

Python網路爬蟲

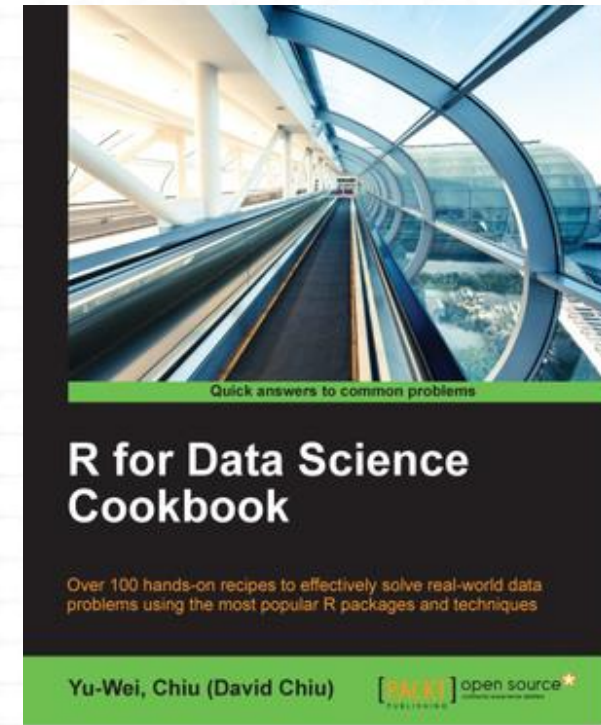
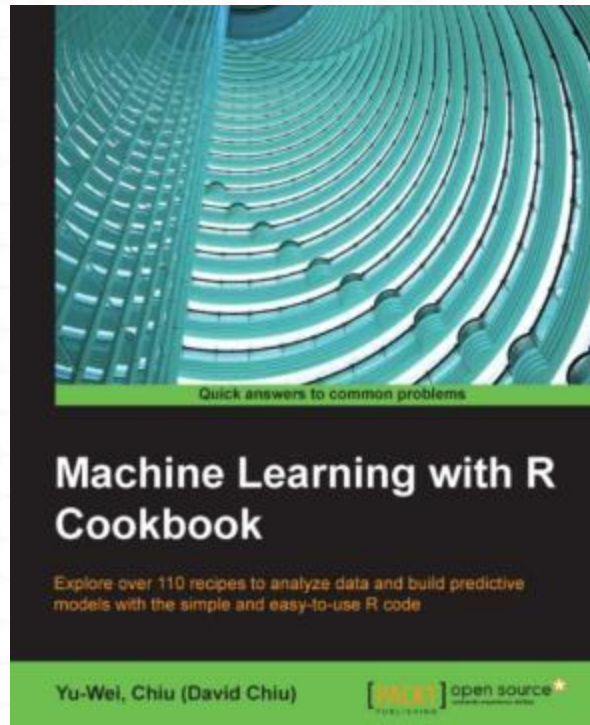
David Chiu

關於我



- 大數軟體有限公司創辦人
- 前趨勢科技工程師
- ywchiu.com
- 大數學堂
<http://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>

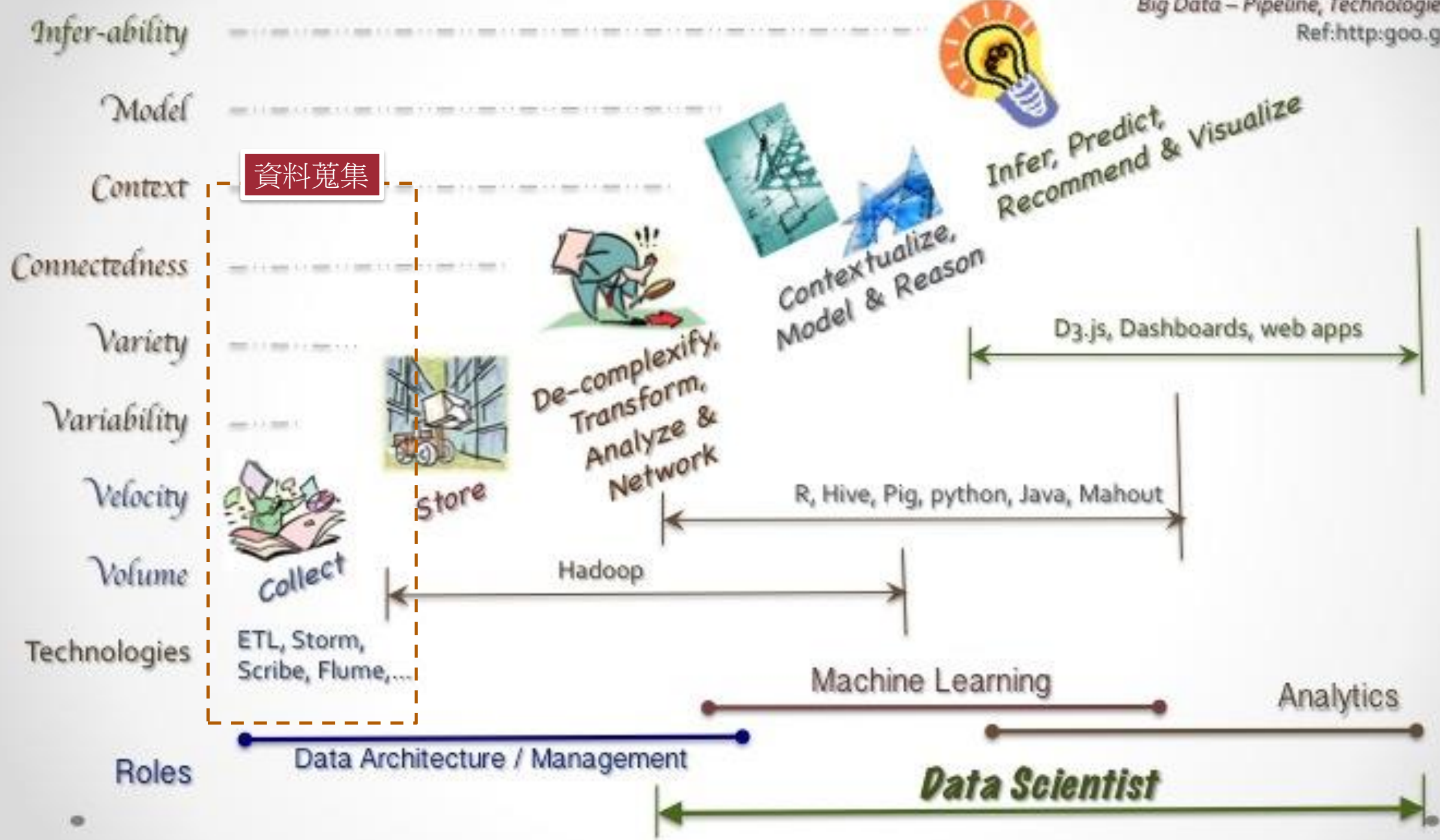
Machine Learning With R Cookbook (机器学习与R语言实战) & R for Data Science Cookbook (数据科学：R语言实现)



