
# Python 網路爬蟲

丘祐瑋 – David Chiu

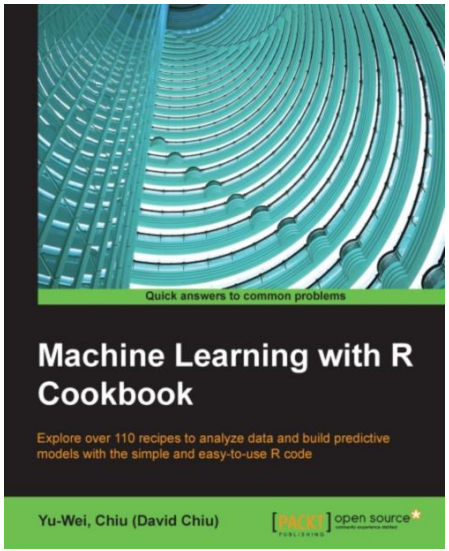
EMAIL: david@largitdata.com
網站: www.largitdata.com
電話: +886929094381

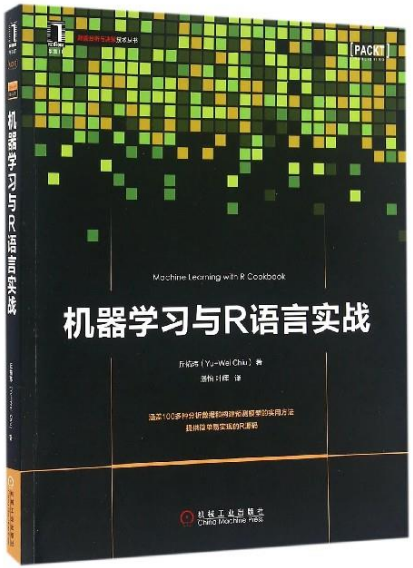
# 關於我

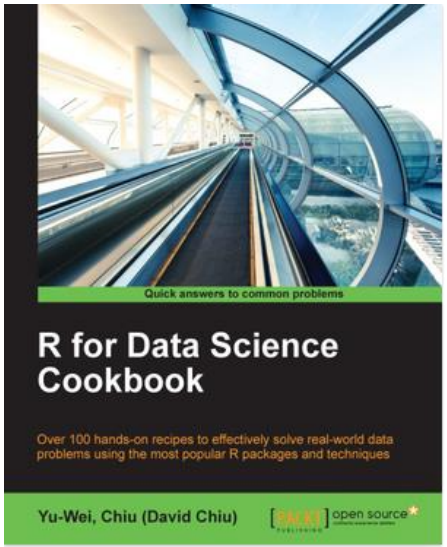

- 大數軟體有限公司創辦人
- 前趨勢科技工程師
- 大數學堂
https://www.largitdata.com/
- 粉絲頁
https://www.facebook.com/largitdata
- R for Data Science Cookbook
https://www.packtpub.com/big-data-and-business-intelligence/r-data-science-cookbook
- Machine Learning With R Cookbook
https://www.packtpub.com/big-data-and-business-intelligence/machine-learning-r-cookbook

LargitData

# 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

LargitData

# 課程補充資料

投影片、程式碼放置於:
https://github.com/ywchiu/nccu_crawler

# 什麼是網路爬蟲?

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

Infer-ability

Model

Context

Connectedness

Variety

Variability

Velocity

Volume

Technologies

Roles

Infer, Predict, Recommend & Visualize

Contextualize, Model & Reason

De-complexify, Transform, Analyze & Network

D3.js, Dashboards, web apps

Store

R, Hive, Pig, python, Java, Mahout

Collect

Hadoop

ETL, Storm, Scribe, Flume,...

