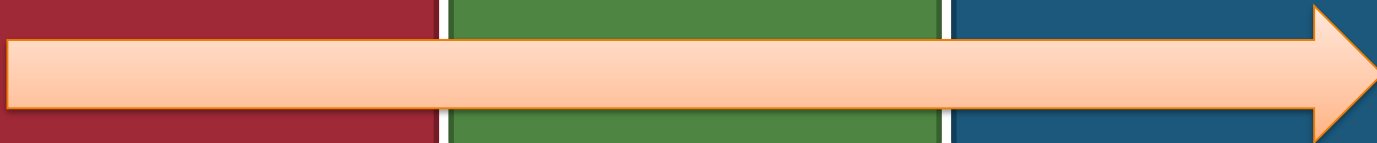
ETL腳本

ETL Script



結構化資料

Tidy Data





# 使用Python 打造網路爬蟲

---

# 目標: 將非結構化數據轉變為結構化數據



透過簡單的SQL語句  
從結構化資料中  
達到簡單的分析目的

Prices							
Date	Open	High	Low	Close	Volume	Adj Close*	
Mar 25, 2016	158.50	159.00	157.00	158.00	10,175,000	158.00	
Mar 24, 2016	158.00	159.00	157.00	158.50	24,853,000	158.50	
Mar 23, 2016	158.50	159.50	158.00	159.50	27,478,000	159.50	
Mar 22, 2016	159.50	159.50	157.00	158.50	25,809,000	158.50	
Mar 21, 2016	160.00	160.00	158.00	160.00	26,100,000	160.00	
Mar 18, 2016	158.50	159.50	158.50	159.50	55,975,000	159.50	
Mar 17, 2016	159.50	160.00	157.50	158.50	48,193,000	158.50	
Mar 16, 2016	155.50	156.00	154.00	156.00	30,962,000	156.00	
Mar 15, 2016	155.00	156.50	153.00	154.50	28,689,000	154.50	
Mar 14, 2016	156.50	157.50	155.50	156.00	32,751,000	156.00	
Mar 11, 2016	154.50	155.00	153.00	155.00	29,566,000	155.00	
Mar 10, 2016	153.00	154.50	151.50	154.50	28,302,000	154.50	
Mar 9, 2016	152.00	153.00	150.50	153.00	24,004,000	153.00	
Mar 8, 2016	151.00	152.00	149.50	152.00	35,683,000	152.00	
Mar 7, 2016	152.50	153.50	151.00	152.00	23,906,000	152.00	
Mar 4, 2016	153.00	153.50	151.50	152.50	32,794,000	152.50	
Mar 3, 2016	154.00	154.50	153.00	154.00	28,822,000	154.00	
Mar 2, 2016	154.00	154.50	153.00	153.00	36,010,000	153.00	