Author: David (YU-WEI CHIU) Chiu

課程資料

- 所有課程補充資料、投影片皆位於
 - ▣ <https://github.com/ywchiu/cathaysitecrawler>



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

結構化資料

■ 資料有固定的欄位與格式

□ 例如：資料庫表格中所存放的資料

| id | title | content | time | view_cnt | category |
|----|--------------------------|----------------------|---------------------|----------|----------|
| 56 | Uni-girls成立經紀公司 將選拔5位... | 統一7-ELEVEn獅隊旗下Uni... | 2016-06-17 16:11:00 | 0 | 體育 |
| 57 | 販售工業用劣質油 鑫好企業判6... | 鑫好企業負責人吳容合被... | 2016-06-17 16:10:00 | 0 | 社會 |
| 58 | 義大回嗆中職會長：聯盟應深切檢討 | 中華職棒會長吳志揚今天... | 2016-06-17 16:09:00 | 0 | 體育 |
| 59 | 【就職滿月】蔡英文滿意度4成7 ... | 總統蔡英文就職即將滿月... | 2016-06-17 16:08:00 | 0 | 政治 |
| 60 | 【有片】素珠之亂有續集 洪家賓... | (更新：新增影片)自稱「... | 2016-06-17 16:08:00 | 36312 | 社會 |
| 61 | 昔日老毛鬥爭工具 「中央專案組... | 香港銅鑼灣書店店長林榮... | 2016-06-17 16:08:00 | 0 | 國際 |
| 62 | 【有片】北市府給雪隧建議 林聰... | (更新:新增動新聞)台北市... | 2016-06-17 16:07:00 | 5067 | 政治 |

■ 可以下SQL 處理與撈取資料

□ `select title, content from newsmain;`

半結構化資料 - XML

- 可以使用欄位存取資料內容
- 欄位不固定，例如Mary 就少了age的欄位
- 可以彈性的存放各種欄位格式的資料

```
<users>
  <user>
    <name>Q00</name>
    <gender>M</gender>
    <age>12</age>
  </user>
  <user>
    <name>Mary</name>
    <gender>F</gender>
  </user>
</users>
```


半結構化資料 - JSON

- 如同XML可以使用欄位存取資料內容
- 使用Key:Value存放資料
- 不用宣告欄位的結尾，可以比XML更快更有效傳輸資料

```
[  
  user:{  
    name:QOO,  
    gender:M,  
    age:12,  
  },  
  user:{  
    name:Mary,  
    gender:F  
  }  
]
```

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571個房屋

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-四維路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 / 98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

