

Python金融爬蟲原理班

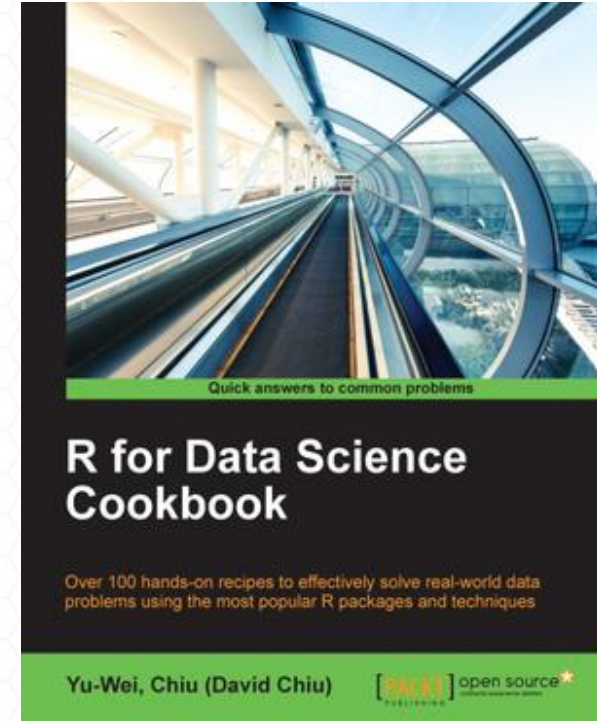
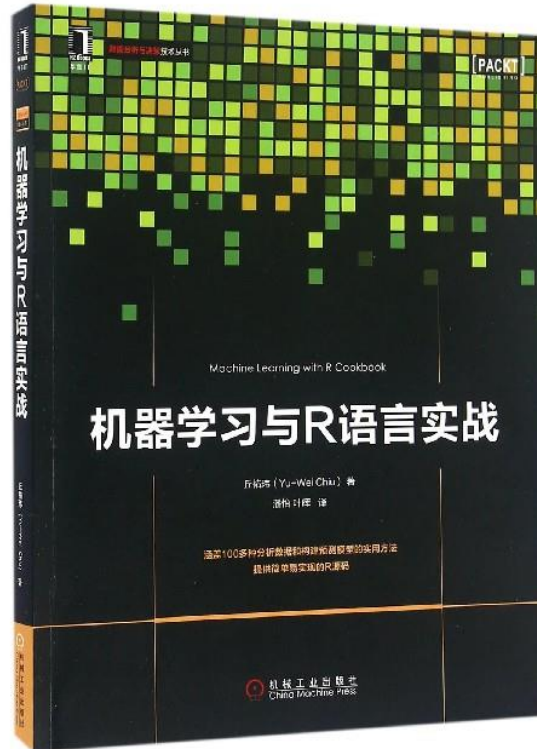
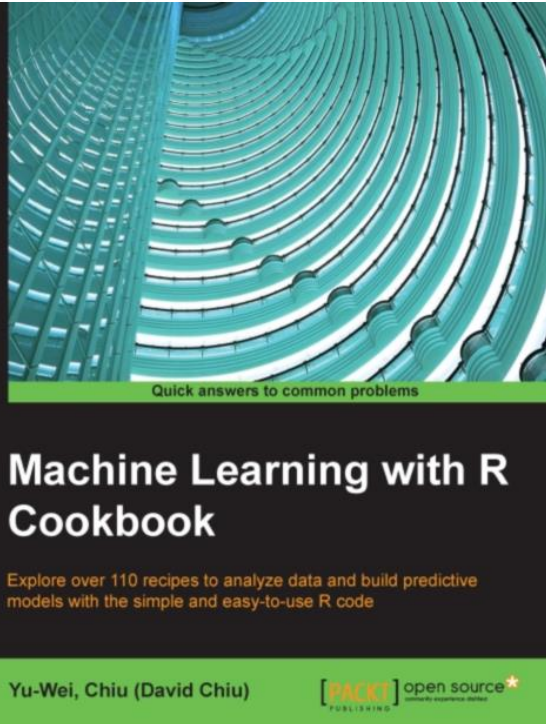
David Chiu
2017/04/16

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



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/pyfinance>

網路爬蟲

透過分析數據擬定策略才能找到聖杯

- 在做任何分析之前，必定要蒐集足夠的數據做分析，才能擬定高勝率策略
- 除了購買數據外，任何人都可以透過撰寫ETL (Extract, Transformation, Loading) 程序自動化蒐集資訊



將非結構化數據轉變為結構化數據

公開資訊觀測站

登錄 | 資訊項目 | 精華版2.0 | 重大訊息 | English

請輸入公司代號或簡稱 搜尋 代號查詢 回首頁

基本資料 負債報表 股東會及股利 公司治理 財務報表 重大訊息與公告 營運概況 投資專區 認購(售)權證 債券 資產證券化

負債報表

- 基本資料
- 股東會及股利
 - TDR股利分派情形(101年起適用)
 - 除權公告
 - 股東會及除權息日曆
 - 法人說明會一覽表
- 財務報表
 - 按IFRSs後
 - 綜合損益表
 - 資產負債表
 - 財務報告經監察人承認情形
 - 會計師查核(核閱)報告
 - 各產業EPS統計資訊
 - 按IFRSs前
 - 財務預測