# 為什麼要使用Python 做網路爬蟲?

---

網路爬蟲



資料分析



網頁製作

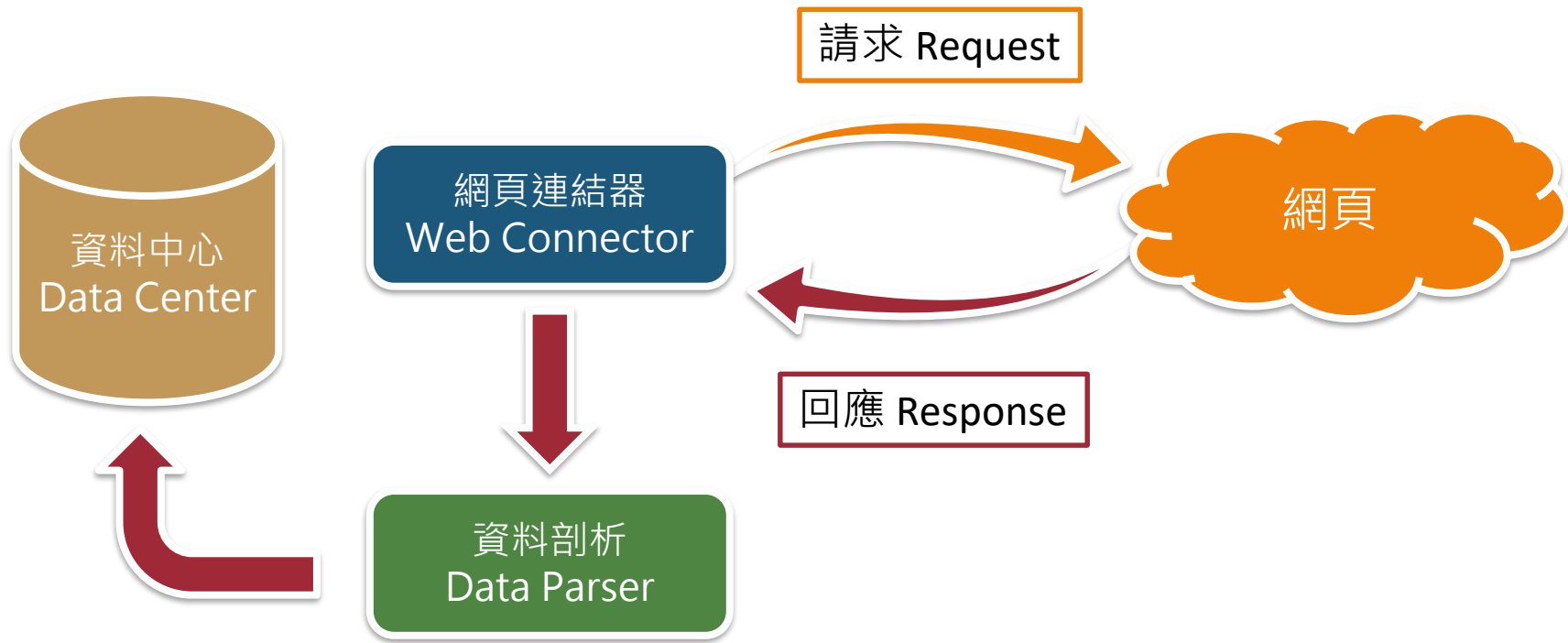


---

打造一條龍服務



# Python 爬蟲起手式



# 如何抓取蘋果即時新聞

<https://tw.appledaily.com/new/realtime>

The screenshot shows the Apple Daily Realtime website interface. At the top, there's a navigation bar with the Apple Daily logo and various menu items like '即時', '新聞', '生活', etc. Below the navigation bar, there's a large headline: '北市府同意遠雄 7月底前完成大巨蛋大底工程' (The Municipal Government agrees to Far Eastern Group to complete the Sun Yat-sen Memorial Hall renovation by the end of July). The main image shows the large, dome-shaped building under construction. To the right of the main article, there's a sidebar with promotional banners, including one for '醫靠 無界' (Medical Reliance, No Boundaries) and another for 'Create an Online Store Today!'. At the bottom, there's a date bar showing '2015/05/25' and a time bar showing '23:00'.

昨日瀏覽量: 17466897 • 結果日報自來水委員會 • 香港

新聞 | 娛樂 | 生活 | 國際 | 財經 | 地產 | 政治 | 論壇 | 陣線

動即時 最新 焦點 熱門 動物 FUN 僑胞 搜尋 影片 正妹 體育

娛樂 時尚 生活 社會 國際 財經 地產 政治 論壇 陣線

Gmail 商用版  
Google 為中小企業客裝化電子郵件。30 天免費試用，即刻體驗！

北市府同意遠雄  
7月底前完成大巨蛋大底工程

醫靠 無界  
無國界醫生之旅 攝影展  
6.12-6.21 敦南誠品  
11:00-21:00 免費入場 詳情 >>

Create an Online Store Today!  
✓ Zero Setup Fee  
✓ Zero Bandwidth Fees  
✓ 0% Transaction Fees\*  
✓ Unlimited SKU's  
✓ Unlimited Storage  
✓ \$100 Adwords Credit  
✓ Discount Codes  
START FREE TRIAL

動即時 按 看蘋果

2015/05/25

23:00 中國明發布 軍事戰略白皮書(0)