Machine Learning

Analytics

Data Architecture / Management

**Data Scientist**

# 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

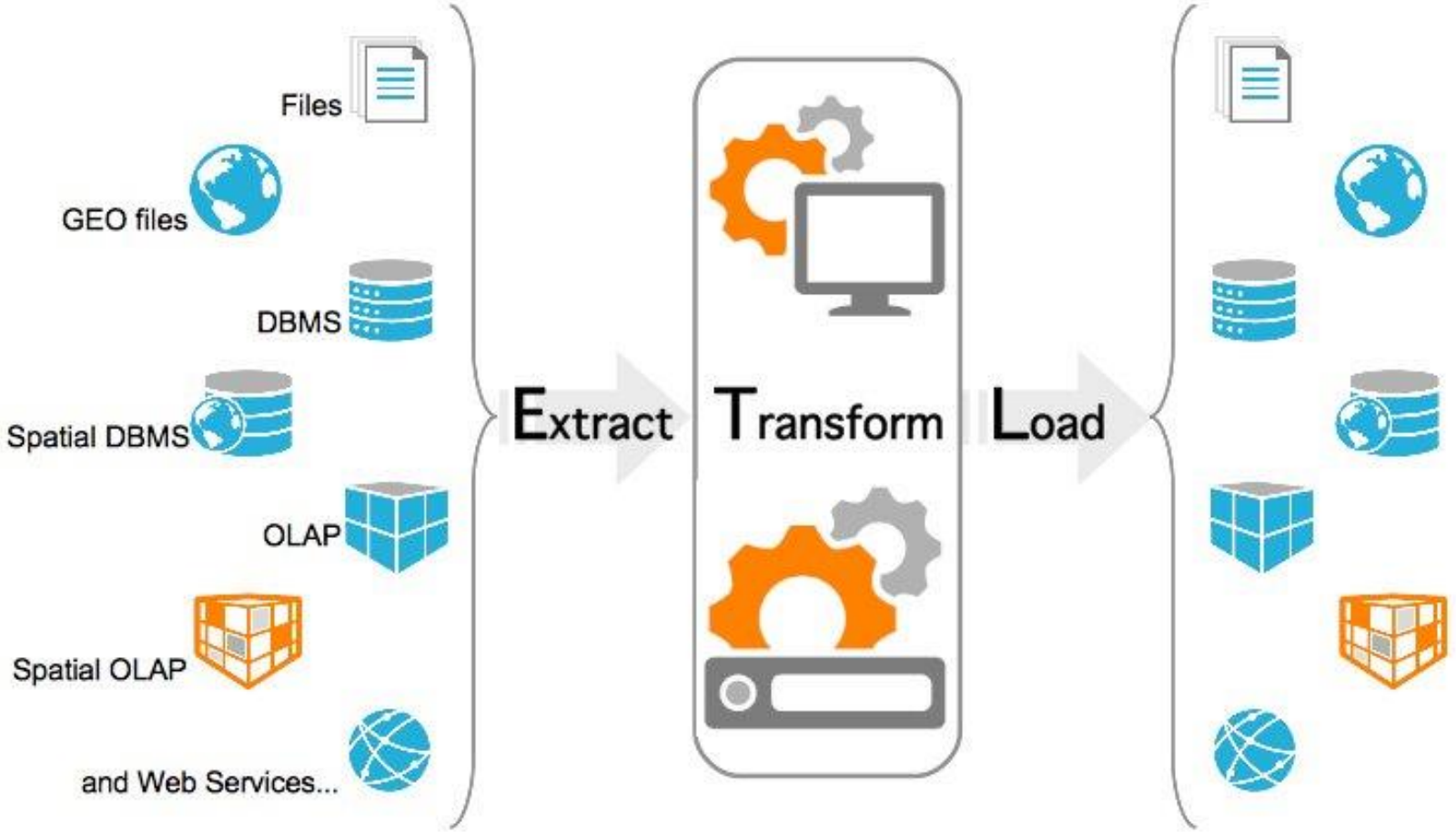
沒有固定的資料格式
例如網頁數據

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


如何從非結構化資料挖出價值資料是一大挑戰

LargitData

# 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


透由簡單的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 |

LargitData

# 為什麼要使用Python 做網路爬蟲?

網路爬蟲

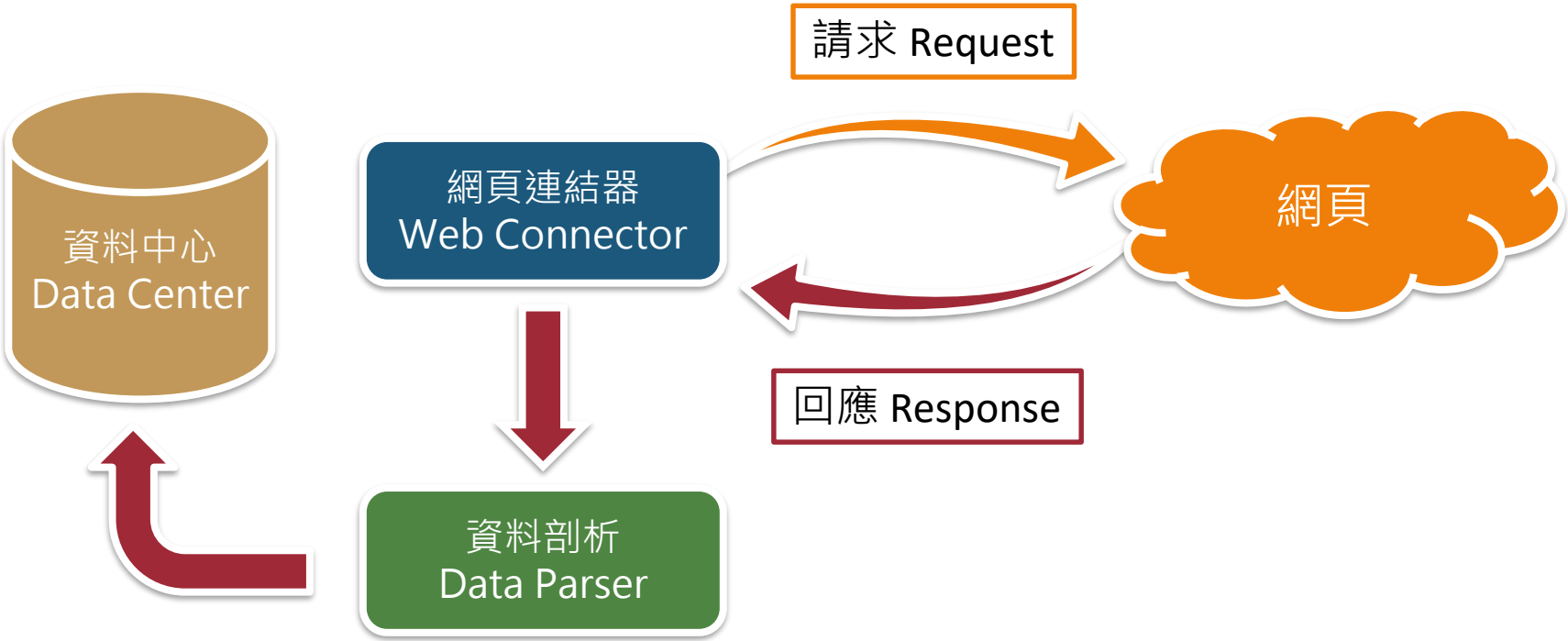

資料分析

網頁製作