綜合損益表

市場別 上市 年度 104 季別 3 搜尋

列印網頁 重新讀取 問題回報 目上頁

上市公司第三季資料

註：依證券交易法第36條及證券商期貨局相關函令規定，財務報告申報期限如下：

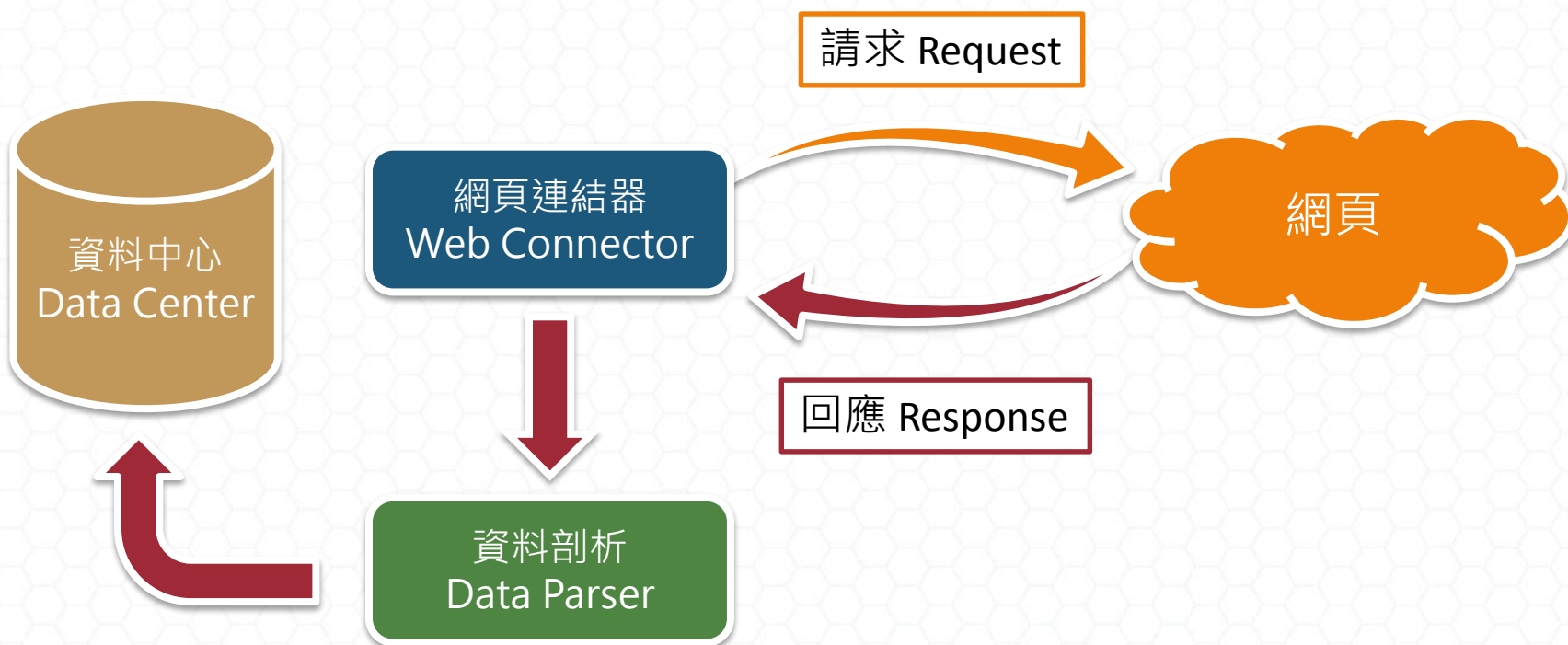
- 1.一般行業申報期限：第一季為5月15日，第二季為8月14日，第三季為11月14日，年度為3月31日。
- 2.金融業申報期限：第一季為5月30日，第二季為8月31日，第三季為11月29日，年度為3月31日。
- 3.銀行及票業申報期限：第一季為5月15日，第二季為8月31日，第三季為11月14日，年度為3月31日。
- 4.保險業申報期限：第一季為5月15日，第二季為8月31日，第三季為11月14日，年度為3月31日。
- 5.證券業申報期限：第一季為5月15日，第二季為8月31日，第三季為11月14日，年度為3月31日。
- 6.申報期限如遇例假日，以證券商期貨局公布者為準。

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



使用Anaconda 開發

安裝Anaconda

Download for Windows

Download for macOS

Download for Linux

Anaconda 4.3.1

For Windows

Anaconda is BSD licensed which gives you permission to use Anaconda commercially and for redistribution.

[Changelog](#)

1. Download the installer
2. Optional: Verify data integrity with [MD5](#) or [SHA-256](#) [More info](#)
3. Double-click the **.exe** file to install Anaconda and follow the instructions on the screen

Behind a firewall? Use these [zipped Windows installers](#)

Python 3.6 version

64-BIT INSTALLER (422M)

32-BIT INSTALLER (348M)

Python 2.7 version

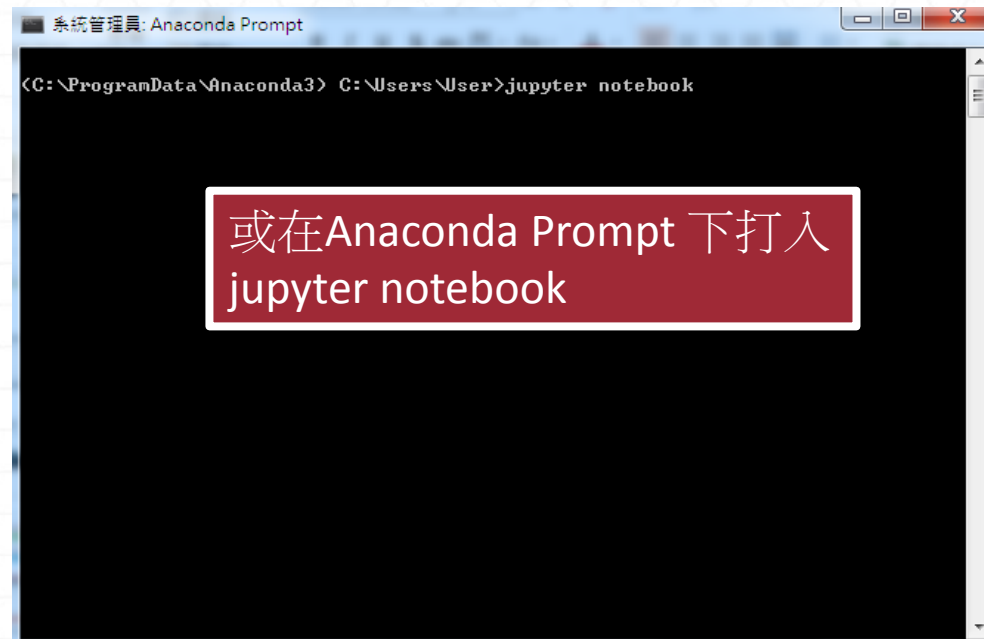
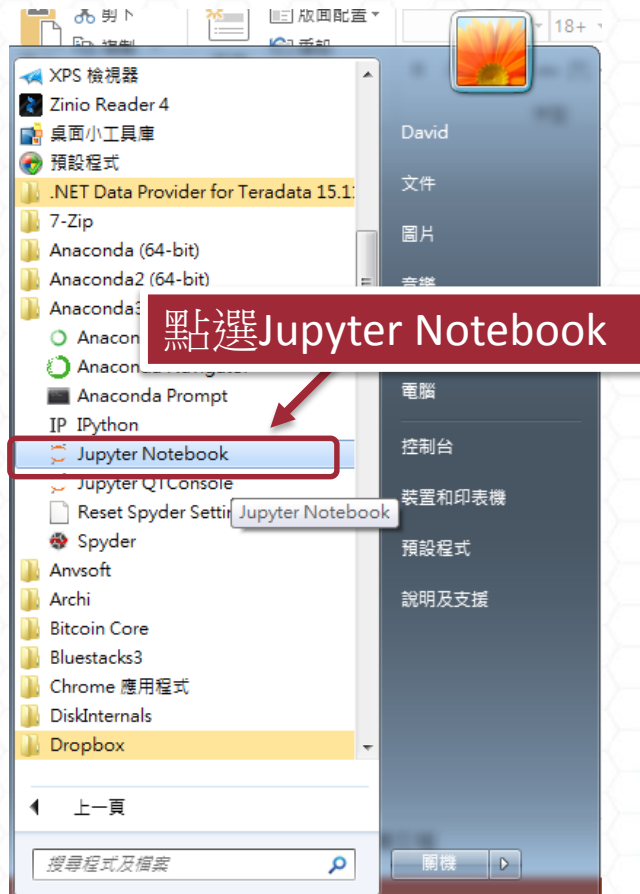
64-BIT INSTALLER (414M)