娛樂最 Hot 看更多

# 使用開發人員工具

於網頁上點選右鍵 -> 檢查

The screenshot shows a news website interface. At the top, there's a navigation bar with a red apple icon and the word '最新' (Latest), followed by various category links: 焦點 (Focus), 熱門 (Popular), 娛樂 (Entertainment), 愛播網 (Love Broadcast), 社會 (Society), 國際 (International), 政治 (Politics), 生活 (Life), 火線 (Frontline), 3C (3C), 動物 (Animals), 副刊 (Supplement), 體育 (Sports), and 財經地產 (Finance and Real Estate). Below this is a large blue banner with the text '原來她感情路是這樣' (It turns out her love life is like this). To the right of the banner are social media links for '蘋果日報' (Apple Daily) and '說這專頁' (Say This Special Page). Below the banner is a list of news items dated '2017 / 11 / 15'. A right-click context menu is open over the list, showing options like '上一頁(B)' (Previous page), '下一頁(F)' (Next page), '重新載入(R)' (Reload), '另存新檔(A)...' (Save as...), '列印(P)...' (Print...), '投放(C)...' (Share...), '翻譯成中文(繁體)(T)' (Translate to Chinese (Traditional) (T)), 'AdBlock', 'OneTab', '檢查(N)' (Inspect), and 'Ctrl+Shift+I'. The '檢查(N)' option is highlighted with a red box. In the bottom right corner, there is an orange button with the text '點選檢查' (Click to inspect).

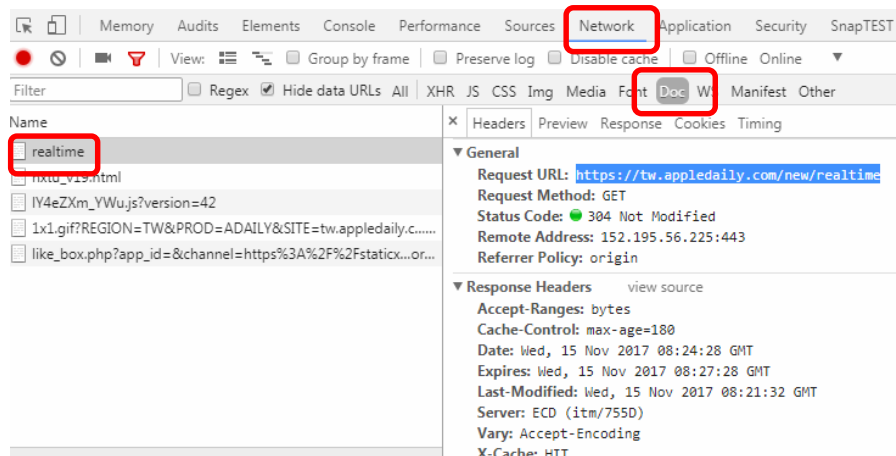
2017 / 11 / 15

- 16:19 **生活** 兩岸高職教育交流 第五屆台蘇論壇聖約大...
- 16:18 **社會** 光電廠解僱工會理事長 桃產總聲援抗議
- 16:17 **生活** 縮胃減去75公斤 換身分證差點被拒
- 16:15 **社會** 賴清德走了 南市開始放焰火還傷3人
- 16:15 **政治** 賴清德：慶富案是在2014年 哪一朝發生...
- 16:14 **社會** 油漆掉落灑5公尺路面 1騎士滑倒幸輕傷
- 16:13 **國際** 美警鬧烏龍！查毒品查到「自己人」還互毆
- 16:13 **政治** 公教新制退撫給與專戶上路 即日起可至台銀...
- 16:13 **娛樂** 佼佼問「復台言承旭了喔」 志玲4天都沒回...

點選檢查

# 觀察HTTP 請求與返回內容

1. 選擇 **Network** 頁籤
2. 點選 **Doc**
3. 點選 **realtime/**



# 什麼是GET?



GET  
內容寫在上頭

<https://tw.appledaily.com/new/realtime>

# Requests

---

## Requests

網路資源(URLs)擷取套件

改善Urllib2 的缺點，讓使用者以最簡單的方式獲取網路資源  
可以使用REST操作(POST, PUT, GET, DELETE)存取網路資源



# 使用requests.get

```
import requests  
res = requests.get('https://tw.appledaily.com/new/realtime')  
print(res)  
#print(res.text)
```