**打造一條龍服務**

LargitData

# Python 爬蟲起手式



# 如何抓取蘋果即時新聞

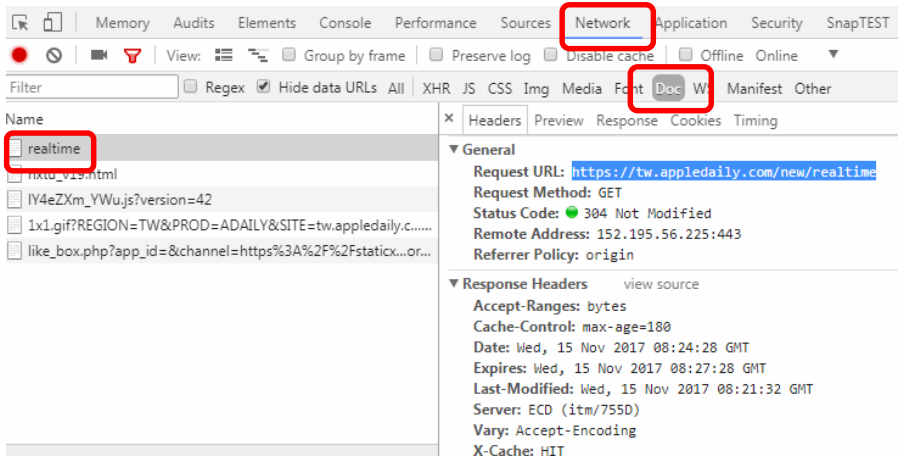
https://tw.appledaily.com/new/realtime

# 使用開發人員工具

於網頁上點選右鍵 -> 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


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
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
到達站：嘉義站
日期：2014/06/18
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暑期時刻表
2013/12/23起適用時刻表
您的查詢結果
台北站 ▸ 嘉義站 2014/06/18(周三) 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