32-BIT INSTALLER (339M)

選擇Python 3.6版

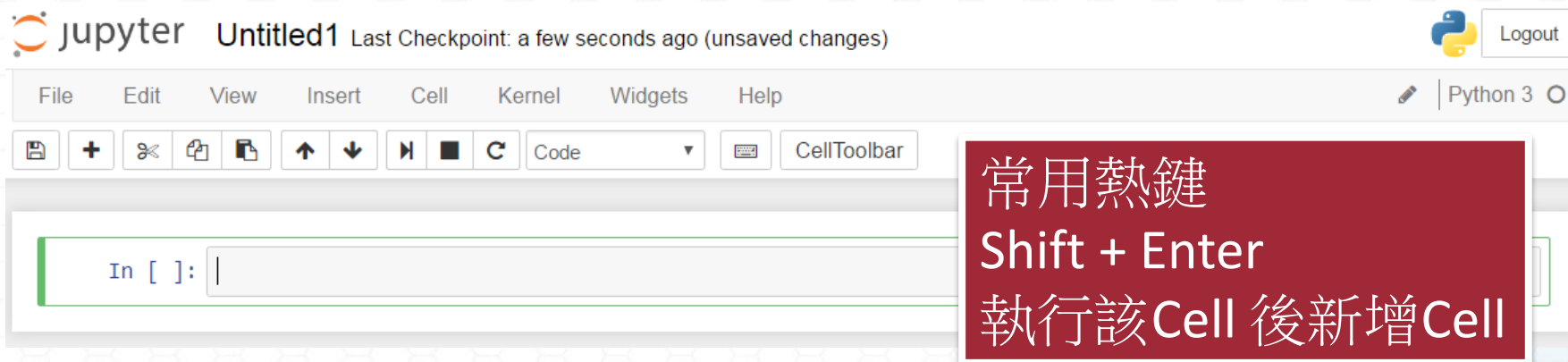
<https://www.continuum.io/downloads>

使用 Jupyter (Ipython Notebook)



啟用 Jupyter (jupyter Notebook)

- 在命令列下打:
 - jupyter notebook
 - 自動開啟瀏覽器後便可瀏覽 (預設為localhost:8888)
- 可匯出.ipynb, .py 各種不同格式檔案
- 瀏覽快捷鍵 Help -> Keyboard Shortcuts



Python v.s. Java

Java

```
/* example 1 */  
public static void main(String[]  
args){  
    for(int i=0; i< 10; i++)  
        System.out.print(i);  
}
```

- 執行速度較Python 為快
- 使用{}分隔區塊
- 需要宣告變數型態
- 可以透過Compiler 檢查錯誤
- 使用/**/做註解

Python

```
"example1"  
for i in range(1,11):  
    print(i)
```

- 開發速度較快
- 使用indent 替代 {}
- 不須宣告變數型態
- 只能在runtime 檢查錯誤
- 以#與"或" 做註解

撰寫第一隻爬蟲

使用開發人員工具

■ 於網頁上點選右鍵 -> 檢查

臺灣證券交易所

線上支援 相關服務平台 全站搜尋

關於證交所 公司治理中心 交易資訊 上市公司 產品與服務 結算服務 市場公告

盤後資訊
臺灣跨市場指數
TWSE自行編製指數
與FTSE合作編製指數
與銳聯合作編製指數
與S&PDJI合作編製指數
升降幅度/首五日無漲跌幅
變更交易
當日沖銷交易標的及統計
融資融券與可借券賣出額度
標價
三大法人
三大法人買賣金額統計表

資料日期：106/04/13

列印 下載HTML 下載CSV

106年04月13日

| 證券代號 | 證券名稱 | 外資買進股數 | 外資賣出股數 | 外資買賣超股數 |
|------|------|-----------|---------|---------|
| 1102 | 亞泥 | 1,571,000 | 663,000 | 908,000 |
| 1104 | 環泥 | 3,000 | 0 | 3,000 |
| 1103 | 嘉泥 | 31,000 | 33,000 | -2,000 |
| 1109 | 信大 | 0 | 0 | 0 |
| 1108 | 幸福 | 2,000 | 60,000 | -58,000 |

檢查(N) Ctrl+Shift+I

點選檢查
或使用ctrl + shift + i

觀察HTTP 請求與返回內容

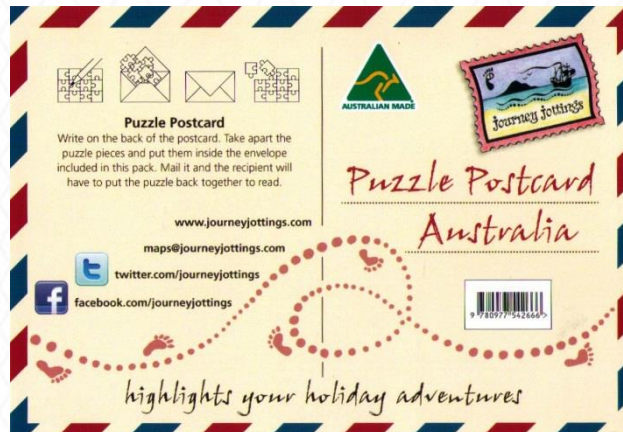
The screenshot shows the Chrome DevTools Network tab. The 'Network' tab is selected in the top bar. The 'Doc' sub-tab is selected in the request details pane. The request 'T86.php' is selected in the list. The details pane shows the following information:

- General**
 - Request URL: <http://www.twse.com.tw/ch/trading/fund/T86/T86.php>
 - Request Method: GET
 - Status Code: 200 OK
 - Remote Address: 163.29.17.130:80
 - Referrer Policy: no-referrer-when-downgrade
- Response Headers**
 - Connection: Keep-Alive
 - Content-Type: text/html; charset=utf-8
 - Date: Fri, 14 Apr 2017 05:17:20 GMT
 - Keep-Alive: timeout=5, max=100
 - Server: Apache/2.2.9 (Unix) DAV/2
 - Transfer-Encoding: chunked
- Request Headers**

Annotations with red boxes and arrows indicate the steps:

1. 點選Network (Click Network)
2. 點選Doc (Click Doc)
3. 點選連結 (Click link) - pointing to the 'T86.php' entry in the list.