請求 Request



回應 Response



# 以台灣高鐵為例

← → ↺ 〡 www.thsrc.com.tw/tw/TimeTable/SearchResult

台灣高鐵  
TAIWAN HIGH SPEED RAIL

繁體中文 | 日本語 | English

優惠活動 購票資訊 乘車指南 關於高鐵 高鐵假期 24h 網路訂票

首頁 > 購票資訊 > 快速查詢 > 時刻表與票價查詢

字體大小 [大] [中] [小] 列印本頁

## 時刻表與票價查詢

### 請選擇查詢條件

出發站: 台北站

日期: 2014/06/18

到達站: 嘉義站

時間: 10:30 出發

立即查詢

#### 一週內時刻表

- 南下時刻表
- 北上時刻表

#### 時刻表下載

- 2014舊版時刻表
- 2013/12/23起適用時刻表

### 您的查詢結果

台北站 ▶ 嘉義站 2014/06/18(周二) 10:30 出發

檢視 2014/06/18 時刻表 (含南下/北上車次)

# 什麼是POST?

**StartStation:**

977abb69-413a-4ccf-a109-0272c24fd490

**EndStation:**

fbd828d8-b1da-4b06-a3bd-680cdca4d2cd

**SearchDate:**

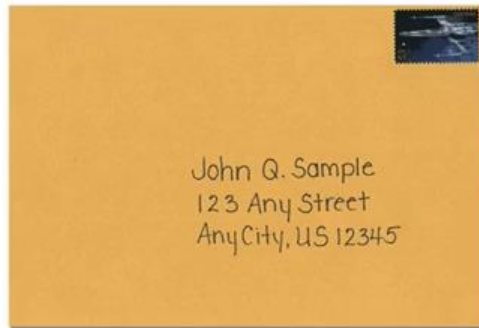
2015/04/19

**SearchTime:**

17:30

**SearchWay:**

DepartureInMandarin



POST

內容寫在信紙，包在信封內

<https://www.thsrc.com.tw/tw/TimeTable/SearchResult>

# 使用requests.post

---

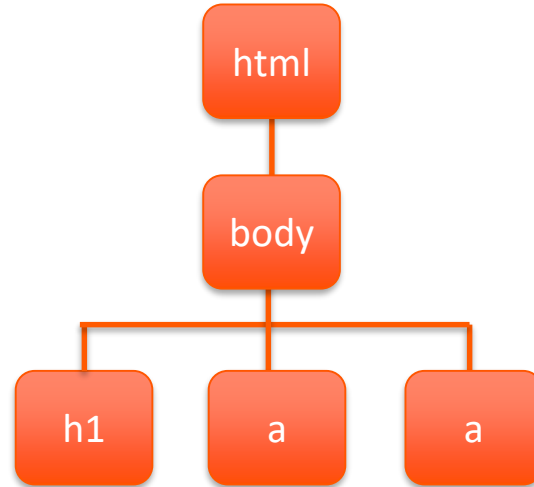
```
import requests
payload = {
    'StartStation':'977abb69-413a-4ccf-a109-0272c24fd490',
    'EndStation':'60831846-f0e4-47f6-9b5b-46323ebdcef7',
    'SearchDate':'2018/05/25',
    'SearchTime':'10:30',
    'SearchWay':'DepartureInMandarin'
}
res = requests.post('http://www.thsrc.com.tw/tw/TimeTable/SearchResult',
    data=payload)
print(res.text)
```

# DOM Tree

---

```
<html>
<body>
<h1 id="title">Hello World</h1>
<a href="#" class="link">This is link1</a>
<a href="# link2" class="link">This is link2</a>
</body>
</html>
```

Document Object Model



# 使用BeautifulSoup4

---

可以用來剖析及萃取 HTML的內容

會自動將讀入的內容轉換成UTF-8編碼

底層使用lxml及html5lib，可以使用不同的剖析函式以取得速度與彈性的平衡  
`BeautifulSoup(html_sample, 'lxml')`



可抽換Parser

<https://www.crummy.com/software/BeautifulSoup/bs4/doc.zh/>



# BeautifulSoup 範例

---

將網頁讀進BeautifulSoup 中

```
from bs4 import BeautifulSoup
```

```
html_sample = '''
```

```
<html>
```

```
<body>
```

```
<h1 id="title">Hello World</h1>
```

```
<a href="#" class="link">This is link1</a>
```

```
<a href="# link2" class="link">This is link2</a>
```

```
</body>
```

```
</html>'''
```

```
soup = BeautifulSoup(html_sample, 'lxml')
```

```
print(soup.text)
```

# 取出h1 標籤的資料

---

使用select\_one 找出唯一含有h1 標籤 的元素

```
soup = BeautifulSoup(html_sample, 'lxml')  
title = soup.select_one('h1')  
print(title)
```