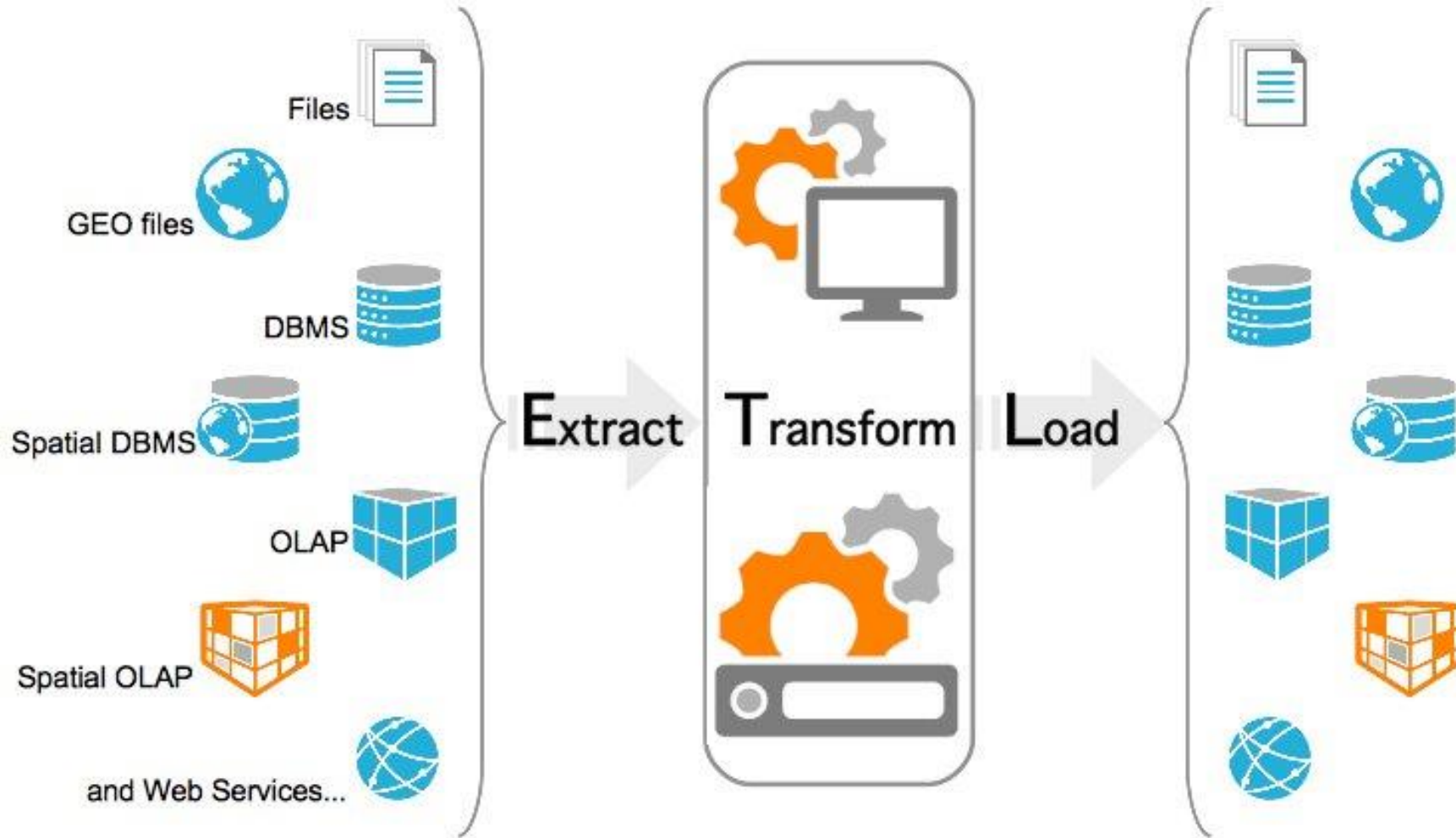
大安區-樂利路

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



網路爬蟲

目標：將非結構化數據轉變為結構化數據

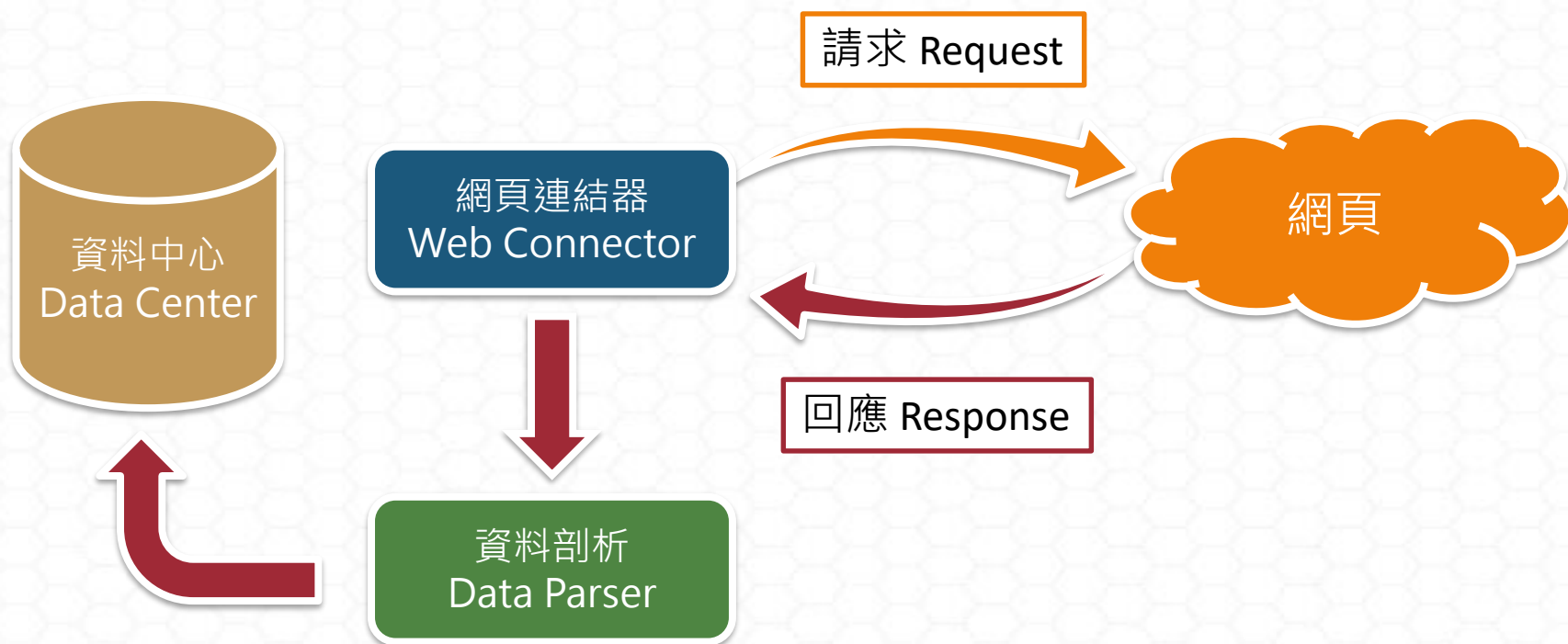


透由簡單的SQL語句
從結構化資料中
達到簡單的分析目的



| 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 |

爬蟲是怎麼運作的





資料蒐集

如何抓取蘋果即時新聞

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

The screenshot shows the Apple Daily Realtime website. At the top, there's a navigation bar with the Apple Daily logo and various category links like 體育 (Sports), 娛樂 (Entertainment), 時尚 (Fashion), 生活 (Life), 社會 (Society), 國際 (International), 財經 (Finance), 地產 (Real Estate), 政治 (Politics), 論壇 (Forum), and 陣線 (Frontline). Below this is a search bar and a navigation menu with categories like 動即時 (Realtime), 最新 (Latest), 焦點 (Focus), 熱門 (Popular), 動物 (Animals), FUN, 僑哈 (Expats), 摺奇 (Curiosities), 影片 (Videos), 正妹 (Hot Girls), and 體育 (Sports). A banner for Gmail 商用版 (Gmail Business) is visible. The main headline is about the Beijing government's agreement to complete the 'Big Egg' project by the end of July. To the right, there's an advertisement for a medical service. Below the main headline, there's a section for '動即時' (Realtime) with a date of 2015/05/25 and a time of 23:00. A button labeled '按 看蘋果' (Click to see Apple Daily) is also present.