**https://www.thsrc.com.tw/tw/TimeTable/SearchResult**


POST
內容寫在信紙，包在信封內

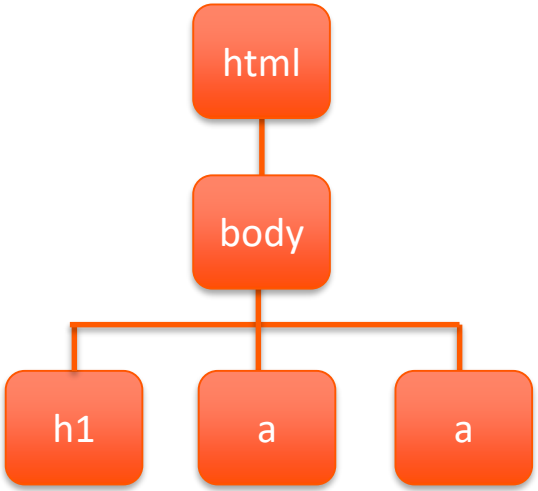
# 使用requests.post

```
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, 'lxml')
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
3C
【更新】中華電4G新頻段開通 300M飆...(10661)
tddt.ccc 640px × 40.7969px
觀星族也愛用(0)
Elements Console Sources Network Timeline Profiles Resources Security Audits Backbone

<h1 class="dddd">…</h1>
<ul class="rtddd slvl">
<li class="rtddt ccc">…</li>
<li class="rtddt life even">…</li>
<li class="rtddt life">…</li>
<li class="rtddt sp_ad even">…</li>
<li class="rtddt life">…</li>
<li class="rtddt inter even
<li class="rtddt local">…</
<li class="rtddt enter even
<li class="rtddt inter">…</
<li class="rtddt busi even">…</li>
<li class="rtddt life">…</li>
<li class="rtddt local even hsv">…</li>
<li class="rtddt life">…</li>
<li class="rtddt polit even">…</li>
<li class="rtddt life hsv">…</li>
<li class="rtddt property even">…</li>
2. 觀察元素所在位置 <li class= “rtddt…” >
3. 下方可以觀察標籤路徑
html body#article.all div.wrapper div.sqrezeer div.soil article#maincontent.vertebrae div.thoracis div.abdominis.rlby.clearmen ul.rtddd.slvl li.rtddt.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.rtddt a的元素
soup = BeautifulSoup(res.text , 'lxml')
for news in soup.select('.rtddt 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


危機管理
即時監控、避免危機擴大
行銷策略
社群活動追蹤,廣告投放方向及精準行銷
品牌抱怨
即時社群消毒
市場探查
了解產業、指標性品牌、或潛在商機
敵情探查
了解對手活動及社群策略
社群挖礦
意見領袖,粉絲獲取,話題關鍵字

# InfoMiner即時輿情分析平台


自動發送警示通知



自設關鍵字組條件

上萬社群媒體收集

視覺化多樣分析
呈現

即時整合分析
發文10分鐘即
掌握

## 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. 每五分鍾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

LargitData

# 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、法人機構及企業搜尋網路資訊、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。

網路科技集所不在，引爆大數據的時代來臨，不管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，政治人物可以藉此洞察選情，企業可以擬訂吸引受眾者眼球的最佳行銷策略，政府機構可以推動符合人民需求的施政方向。

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，協助各種組織分析網路資訊，找到正確的策略方向，符合市場需求的創業主題，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強力新星。

### 找到最適夥伴 走上創業坦途

大數軟體創辦人丘祐瑋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「台灣雲谷雲的育成計畫」競賽，團隊所發出的「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，也讓團隊開起創業的念頭。

就這樣，原本都在大公司擁有穩定工作的幾個人，辭去工作、共同成立瑜源資訊，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，更獲獲得競賽肯定，原就走上創業的坦途，沒想到後來因為創業團隊理念不合，最後走向解散的命運。

當時丘祐瑋覺得「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，「為了讓大家知道我還在創業

| 大數軟體有限公司 | |
|---|---|
| 型　別 | 新創企業類 |
| 負 責 人 | 丘祐瑋 |
| 成立時間 | 2014 年 |
| 主要業務 | InfoMiner<br>即時輿情分析平台 |
| 員工人數 | 8 人 |
| 孵育育成中心 | 財團法人時代基金會<br>Garage+ 育成中心 |

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，源源著採用 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

協媒路上，也都跟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，」丘祐瑋回憶，酒會來了許多好友，大數軟體技術長（CTO）樊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，丘祐瑋就問樊百祥要不要加入大數軟體，樊百祥表示有意願，不過原本的工作還走不開，過了一段時間回國，才加入大數軟體。

丘祐瑋回憶，大數軟體剛成立時，自己一人單打獨鬥，一開始先以工程師背景開設大數據、機器學習等教育訓練課程賺取營收維持公司運作，也與開始與樊百祥接觸，並且有一些合作，也一起討論輿情系統如何搜集資料等，後來樊百祥正式加入團隊，兩

LargitData

# We Are Hiring – 資料爬取實習生

104人力銀行

## 資料爬取實習生

儲存　應徵

大數軟體有限公司　本公司其他工作

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


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

LargitData

使用網路爬蟲
征服宇宙
用資料發大財

# THANK YOU

EMAIL: david@largitdata.com
網站: www.largitdata.com
電話: 0929094381