# 找出所有含a tag 的HTML 元素

---

使用 select 找出(第一個)含有a tag 的元素

```
soup = BeautifulSoup(html_sample, 'xml')  
alink = soup.select('a')  
print(alink)
```

select 的結果會存放在list 中

# 取得含有特定ID的元素

---

使用select 找出所有id為title的元素

```
alink = soup.select('#title')  
print(alink)
```

id 前面必須加上 #

# 取得含有特定class的元素

---

使用select 找出所有class為link的元素

```
soup = BeautifulSoup(html_sample, 'lxml')  
for link in soup.select('.link'):  
    print(link)
```

class 前面必須加上 .

# 取得所有a tag 內的連結

---

使用select找出所有a tag 的href 連結

```
alinks = soup.select('a')  
for link in alinks:  
    print(link['href'])
```



# 觀察元素抓取位置

## 1. 點選元素觀測

2016 / 03 / 24

【更新】中華電4G新頻段開通 300M飆... (10661)

海預報APP 觀星族也愛用(0)

2. 觀察元素所在位置 `<li class= "rtdtdt..." >`

3. 下方可以觀察標籤路徑

html > body > #article.all > div.wrapper > div.sqzezer > div.soil > article > #maincontent > vertebrae > div.thoracis > div.abdominis > rlyby > clearmen > ul.rtdtdt.slv1 > li.rtdtdt.ccc > a

# 抓取第一頁的新聞

---

#抓取新聞列表原始碼

```
import requests
```

```
from bs4 import BeautifulSoup
```

```
res = requests.get('https://tw.appledaily.com/new/realtime')
```

#用迴圈遍歷含有.rtdtd a的元素

```
soup = BeautifulSoup(res.text , 'lxml')
```

```
for news in soup.select('.rtdtd a'):
```

```
    print(news)
```

# 根據不同HTML標籤取得對應內容

---

```
for news in soup.select('.rtddt a'):
    if news.select_one('h1'):
        h1      = news.select('h1')[0].text
        h2      = news.select('h2')[0].text
        time    = news.select('time')[0].text
        print(h1,h2,time)
```

標題: h1  
類別: h2  
時間: time

# 用 Python 破解各式網站 – 小菜一碟

- 1 AJAX 生成頁面
- 2 需要登入的頁面
- 3 需要Header 或 Cookie 的頁面
- 4 會擋IP的頁面
- 5 會擋使用者頻繁存取的頁面
- 6 需要Hidden Input 的頁面
- 7 抓取圖表、影音資料
- 8 需要圖形辨識的頁面
- 9 需要輸入驗證碼的頁面

爬不是問題  
只要看的到，就能爬的到



大數學堂

<https://www.largitdata.com>



# 輿情分析系統

---

# 智能監測



# InfoMiner即時輿情分析平台



## 全球範圍監測

- 台灣來源超過3萬以上頻道
- 涵蓋中、美、歐、日本等國外來源

## 自主彈性最高

- 不限定關鍵字修改次數
- 可新增指定來源



## 數據提供速度最快

- 發文10分鐘即收錄
- 每日爬文不中斷



# 利用分散式架構爬蟲的提升服務品質

# 解決問題

- 一有相關訊息，就能告知權責單位相關情資，避免公關危機產生
- 能摘要相關情資，方便使用者快速理解發展情勢



## 需求

- 1 每五分鐘更新一次所有網站資料
- 2 即時能使用不同關鍵字搜尋相關輿情
- 3 分散式機器學習運算

# 優秀駭客團隊

大數軟體開發 > 「InfoMiner 即時輿情分析平台」

## 助客戶迎接大數據時代 開啟市場商機



掌握「數據為王」的時代來臨，台灣新創公司大數軟體深厚的技術背景，創新研發「InfoMiner 即時輿情分析平台」，成功協助政府、法人機構及企業精準網路資訊、進行輿情分析，進而做出正確決策，符合市場需求的大數據分析平台，使大數軟體快速打開知名度與業務，贏得數據分析商機。

網路科技無所不在，引領大數據的時代來臨。不論是政府機構或民間企業，都知道掌握網路的龐大數據資訊和輿論風向，才能正確掌握組織的各種策略。政治人物可以藉此判斷選情，企業可以挖掘吸引忠實顧客的最佳行銷策略，政府機構可以推助符合人民需求的政策方向。