什麼是GET?



GET
內容寫在上頭

<http://www.twse.com.tw/ch/trading/fund/T86/T86.php>

Python 抓取網頁的主流套件

■ Requests

- ▣ 改善Urllib2 的缺點，讓使用者以最簡單的方式獲取網路資源
- ▣ 使用**REST** 操作，可以調用GET,POST, PUT, DELETE

使用GET 抓取頁面資訊

```
import requests  
res = requests.get('http://www.twse.com.tw/ch/trading/fund/T86/T86.php')  
res.text
```

```
▼ <table border="1" align="center" style="width:1400px;" id="tbl-sortable" class="sortable">  
  ▼ <thead>  
    ▼ <tr>  
      ▼ <th>  
        "證券"  
        <br>  
        "代號"  
      </th>  
      ▶ <th>...</th>  
      ▶ <th>...</th>  
      ▶ <th>...</th>
```

使用Help 與 dir 查詢套件與函式

```
import requests
```

```
help(requests)
```

使用help 查詢文件

```
dir(requests)
```

使用dir 表列可用屬性
與方法

```
help(requests.get)
```

```
? requests.get
```

不確定該方法的功能
使用help 或 ?

抓取三大法人買賣超日報


臺灣證券交易所

[ENGLISH](#)
[日本語](#)
[t](#)
[f](#)
[p](#)

[線上支援](#)
[相關服務平台](#)

[關於證交所](#)
[公司治理中心](#)
[交易資訊](#)
[上市公司](#)
[產品與服務](#)
[結算服務](#)
[市場公告](#)
[法令規章](#)

[盤後資訊](#)
[臺灣跨市場指數](#)
[TWSE自行編製指數](#)
[與FTSE合作編製指數](#)
[與銳聯合作編製指數](#)
[與S&PDJI合作編製指數](#)
[升降幅度/首五日無漲跌幅](#)
[變更交易](#)
[當日沖銷交易標的及統計](#)
[融資融券與可融券賣出額度](#)
[標借](#)
[三大法人](#)
[三大法人買賣金額統計表](#)
[三大法人買賣超日](#)

[首頁](#) > [交易資訊](#) > [三大法人](#) > [三大法人買賣超日報](#)
[回首頁](#)



資料日期：
分類項目：

☒ 依買賣超股數排列
 ☐ 依證券代號排列

本資訊自民國101年5月2日起提供

| 證券代號 | 證券名稱 | 外資買進股數 | 外資賣出股數 | 外資買賣超股數 | 投信買進股數 | 投信賣出股數 | 投信買賣超股數 | 自營商買進股數 | 自營商賣出股數 (自行買賣) | 自營商買賣超股數 (自行買賣) |
|------|------|-----------|-----------|---------|--------|--------|---------|---------|----------------|-----------------|
| 1102 | 亞泥 | 4,774,896 | 4,036,737 | 738,159 | 0 | 0 | 0 | 185,000 | 0 | 1,00 |
| 1109 | 信大 | 6,000 | 10,000 | -4,000 | 0 | 0 | 0 | 0 | 0 | |
| 1110 | 東泥 | 0 | 19,000 | -19,000 | 0 | 0 | 0 | 0 | 0 | |
| 1104 | 環泥 | 1,000 | 85,000 | -84,000 | 70,000 | 6,000 | 64,000 | 0 | 0 | |
| 1108 | 幸福 | 1,000 | 46,000 | -45,000 | 0 | 0 | 0 | 23,000 | 23,000 | |

<http://www.twse.com.tw/ch/trading/fund/T86/T86.php>

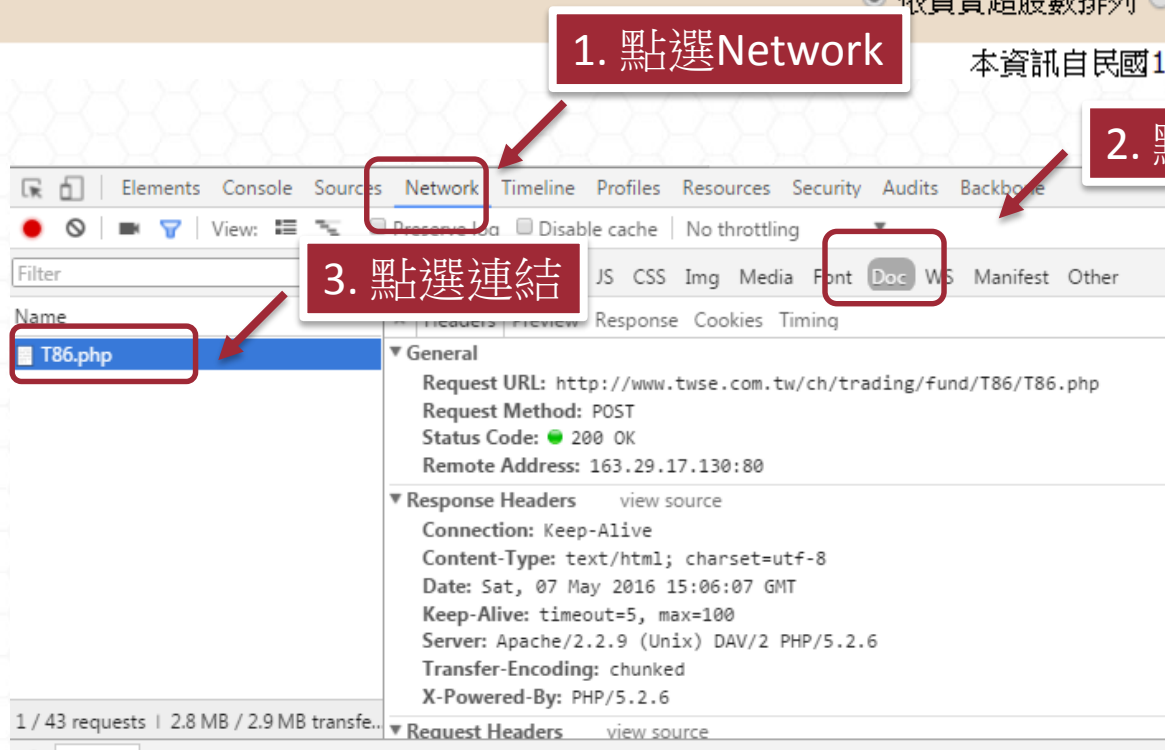
找尋抓取三大法人買賣超日報資訊

■ 填入資訊後按查詢

資料日期： 分類項目：

☒ 依買賣超股數排列 ☐ 依證券代號排列

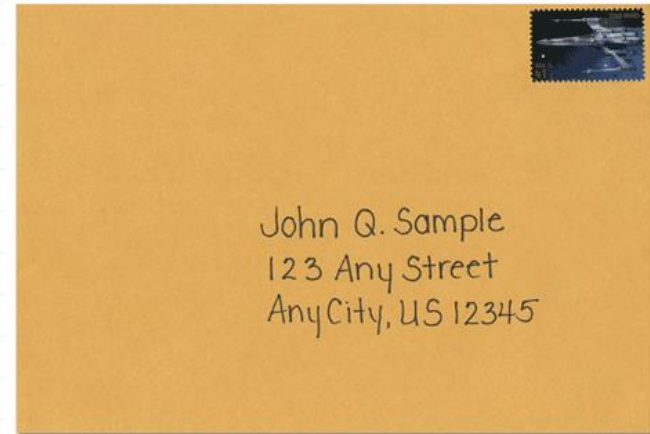
本資訊自民國101年5月2日起提供



什麼是POST?