昨日瀏覽量: 17466897 • 蘋果日報自律委員會 • 香港

動即時 最新 焦點 熱門 動物 FUN 僑哈 摺奇 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 為初代客製成電子郵件。30 天免費試用，即刻體驗！

北京市府同意遠雄
7月底前完成大巨蛋大底工程

醫靠 無界
熟醫界醫生之旅 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 Standstill Fees
✓ 0% Transaction Fees*
✓ Unlimited SKU's
✓ Unlimited Storage
✓ \$100 Accounts 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 the Apple Daily (蘋果日報) website. At the top, there is a navigation bar with categories like 最新 (Latest), 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 (Hot Line), 3C, 動物 (Animals), 副刊 (Supplement), 體育 (Sports), and 財經地產 (Finance and Real Estate). The main headline is "原來她感情路是這樣" (It turns out her love life is like this). Below the headline, there is a list of news items with timestamps and category tags. A right-click context menu is open over the news list, showing various browser actions. The "檢查 (N)" option, which corresponds to the keyboard shortcut Ctrl+Shift+I, is highlighted with a red box. To the right of the menu, there is a small sidebar with the Apple Daily logo and a "說這專頁" (Say this page) button. Below the sidebar, there is a small text "其實沒差多少，但..." (Actually not much difference, but...). At the bottom right of the screenshot, there is an orange button with the text "點選檢查" (Click to inspect).

2017 / 11 / 15

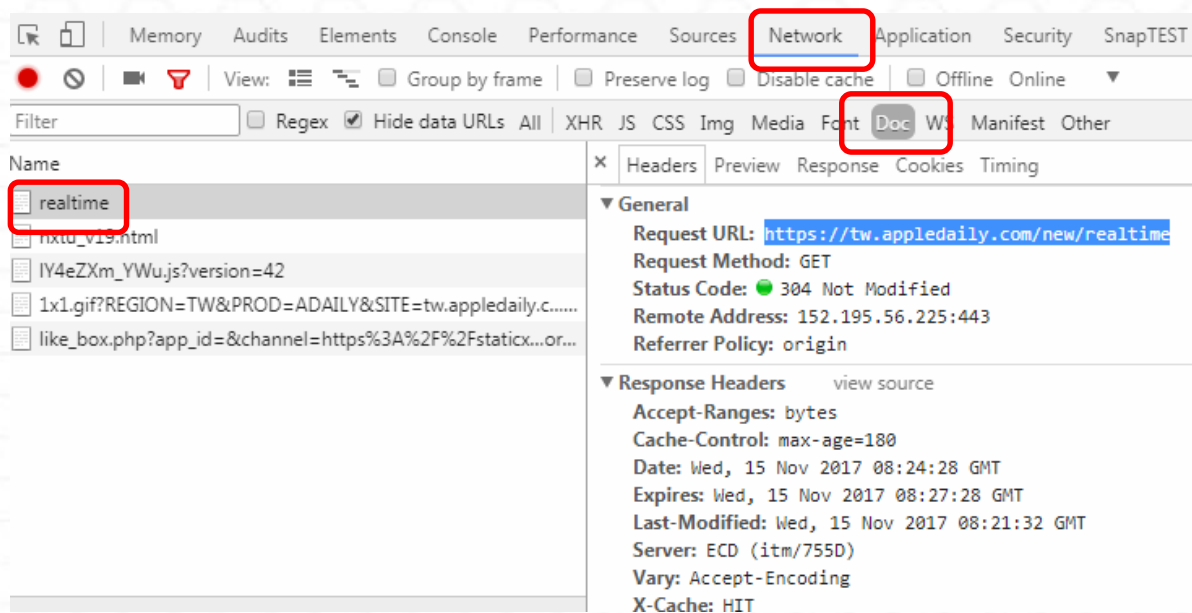
- 16:19 **生活** 兩岸高職教育交流 第五屆台蘇論壇聖約大...
- 16:18 **社會** 光電廠解雇工會理事長 桃產總聲援抗議
- 16:17 **生活** 縮胃減去75公斤 換身分證差點被拒
- 16:15 **社會** 賴清德走了 南市開始放焰火還傷3人
- 16:15 **政治** 賴清德：慶富案是在2014年 哪一朝發生...
- 16:14 **社會** 油漆掉落灑5公尺路面 1騎士滑倒幸輕傷
- 16:13 **國際** 美警鬧烏龍！查毒品查到「自己人」還互毆
- 16:13 **政治** 公教新制退撫給與專戶上路 即日起可至台銀...
- 16:13 **娛樂**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

檢查 (N) Ctrl+Shift+I

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

← → C 台 www.thsrc.com.tw/tw/TimeTable/SearchResult

台灣高鐵路
TAIWAN HIGH SPEED RAIL

繁體中文 | 日本語 | English

優惠活動 購票資訊 乘車指南 關於高鐵路 高鐵路假期 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 f 世 台 P