了解到輿情分析的重要性，創立於 2014 年的大數軟體創新研發「InfoMiner 即時輿情分析平台」，透過提供即時輿情分析的 B2B 服務，協助各類組織分析網路資訊，找

到正確的策略方向。符合市場需求的創業主願，讓大數軟體快速打開數據分析商機，成為國內平台新創又一潛力新星。

### 找到最合適夥伴 走上創業坦途

大數軟體創辦人丘祐璋原本任職於趨勢科技，當時基於創新實驗應用的興趣，與好友組隊參加台灣雲端運算產業協會舉辦的第一屆「台灣雲谷雲谷育成計畫」競賽，團隊開發出的「InfoLite 即時網頁蒐集系統」深受評審青睞，獲得台灣大哥大的輔導，也讓他們興起創業的念頭。

就這樣，原本都在大公司擁有穩定工作的幾個人，辭去工作、共同成立碩源資訊。隔年又以「InfoLite」系統前往北京中關村參加中國雲計算資料創新大會，一舉獲得首獎殊榮，屢獲媒體競賽肯定，原該走上創業的坦途，沒想到後來因為創業團隊理念不合，最後走向解散的命運。

當時丘祐璋覺得「InfoLite」系統具備市場發展潛力，公司就此解散非常可惜，因為不想放棄創業之路，於是在 2014 年 7 月成立大數軟體。「為了讓大家知道我們還在創業



使用者只需透過帳號訂閱，以月繳方式，便可以使用 InfoMiner 即時輿情分析平台的技術分析網路輿情。由於採取月租收費的商業模式，讓大數軟體的營收，跟著採用 InfoMiner 即時輿情分析平台的客戶數增加，不斷向上攀升。

—大數軟體有限公司

### 大數軟體 有限公司

類別	新創企業獎
負責人	丘祐璋
成立時間	2014 年
主要業務	InfoMiner 即時輿情分析平台
員工人數	8 人
獲獎育成中心	財團法人時代基金會 Garage+ 育成中心



LargeData

這條路上，也想到自己宣告大數軟體是一個新的開始。大數軟體成立時特別舉辦一場酒會。」丘祐璋回憶，酒會來了很多好友，大數軟體技術長（CTO）梁百祥當時還在菲律賓一家新創公司任職，剛好回台也來參加酒會，丘祐璋詢問梁百祥要不要加入大數軟體，梁百祥表示有意願，不過原本的工作無法走

開，過了一段時間回國，才加入大數軟體。

丘祐璋回憶，大數軟體剛成立時，自己一人單打獨鬥。一開始先以工程師為對象開設數據、機密學習等教育訓練課程，收費收維持公司運作，這段期間持續與梁百祥接觸，並且有一些合作，也一起討論輿情系統如何搜集資料等，後來梁百祥正式加入團隊，兩

# We Are Hiring – 資料爬取實習生

104人力銀行

## 資料爬取實習生

大數軟體有限公司 本公司其他工作

☆ 儲存

✉ 應徵

30人以上應

計算屬於你的工作適合度 [請點此登入](#)

### 工作內容

對資料分析有興趣嗎? 想要更深入了解該怎麼利用數據分析、機器學習，從資料中掏金嗎?  
加入大數軟體或許正就是協助你/妳快速如何運用大數據商業應用的最佳途徑。

大數軟體是一以大數據服務為主軸的公司，建立輿情觀測服務InfoMiner，協助企業、政府客戶蒐集、分析網路輿情，讓客戶掌握第一手資訊，洞燭先機。我們另有協助企業與政府機關分析數據以及建立資料服務流程。客戶涵蓋政府機關、學校、金融、壽險、科技業、資訊服務業。

經濟日報報導: <https://money.udn.com/money/story/8889/2321156>  
數位時代報導: <http://www.bnext.com.tw/article/view/id/39110>

**WE'RE  
HIRING!**

InfoMiner 輿情分析系統

瀏覽工作紀錄

[清除](#)

資料爬取實習生  
大數軟體有限公司

資料爬取實習生



使用網路爬蟲  
征服宇宙  
用資料發大財

---



# THANK YOU

---

EMAIL: [david@largitdata.com](mailto:david@largitdata.com)

網站: [www.largitdata.com](http://www.largitdata.com)

電話: 0929094381