download:
qdate: 106/04/12
select2: ALL
sorting: by_issue

<http://www.twse.com.tw/ch/trading/fund/T86/T86.php>



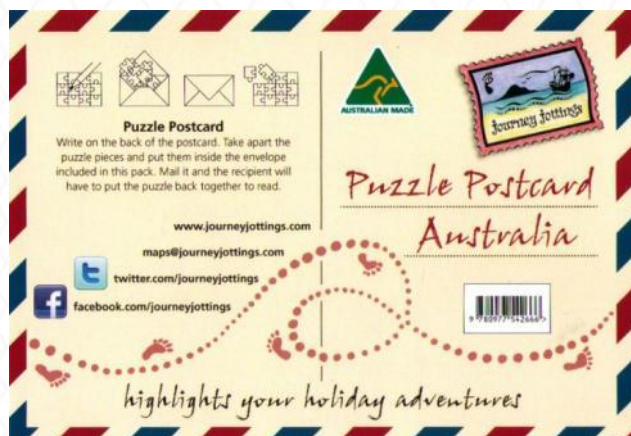
POST
內容寫在信紙，包在信封內

使用POST 取得三大法人買賣超日報資訊

```
import requests  
payload = {  
    'qdate':'106/04/12',  
    'select2':'ALL',  
    'sorting':'by_issue'  
}
```

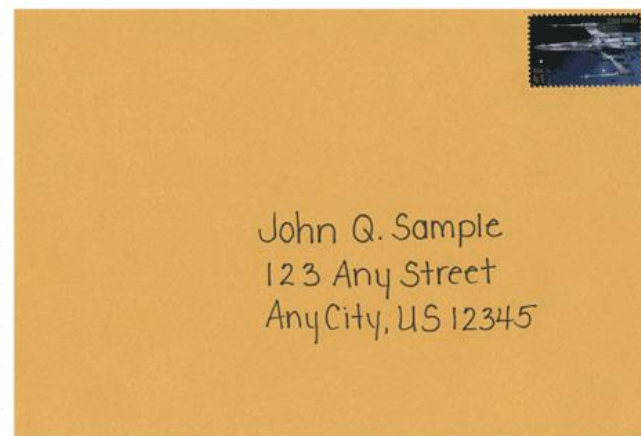
```
res = requests.post('http://www.twse.com.tw/ch/trading/fund/T86/T86.php',  
    data=payload)
```

GET V.S. POST



GET

內容寫在上頭



POST

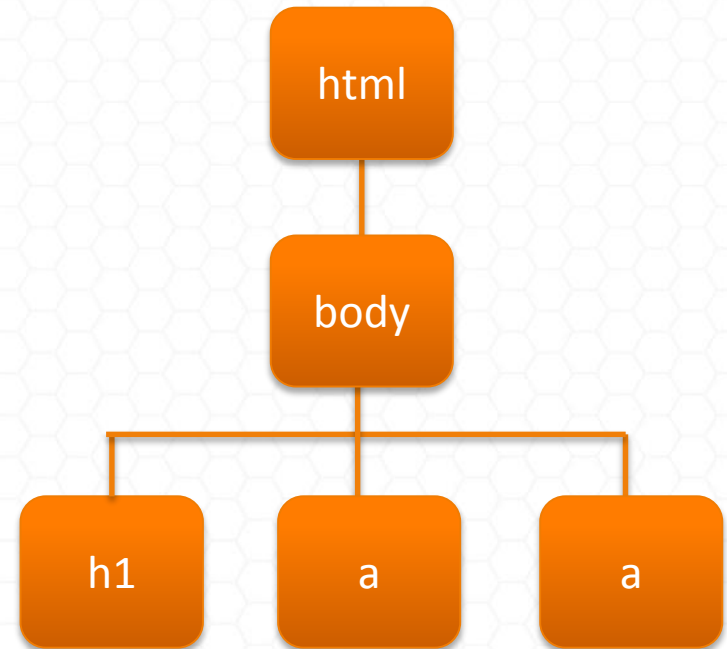
內容寫在信紙，包在信封內

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, 'html.parser')



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, 'html.parser')
```

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


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, 'html.parser')  
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, 'html.parser')
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'])
```

試著抓取表格資料

2. 點選要抓取的區塊

- 變更交易
- 當日沖銷交易標的及統計
- 融資融券與可借券賣出額度
- 標債
- 三大法人
 - 三大法人買賣金額統計表
 - 三大法人買賣超日報

100407月14日 三大法人買賣超日報

| 證券代號 | 證券名稱 | 外資買進股數 | 外資賣出股數 | 外資買賣超股數 | 投信買進股數 | 投信賣出股數 | 投信買賣超股數 | 自營商買進股數 | 自營商買賣超股數(自行買賣) |
|------|------|---------|---------|----------|--------|--------|---------|---------|----------------|
| 1110 | 東泥 | 1,000 | 0 | 1,000 | 0 | 0 | 0 | 0 | |
| 1104 | 環泥 | 10,000 | 44,000 | -34,000 | 0 | 2,000 | -2,000 | 0 | |
| 1108 | 幸福 | 2,000 | 42,000 | -40,000 | 0 | 2,000 | -2,000 | 0 | |
| 1103 | 嘉泥 | 19,000 | 119,000 | -100,000 | 0 | 13,000 | -13,000 | 0 | |
| 1101 | 台泥 | 463,000 | 442,000 | 21,000 | 0 | 58,000 | -58,000 | -86,000 | |
| 1102 | 亞泥 | 291,000 | 423,000 | -132,000 | 0 | 7,000 | -7,000 | 0 | |

1. 點選觀察元素

說明：

1. 自營商表示證券自營商專戶。
2. 投信表示本國投資信託基金。
3. 外資及陸資表示依「華僑及外國人投資證券管理辦法」及「大陸地區投資人來臺從事證券投資及期貨交易管理辦法」辦理登記之投資人。