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 立即查詢

到達站: 嘉義站 時間: 10:30 出發

一週內時刻表 時刻表下載

- 南下時刻表
- 北上時刻表
- 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:

fb828d8-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':'2018/10/29',
    'SearchTime':'10:30',
    'SearchWay':'DepartureInMandarin'
}
res = requests.post('http://www.thsrc.com.tw/tw/TimeTable/SearchResult', data=payload)
print(res.text)
```

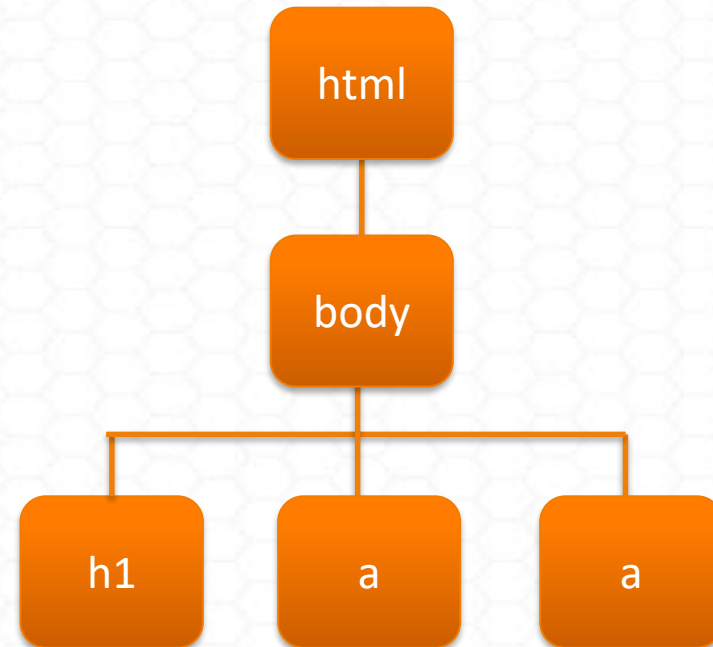


資料剖析

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

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'])
```

尋找CSS 的定位

- Chrome 開發人員工具

- Firefox 開發人員工具

- InfoLite

- <https://chrome.google.com/webstore/detail/infolite/ipjbadabbpedegielkhgpi/ekdlmfpgal>

資料蒐集實務

連結到蘋果即時新聞的頁面

使用requests.get 取得頁面內容



使用開發者工具檢視每則新聞的分隔
發現都以.rtd dt a做分隔

觀察元素抓取位置

1. 點選元素觀測

2016 / 03 / 24

16:11 3C 【更新】中華電4G新頻段開通 300M頻..(10661)

海預報APP 觀星族也愛用(0)

Elements Console Sources Network Timeline Profiles Resources Security Audits Backbone

```
<div class="abdominis rlby clearmen">
  <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>
    <li class="rtddt local">...</li>
    <li class="rtddt enter even">...</li>
    <li class="rtddt inter">...</li>
    <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>
    <li class="rtddt local">...</li>
  </ul>
</div>
```

html body#article.all div.wrapper div.sqzezer div.soil article#maincontent.vertebrae div.thoracis div.abdominis.rlby.clearmen ul.rtddd.slvl li.rtddt.ccc a

2. 觀察元素所在位置 `<li class= "rtddt..." >`

3. 下方可以觀察標籤路徑

使用InfoLite

- <https://chrome.google.com/webstore/detail/infolite/ipjbadabbpedegielkhgpiekdmlmfpgal>



綠色: 目前選取得區塊
黃色: 符合樣式的區塊
紅色: 排除的曲塊
Clear旁邊的數字: 符合區塊的數目

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('http://www.appledaily.com.tw/realtimenews/section/new/')
```

#用迴圈遍歷含有.rtdtdt 的元素

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

```
for news in soup.select('.rtdtdt'):
```

```
    print(news)
```

取得標題

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

```
▼<a href="/realtime/news/article/3c/20160324/823044/【更新】中華電4G新頻段開通 300M纜網" target="_blank" class>
  <time>16:11</time>
  <h2>3C</h2>
  ▼<h1>
    <font color="#383c40">【更新】中華電4G新頻段開通 300M纜... (10661)</font>
  </h1>
</a>
```

標題: h1

任務:根據不同標籤取得不同細部資料

■ 請抓取蘋果新聞列表中每篇新聞的:

- 新聞類別
- 新聞連結
- 時間

回顧抓取清單列表流程詳解

1. 找尋資料清單頁面連結

2. 使用requests抓取該清單頁面原始碼

3. 根據列表(或區塊)分隔 .rtddt a

- ❑ 抓取標題 h1
- ❑ 抓取時間 time
- ❑ 抓取類別 h2
- ❑ 抓取連結

抓取列表的時間、種類、標題、連結

| | | |
|-------|----|---------------------------------|
| 16.11 | 3C | 【更新】中華電4G新頻段開通 300M顯... (10661) |
| 16.10 | 生活 | 首創雲海預報APP 觀星族也愛用(0) |
| 16.10 | 生活 | 【民報】被吳念真改變人生的金鐘男主角陳竹... (0) |
| 16.10 | 特企 | 【特企】宏正重操舊業 靠銀彈網技現身新莊... (2645) |
| 16.09 | 生活 | 梅山鄉櫻花生病 竟是人禍導致(0) |
| 16.08 | 國際 | 反難民政客 反被難民救(0) |
| 16.07 | 社會 | 【更新】霧氣朱雪瑋4時軟掉 因開竊聽到這... (60570) |
| 16.05 | 娛樂 | 黃仲嘉父子曾同台鋼鐵 竟看上她收場遺憾(234) |
| 16.04 | 國際 | 【更新】比利時媒體：地獄監視器 拍到另一... (40703) |
| 16.04 | 財經 | 央行降準預期心理 台股逆勢1角收32.7... (572) |
| 16.04 | 生活 | 【央廣RT】長者聽力損失 日後失智比例... (0) |
| 16.02 | 社會 | 【更新】拖板車飛跨安全島 大貨車側翻被... (2259) |

取得蘋果新聞內文

列表的資訊比較簡單

| | | |
|----------------|----|-----------------------------|
| 2016 / 04 / 05 | | |
| 16:40 | 體育 | 陳冠宇明一軍先發對軟銀鷹(0) |
| 16:34 | 國際 | 【法廣RFI】菲美軍演 美機動火箭系統初...(0) |
| 16:33 | 生活 | 清明連假最後一天 國道桃園路段順暢(6) |
| 16:33 | 體育 | 統一獅vs.Lamigo桃猿 桃猿尋求今...(59) |
| 16:32 | 生活 | 【更新】換肝更有效 術前先殺C肝病毒(1670) |
| 16:31 | 生活 | 妻與兩子都小腦萎縮 他用太鼓抗病魔(2570) |
| 16:30 | 社會 | 4P同志性愛趴 情趣環讓警開眼界(1073) |
| 16:30 | 娛樂 | 希臘拍《太陽》 宋仲基當成跟喬妹旅行(61) |
| 16:29 | 動物 | 【更新】哈利波特信差落巢 睜大眼賣萌(1980) |

生活

20160405 16:42
喝酒、吃嘉辣鍋第二天狂拉肚子 原因是這個(0)

20160405 16:33
清明連假最後一天 國道桃園路段順暢(6)

20160405 16:32
【更新】換肝更有效 術前先殺C肝病毒(1670)

20160405 16:31
妻與兩子都小腦萎縮 他用太鼓抗病魔(2570)

字級：A- A A+

f 分享FB g+ 分享g+ p 分享Plurk t 分享Twitter

2016年04月05日16:32 推 讚 2 G+ 0

(更新：新增影片)

高雄長庚醫院預計明天進行肝臟移植團隊的第1500例肝臟移植手術，患者是65歲來自嘉義的謝張女士，她自2001年就被診斷出患有C型肝炎，雖定期追蹤，也曾在2011年接受干擾素治療，但因復發宣告失敗，去年3月發現惡化罹患肝癌，5月到高雄長庚以腹腔鏡切除肝癌，但因有嚴重肝硬化及肝癌，醫師建議換肝保命，經比對決定由37歲的小兒子捐出約三分之一肝臟。

內文資訊比較豐富

摘取新聞標頭

主頁 即時 日報 新聞



最新 焦點 熱門 娛樂 藝攝樂 社會 國際 政治 生活 火線 3C 動

美商砸5百億向空巴訂430架機 史上最大

hgroup h1

建立時間：2017/11/15 16:27



美國私募基金公司「Indigo Partners」砸5百億向空巴訂430架飛機。法新社

位在h1 中

InfoLite Username Password Login X ?

Submit

h1 Clear (2) Add

- 1 台灣高中生一記魔球 引發國外網友討
- 2 台幻象戰機投共？陸國台辦：也可能
- 3 【有片】慘烈慎入！中國發生30車

看更多

最新的火線話題

國民年金改革爭議

擷取新聞標頭

```
detailurl = 'detailurl'  
res = requests.get(detailurl)  
soup = BeautifulSoup(res.text, 'lxml')  
print(soup.select('h1')[0].text)
```

美商砸5百億向空巴訂430架機 史上最大單

每日新聞摘要

[最新](#) [焦點](#) [熱門](#) [娛樂](#) [愛播網](#) [社會](#) [國際](#) [政治](#) [生活](#) [火線](#) [3C](#) [動物](#)



美國私募基金公司「Indigo Partners」砸5百億向空巴訂430架飛機。法新社

美國私募基金公司「Indigo Partners」與歐洲空中巴士（Airbus）將簽下430架飛機、近5百億美元（1.5兆元台幣）的破天荒大訂單，這也讓空巴在杜拜航空展與波音（BOEING）的較勁中「揚眉吐氣」，徹底逆轉勝。

根據路透及《彭博》報導，「Indigo Partners」創辦人法蘭克（Bill Franke）在周三（15日）飛往阿拉伯聯合大公國，消息人士指出，法蘭克會在當天的杜拜航空展上與空巴簽約，訂下430架A320neo窄體客機，將會是空巴史上最大筆的訂單。

「Indigo Partners」公司位在亞利桑那州鳳凰城，旗下擁有西哥廉航。（陳其仙／綜合外電報導）

跟上國際脈動，快來[蘋果國際](#)按讚

InfoLite

Username

Password

Login

X

?

▼

Submit

.ndArticle_margin p

Clear (1)

Add



新年懶人旅遊秘笈



賴揆新政

位在.ndArticle_margin p中

 蘋果粉絲團

擷取新聞摘要

```
print(soup.select('.ndArticle_margin p')[0].text)
```

美國私募基金公司「Indigo Partners」與歐洲空中巴士（Airbus）將簽下430架飛機、近5百億美元（1.5兆元台幣）的破天荒大訂單，這也讓空巴在杜拜航空展與波音（BOEING）的較勁中「揚眉吐氣」，徹底逆轉勝。根據路透及《彭博》報導，「Indigo Partners」創辦人法蘭克（Bill Franke）在周三（15日）飛往阿拉伯聯合大公國，消息人士指出，法蘭克會在當天的杜拜航空展上與空巴簽約，訂下430架A320neo窄體客機，將會是空巴史上最大筆的訂單。「Indigo Partners」公司位在亞利桑那州鳳凰城，旗下投資邊境航空及另一家墨西哥廉航。（陳其仙／綜合外電報導） 跟上國際脈動，快來蘋果國際按讚想知道更多，一定要看.....【杜拜航空展】空巴接1.5兆大單 史上最驚人【杜拜航空展】空巴超羞恥 等3天終於有訂單

取出時間資料

主頁 即時 日報 動新聞

蘋果即時 最新 焦點 熱門 娛樂 愛播網 社會 國際 政治 生活 火線 3C 動

美商砸5百億向空巴訂430架機 史上最大單

2017/11/15 16:21



InfoLite Username Password Login Submit

ndArticle_creat Clear (1) Add

1 台灣高中生一記魔球 引發國外網友討

2 台幻象戰機投共？陸國台辦：也可能

3 入！中國發生30車

看更多

最新的火線話題

國民年金改革爭議

位在.ndArticle_creat中

時間跟字串轉換範例

■ 引用套件

```
from datetime import date,datetime
```

■ 將時間轉變為字串

```
currenttime = datetime.now()  
print(currenttime.strftime("%Y-%m-%d"))
```

■ 將字串轉變為時間

```
a = '2018-10-26 14:00'  
print(datetime.strptime(a, "%Y-%m-%d %H:%M"))
```

時間格式轉換