3. 檢視css path

```
Elements Profiles Console Sources Network Timeline Application Backbone Security Adblock Audits
</thead>
</table>
<table border="1" align="center" style="width:1400px;" id="tbl-sortable" class="sortable"> == $0
  <thead>
    <tr>
      <th>...</th>
      <th>...</th>
      <th>...</th>
      <th>...</th>
      <th>...</th>
      <th>...</th>
    </tr>
  </thead>
  <tbody>
    <tr>
      <td>1110</td>
      <td>東泥</td>
      <td>1,000</td>
      <td>0</td>
      <td>1,000</td>
      <td>0</td>
      <td>0</td>
      <td>0</td>
      <td>0</td>
    </tr>
    <tr>
      <td>1104</td>
      <td>環泥</td>
      <td>10,000</td>
      <td>44,000</td>
      <td>-34,000</td>
      <td>0</td>
      <td>2,000</td>
      <td>-2,000</td>
      <td>0</td>
    </tr>
    <tr>
      <td>1108</td>
      <td>幸福</td>
      <td>2,000</td>
      <td>42,000</td>
      <td>-40,000</td>
      <td>0</td>
      <td>2,000</td>
      <td>-2,000</td>
      <td>0</td>
    </tr>
    <tr>
      <td>1103</td>
      <td>嘉泥</td>
      <td>19,000</td>
      <td>119,000</td>
      <td>-100,000</td>
      <td>0</td>
      <td>13,000</td>
      <td>-13,000</td>
      <td>0</td>
    </tr>
    <tr>
      <td>1101</td>
      <td>台泥</td>
      <td>463,000</td>
      <td>442,000</td>
      <td>21,000</td>
      <td>0</td>
      <td>58,000</td>
      <td>-58,000</td>
      <td>-86,000</td>
    </tr>
    <tr>
      <td>1102</td>
      <td>亞泥</td>
      <td>291,000</td>
      <td>423,000</td>
      <td>-132,000</td>
      <td>0</td>
      <td>7,000</td>
      <td>-7,000</td>
      <td>0</td>
    </tr>
  </tbody>
</table>
```

html body:twse div#body div.page-width div#main-content table#tbl-sortable.sortable thead tr th

抓取交易表格

```
import requests
from bs4 import BeautifulSoup
res = requests.get('http://www.twse.com.tw/ch/trading/fund/T86/T86.php')
soup = BeautifulSoup(res.text, 'html.parser')
soup.select('#tbl-sortable')
```

| 106年04月14日 三大法人買賣超日報 | | | | | | | | | |
|----------------------|------|---------|---------|----------|--------|--------|---------|----------|-------------------|
| 證券代號 | 證券名稱 | 外資買進股數 | 外資賣出股數 | 外資買賣超股數 | 投信買進股數 | 投信賣出股數 | 投信買賣超股數 | 自營商買賣超股數 | 自營商買進股數 (自行買賣) |
| 1110 | 東泥 | 1,000 | 0 | 1,000 | 0 | 0 | 0 | 0 | 0 |
| 1104 | 環泥 | 10,000 | 44,000 | -34,000 | 0 | 2,000 | -2,000 | 0 | 0 |
| 1108 | 幸福 | 2,000 | 42,000 | -40,000 | 0 | 2,000 | -2,000 | 0 | 0 |
| 1103 | 嘉泥 | 19,000 | 119,000 | -100,000 | 0 | 13,000 | -13,000 | 0 | 0 |
| 1101 | 台泥 | 463,000 | 442,000 | 21,000 | 0 | 58,000 | -58,000 | -86,000 | 0 |
| 1102 | 亞泥 | 291,000 | 423,000 | -132,000 | 0 | 7,000 | -7,000 | 0 | 0 |

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/ipjbadabbpedegielkhgpiekdImfpgal>

使用InfoLite 點選抓取區域

列印 下載HTML 下載CSV

| 106年04月14日 三大法人買賣超 | | | | | | | | | | |
|--------------------|------|---------|---------|----------|--------|--------|---------|---------|--------|---------|
| 證券代號 | 證券名稱 | 外資買進股數 | 外資賣出股數 | 外資買賣超股數 | 投信買進股數 | 投信賣出股數 | 投信買賣超股數 | 自營買賣超股數 | (自行買賣) | (自行買賣) |
| 1110 | 東泥 | 1,000 | 0 | 1,000 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1104 | 環泥 | 10,000 | 44,000 | -34,000 | 0 | 2,000 | -2,000 | 0 | 0 | 0 |
| 1108 | 幸福 | 2,000 | 42,000 | -40,000 | 0 | 2,000 | -2,000 | 0 | 0 | 0 |
| 1103 | 嘉泥 | 19,000 | 119,000 | -100,000 | 0 | 13,000 | -13,000 | 0 | 0 | 0 |
| 1101 | 台泥 | 463,000 | 442,000 | 21,000 | 0 | 58,000 | -58,000 | -86,000 | 0 | 130,000 |
| 1102 | 亞泥 | 291,000 | 423,000 | -132,000 | 0 | 7,000 | -7,000 | 0 | 0 | 0 |

div table

InfoLite Username Password Login X ?
Submit
#tbl-sortable Clear (1) Add

綠色: 目前選取得區塊

黃色: 符合樣式的區塊

紅色: 排除的區塊

Clear旁邊的數字: 符合區塊的數目

使用Pandas處理表格資料

Pandas

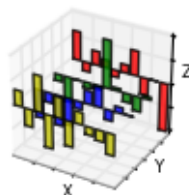
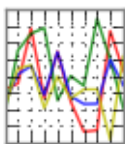
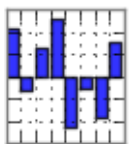
■ Python for Data Analysis

- 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 titled "三大法人五月多單紀錄.csv". The spreadsheet contains a table with 12 rows of data. The first column (A) lists categories like "自營商", "投信", and "外資". The next three columns (B, C, D) contain numerical values. The fourth column (E) contains dates. The rest of the columns (F-J) are empty.

| | A | B | C | D | E | F | G | H | I | J |
|----|-----|--------|-------|-----------|---|---|---|---|---|---|
| 1 | 自營商 | 250996 | 23505 | 2014/5/30 | | | | | | |
| 2 | 投信 | 61 | 81 | 2014/5/30 | | | | | | |
| 3 | 外資 | 63708 | 39940 | 2014/5/30 | | | | | | |
| 4 | 自營商 | 208735 | 22713 | 2014/5/29 | | | | | | |
| 5 | 投信 | 31 | 16 | 2014/5/29 | | | | | | |
| 6 | 外資 | 62902 | 33782 | 2014/5/29 | | | | | | |
| 7 | 自營商 | 361487 | 32194 | 2014/5/28 | | | | | | |
| 8 | 投信 | 349 | 247 | 2014/5/28 | | | | | | |
| 9 | 外資 | 83194 | 52150 | 2014/5/28 | | | | | | |
| 10 | 自營商 | 175380 | 15615 | 2014/5/27 | | | | | | |
| 11 | 投信 | 52 | 94 | 2014/5/27 | | | | | | |
| 12 | 外資 | 43136 | 33329 | 2014/5/27 | | | | | | |
| 13 | 自營商 | 228125 | 10556 | 2014/5/26 | | | | | | |

使用read_html 讀取表格

```
table = """
<table>
  <thead>
    <tr>
      <th>Month</th>
      <th>Savings</th>
    </tr>
  </thead>
  <tbody>
    <tr> <td>January</td> <td>$100</td></tr>
    <tr> <td>February</td> <td>$80</td></tr>
  </tbody>
  <tfoot>
    <tr> <td>Sum</td> <td>$180</td> </tr>
  </tfoot>
</table>
"""
```

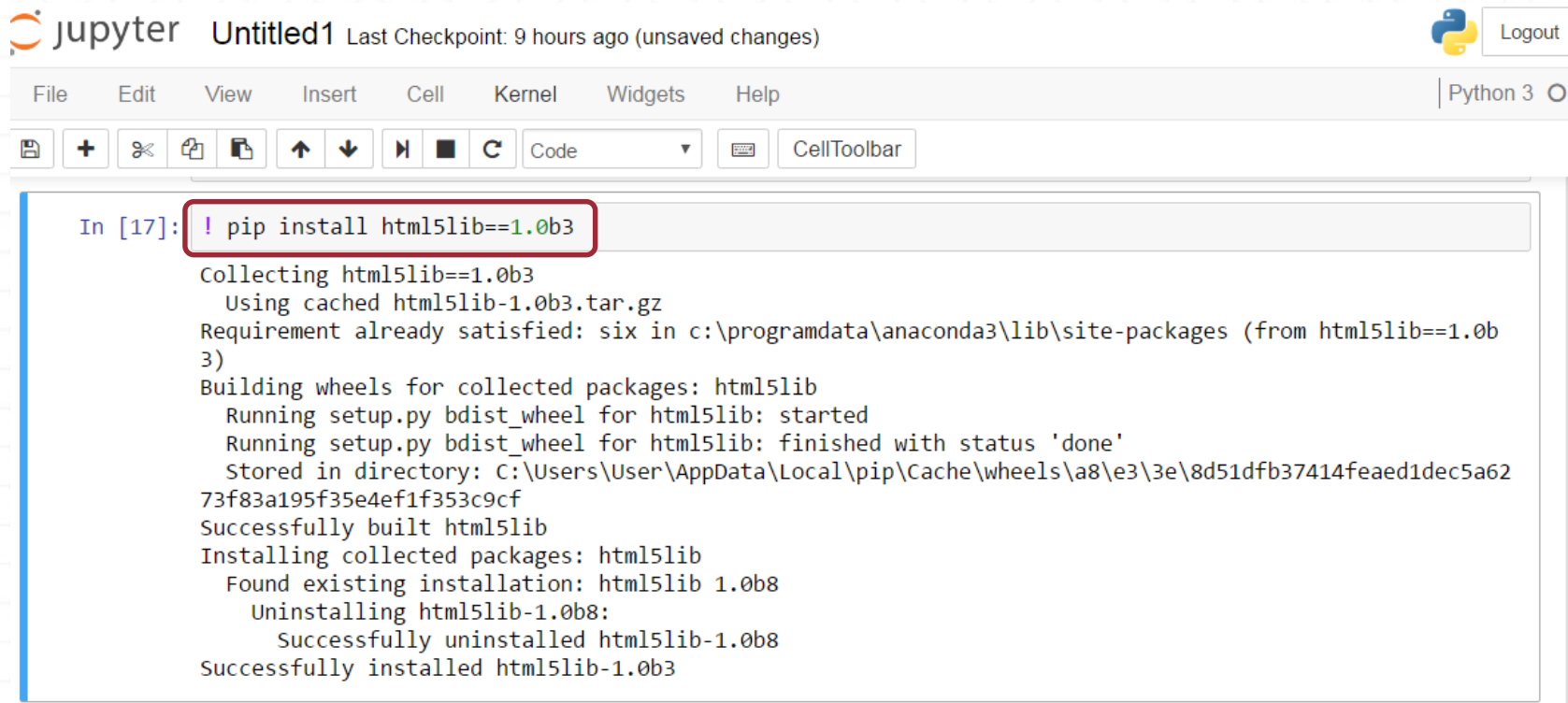
```
import pandas as pd
dfs = pd.read_html(table)
dfs[0]
```

| | Month | Savings |
|---|----------|---------|
| 0 | January | \$100 |
| 1 | February | \$80 |
| 2 | Sum | \$180 |

安裝html5lib

■ 在Jupyter Notebook 的 Cell 下打

□! pip install html5lib==1.0b3



The screenshot shows a Jupyter Notebook window titled 'Untitled1' with a 'Last Checkpoint: 9 hours ago (unsaved changes)' message. The interface includes a top menu bar with 'File', 'Edit', 'View', 'Insert', 'Cell', 'Kernel', 'Widgets', and 'Help'. Below the menu is a toolbar with icons for saving, adding cells, and other functions. The main area displays a code cell with the command `! pip install html5lib==1.0b3` highlighted by a red box. The output of the command is shown below the input, detailing the collection, building, and installation of the package.

```
In [17]: ! pip install html5lib==1.0b3
Collecting html5lib==1.0b3
  Using cached html5lib-1.0b3.tar.gz
Requirement already satisfied: six in c:\programdata\anaconda3\lib\site-packages (from html5lib==1.0b3)
Building wheels for collected packages: html5lib
  Running setup.py bdist_wheel for html5lib: started
  Running setup.py bdist_wheel for html5lib: finished with status 'done'
  Stored in directory: C:\Users\User\AppData\Local\pip\Cache\wheels\a8\e3\3e\8d51dfb37414feaed1dec5a6273f83a195f35e4ef1f353c9cf
Successfully built html5lib
Installing collected packages: html5lib
  Found existing installation: html5lib 1.0b8
  Uninstalling html5lib-1.0b8:
    Successfully uninstalled html5lib-1.0b8
Successfully installed html5lib-1.0b3
```

整理三大法人買賣超日報資訊

整理三大法人買賣超日報資訊


臺灣證券交易所

[ENGLISH](#)
[日本語](#)
[t](#)
[f](#)
[p](#)

[線上支援](#)
[相關服務平台](#)