```
from datetime import date,datetime
dt = soup.select('.ndArticle_creat')[0].text.split(' : ')[1]
print(datetime.strptime(dt, '%Y/%m/%d %H:%M') )
```

建立時間：2017/11/15 16:27



2017-11-15 16:27:00

取出人氣資料

主頁 即時 日報 動新聞

蘋果即時 最新 焦點 熱門 娛樂 愛播網 社會 國際 政治 生活

佼佼問「復合言承旭了喔」 志玲4 回...

1530 建立時間：2017/11/15 16:13

InfoLite Username Password Login X ?
Submit
_ndArticle_view Clear (1) Add

位在.ndArticle_view中

包偉銘兒開賓士撞6機車 賠200萬擺...

鼓鼓創作愷樂新歌 韓粉越聽越怒「太

看更多

最新的火線話題



任務:取出人氣與類別資訊

- 請抓取蘋果新聞內容頁中的:
 - 人氣資訊
 - 類別資訊

任務:完成完整爬蟲

- 請根據列表頁的連結抓取內文資料
- 將蘋果新聞前五頁的內文抓取下來

列表的資訊比較簡單

| | | | |
|----------------|----|-----------------------------|---|
| 2016 / 04 / 05 | | | |
| 16:40 | 體育 | 陳冠宇明一軍先發對軟銀鷹(0) | |
| 16:34 | 國際 | 【法廣RFI】菲美軍演 美機動火箭系統初...(0) | |
| 16:33 | 生活 | 清明連假最後一天 國道桃園路段順暢(6) | |
| 16:33 | 體育 | 統一獅vs.Lamigo桃猿 桃猿尋求今...(59) | |
| 16:32 | 生活 | 【更新】換肝更有效 術前先殺C肝病毒(1670) | ▶ |
| 16:31 | 生活 | 妻與兩子都小腦萎縮 他用太鼓抗病魔(2570) | ▶ |
| 16:30 | 社會 | 4P同志性愛臥 情趣環讓警開眼界(1073) | |
| 16:30 | 娛樂 | 希臘拍《太陽》 宋仲基當成跟喬妹旅行(61) | |
| 16:29 | 動物 | 【更新】哈利波特信差落巢 睜大眼賣萌(1980) | ▶ |



內文資訊比較豐富

生活 字級: A- A A+

2016/04/05 16:42
喝酒、吃麻辣鍋第二天狂拉肚子 原因是這(0)

2016/04/05 16:33
清明連假最後一天 國道桃園路段順暢(6)

2016/04/05 16:32
【更新】換肝更有效 術前先殺C肝病毒(1670)

2016/04/05 16:31
妻與兩子都小腦萎縮 他用太鼓抗病魔(2570)

分享FB 分享g+ 分享Plurk 分享Twitter

2016年04月05日16:32 傳送 讚 2 G+ 0

(更新:新增影片)

高雄長庚醫院預計明天進行肝臟移植團隊的第1500例肝臟移植手術，患者是65歲來自嘉義的謝張女士，她自2001年就被診斷出患有C型肝炎，雖定期追蹤、也曾在2011年接受干擾素治療，但因復發宣告失敗，去年3月發現惡化罹患肝癌，5月到高雄長庚以腹腔鏡切除肝癌，但因有嚴重肝硬化及肝癌，醫師建議換肝保命，經比對決定由37歲的小兒子捐出約三分之一肝臟，將在明天進行活體肝臟移植。

將資料轉到Pandas 的DataFrame

```
import pandas
newsdf = pandas.DataFrame(newsary)
newsdf.head()
```

抓取前五筆資料

| | category | popularity | time | title | url |
|---|----------|------------|-------|-------------------------|---|
| 0 | FUN | 0 | 00:28 | 神設計餐盤 讓料理好有戲 | http://www.appledaily.com.tw/realtimenews/arti... |
| 1 | 娛樂 | 0 | 00:28 | 梁詠琪噴嚏嚇哭女兒 愛犬舔baby腳底呼... | http://www.appledaily.com.tw/realtimenews/arti... |
| 2 | 動物 | 0 | 00:28 | 好不容易獲救 隔天後卻馬上... | http://www.appledaily.com.tw/realtimenews/arti... |
| 3 | 生活 | 0 | 00:27 | 美妝即期品開賣 每人只能買2件 | http://www.appledaily.com.tw/realtimenews/arti... |
| 4 | 娛樂 | 90 | 00:25 | 準爸爸 | |

Pandas 是Python 的資料處理套件
可與任意介面輕易接軌

Pandas

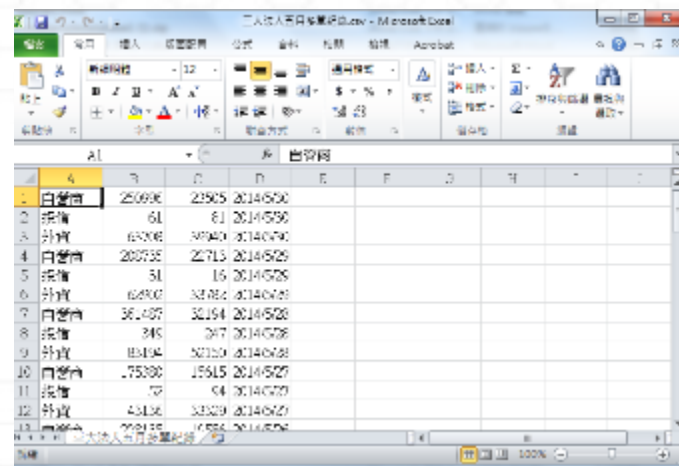
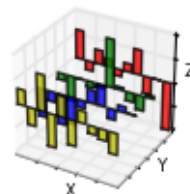
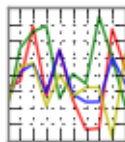
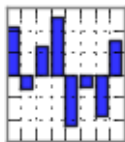
■ Python for Data Analysis

▣ 源自於R

▣ Table-Like 格式

▣ 提供高效能、簡易使用的資料格式(Data Frame)讓使用者可以快速操作及分析資料

pandas
 $y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

A screenshot of a Microsoft Excel spreadsheet. The data table has 13 rows and 4 columns. The first column contains labels in Chinese, the second and third columns contain numerical values, and the fourth column contains dates. The first row is highlighted in yellow.

| | A | B | C | D | E | F | G | H | I |
|----|----|--------|-------|-----------|---|---|---|---|---|
| 1 | 白雲 | 250996 | 23805 | 2014/5/20 | | | | | |
| 2 | 綠 | 61 | 81 | 2014/5/20 | | | | | |
| 3 | 外資 | 65006 | 30940 | 2014/5/20 | | | | | |
| 4 | 白雲 | 200735 | 25715 | 2014/5/20 | | | | | |
| 5 | 綠 | 51 | 15 | 2014/5/20 | | | | | |
| 6 | 外資 | 65006 | 30940 | 2014/5/20 | | | | | |
| 7 | 白雲 | 361487 | 32194 | 2014/5/20 | | | | | |
| 8 | 綠 | 245 | 247 | 2014/5/20 | | | | | |
| 9 | 外資 | 184194 | 32120 | 2014/5/20 | | | | | |
| 10 | 白雲 | 175300 | 15615 | 2014/5/20 | | | | | |
| 11 | 綠 | 72 | 54 | 2014/5/20 | | | | | |
| 12 | 外資 | 43436 | 33029 | 2014/5/20 | | | | | |
| 13 | 白雲 | 220135 | 17554 | 2014/5/20 | | | | | |

根據人氣指標排序

```
newsdf.sort_values(['clicked'], ascending=False).head()
```

| | category | clicked | dt | link | summary | title |
|----|----------|---------|---------------------|---|---|--------------------------|
| 47 | 娛樂 | 8054 | 2016-09-23 00:04:00 | http://www.appledaily.com.tw/realtimenews/arti... | 55歲的庾澄慶（哈林）7月底證實和交往10個月的民視主播張嘉欣公證再婚，之後張嘉欣請長假出國... | 【狗仔偷拍】哈林主播妻收服小哈利 新婚甜牽回愛巢 |
| 13 | 社會 | 5161 | 2016-09-23 00:28:00 | http://www.appledaily.com.tw/realtimenews/arti... | 輔大風波持續延燒！輔仁大學社科院長夏林清昨在臉書發文後，引發各界熱議，輔大昨學召開緊急會議，... | 輔大議題延燒 社科院長夏林清暫時停職 |
| 51 | 副刊 | | 2016- | m.tw/realtimenews/arti... | 日前一名大陸女明星，因為入住甲醛超標的新房子，進而引發基因病變，最後因罹患淋巴癌而死亡，讓大... | 這東西有夠毒 食衣住行竟然都有 |

由大到小做排序

依據各分類排行

```
df1 = newsdf.groupby(['category'], sort=False)['clicked'].max()  
df1
```

```
category  
體育      54  
生活     1921  
娛樂     8054  
國際     1205  
政治       792  
動物     2710  
副刊     3049  
社會     5161  
3C       1364  
論壇     1585  
財經     1614  
地產       332  
Name: clicked, dtype: int64
```

根據新聞類別抓出
各分類人氣最旺的指標

抓出各分類最熱門文章

```
idx = newsdf.groupby(['category'])['clicked'].transform(max) == newspd['clicked']
newspd[idx].head()
```

| | category | clicked | dt | link | summary | title |
|----|----------|---------|---------------------|---|---|---------------------|
| 0 | 體育 | 54 | 2016-09-23 01:02:00 | http://www.appledaily.com.tw/realtimenews/arti... | 2017經典賽資格賽最後一組，台灣時間今天凌晨一點在紐約布魯克林區康尼島的MCU球場開打，由... | 經典賽資格賽韓荷情蒐就位 台灣靠鍵盤啦 |
| 13 | 社會 | 5161 | 2016-09-23 00:28:00 | http://www.appledaily.com.tw/realtimenews/arti... | 輔大風波持續延燒！輔仁大學社科院長夏林清昨在臉書發文後，引發各界熱議，輔大昨學召開緊急會議，... | 輔大議題延燒 社科院長夏林清暫時停職 |
| 17 | 3C | 1364 | 2016-09-23 00:23:00 | http://www.appledaily.com.tw/realtimenews/arti... | 每次要對發票的時候都要一張一張對，實在是很麻煩，但現在其實有很多方便的APP可以幫助你對發 | 對發票好麻煩 下載這個就對了！ |

文章的人氣數必須等於最大人氣數

使用Pandas 匯出資料

將資料存進Excel

```
newsdf.to_excel('appledaily.xlsx')
```

| | category | popularity | time | title | url |
|---|----------|------------|-------|-------------------------|---|
| 0 | FUN | 0 | 00:28 | 神設計餐盤 讓料理好有戲 | http://www.appledaily.com.tw/realtimenews/arti... |
| 1 | 娛樂 | 0 | 00:28 | 梁詠琪噴嚏嚇哭女兒 愛犬舔baby腳底呼... | http://www.appledaily.com.tw/realtimenews/arti... |
| 2 | 動物 | 0 | 00:28 | 好不容易獲救 隔天後卻馬上... | http://www.appledaily.com.tw/realtimenews/arti... |
| 3 | 生活 | 0 | 00:27 | 美妝即期品開賣 每人只能買2件 | http://www.appledaily.com.tw/realtimenews/arti... |
| 4 | 娛樂 | 90 | 00:25 | 準爸爸池城 今搶百想戲王 | http://www.appledaily.com.tw/realtimenews/arti... |

The background features a light blue and white hexagonal grid pattern. Overlaid on this is a large, faint, circular graphic composed of concentric rings and radial lines, resembling a stylized sun or a target. The text "THANK YOU" is centered in a bold, dark blue, sans-serif font.

THANK YOU