[關於證交所](#)
[公司治理中心](#)
[交易資訊](#)
[上市公司](#)
[產品與服務](#)
[結算服務](#)
[市場公告](#)
[法令規章](#)

- 盤後資訊
- 臺灣跨市場指數
- TWSE自行編製指數
- 與FTSE合作編製指數
- 與銳聯合作編製指數
- 與S&PDJI合作編製指數
- 升降幅度/首五日無漲跌幅
- 變更交易
- 當日沖銷交易標的及統計
- 融資融券與可融券賣出額度
- 標借
- 三大法人
 - 三大法人買賣金額統計表
 - 三大法人買賣超口

[首頁](#) > [交易資訊](#) > [三大法人](#) > [三大法人買賣超日報](#)
[回首頁](#)



資料日期：
 分類項目：

☒ 依買賣超股數排列
 ☐ 依證券代號排列

本資訊自民國101年5月2日起提供

| 105年05月06日 三大法人買賣超日報 | | | | | | | | | | |
|----------------------|------|-----------|-----------|---------|--------|--------|---------|----------|---------------|---------------|
| 證券代號 | 證券名稱 | 外資買進股數 | 外資賣出股數 | 外資買賣超股數 | 投信買進股數 | 投信賣出股數 | 投信買賣超股數 | 自營商買賣超股數 | 自營商買進股數(自行買賣) | 自營商賣出股數(自行買賣) |
| 1102 | 亞泥 | 4,774,896 | 4,036,737 | 738,159 | 0 | 0 | 0 | 185,000 | 0 | 1,00 |
| 1109 | 信大 | 6,000 | 10,000 | -4,000 | 0 | 0 | 0 | 0 | 0 | |
| 1110 | 東泥 | 0 | 19,000 | -19,000 | 0 | 0 | 0 | 0 | 0 | |
| 1104 | 環泥 | 1,000 | 85,000 | -84,000 | 70,000 | 6,000 | 64,000 | 0 | 0 | |
| 1108 | 幸福 | 1,000 | 46,000 | -45,000 | 0 | 0 | 0 | 23,000 | 23,000 | |

抓取三大法人買賣超日報資訊

```
import requests  
payload = {  
    'qdate':'106/04/14',  
    'select2':'24',  
    'sorting':'by_issue'  
}
```

```
res = requests.post('http://www.twse.com.tw/ch/trading/fund/T86/T86.php', data=payload)
```


使用Pandas 讀取資料

```
from bs4 import BeautifulSoup
import pandas
soup = BeautifulSoup(res.text, 'lxml')
tbl = soup.select('#tbl-sortable')[0]
dfs = pandas.read_html(tbl.prettify('utf-8'), encoding='utf-8')
stockdf = dfs[0]
```

Out[30]:

| | 證券代號 | 證券名稱 | 外資買進股數 | 外資賣出股數 | 外資買賣超股數 | 投信買進股數 | 投信賣出股數 | 投信買賣超股數 | 自營商買超股數 | 自營商買進股數(自行買賣) | 自營商賣出股數(自行買賣) | 自營商買賣行單 |
|---|--------|-----------|---------|--------|---------|---------|---------|----------|----------|---------------|---------------|---------|
| 0 | 00632R | T50 反1 | 350000 | 391000 | -41000 | 0 | 5243000 | -5243000 | 65286000 | 1195000 | 1030000 | 165 |
| 1 | 042800 | 永豐 EX | 0 | 45000 | -45000 | 0 | 0 | 0 | 9533000 | 0 | 0 | 0 |
| 2 | 2349 | 鍊德 | 9648000 | 997000 | 8651000 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 2345 | 智邦 | 2484000 | 804000 | 1680000 | 1854000 | 0 | 1854000 | 509000 | 472000 | 66000 | 406 |

猜猜哪隻股票外資買賣超最多？

```
stockdf[stockdf['外資 買賣超股數'].decode('utf-8')]==stockdf['外資 買賣超股數'].decode('utf-8')].max()]
```

| | 證券代號 | 證券名稱 | 外資買進股數 | 外資賣出股數 | 外資買賣超股數 | 投信買進股數 | 投信賣出股數 | 投信買賣超股數 | 自營商買賣超股數 | 自營商買進股數(自行買賣) | 自營商賣出股數(自行買賣) | 自營商買賣超股數(自行買賣) | 自營商買進股數(避險) | 自營商賣出股數(避險) | 自營商買賣超股數(避險) | 三大法人買賣超股數 |
|---|------|------|---------|--------|---------|--------|--------|---------|----------|---------------|---------------|----------------|-------------|-------------|--------------|-----------|
| 2 | 2349 | 鍊德 | 9648000 | 997000 | 8651000 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 8651000 |

根據買賣超排序

```
stockdf.sort_values( by= '外資 買賣超股數'  
' .decode('utf-8'), ascending=False).head()
```

| | 證券代號 | 證券名稱 | 外資買進股數 | 外資賣出股數 | 外資買賣超股數 | 投信買進股數 | 投信賣出股數 | 投信買賣超股數 | 自營商買賣超股數 | 自營商買進股數(自行買賣) | 自營商賣出股數(自行買賣) | 自營商買賣超股數(自行買賣) | 自營商買進股數(避險) | 自營商賣出股數(避險) |
|----|------|------|---------|---------|---------|---------|---------|----------|----------|---------------|---------------|----------------|-------------|-------------|
| 2 | 2349 | 鍊德 | 9648000 | 997000 | 8651000 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | 3023 | 信邦 | 4756000 | 1194000 | 3562000 | 214000 | 0 | 214000 | -164000 | 118000 | 348000 | -230000 | 340000 | 274 |
| 18 | 1312 | 國喬 | 3830000 | 713000 | 3117000 | 0 | 1326000 | -1326000 | 198000 | 75000 | 0 | 75000 | 249000 | 126 |
| 5 | 1605 | 華新 | 3451000 | 1257000 | 2194000 | 1500000 | 0 | 1500000 | -61000 | 1000 | 100000 | -99000 | 210000 | 172 |

由大到小做排序

定義函式

```
def getTradingVolume(dt):  
    payload['qdate'] = getTWDate(dt)  
    res = requests.post('http://www.twse.com.tw/ch/trading/fund/T86/T86.php', data=payload)  
    soup = BeautifulSoup(res.text, 'lxml')  
    tbl = soup.select('#tbl-sortable')[0]  
    dfs = pd.read_html(tbl.prettify('utf-8'), encoding='utf-8')  
    stockdf = dfs[0]  
    stockdf['ymd'] = dt  
    return stockdf
```

增加日期欄位

時間跟字串轉換

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

```
a = '2017-04-16 14:00'  
print datetime.strptime(a, "%Y-%m-%d %H:%M")
```

產生日期

```
from datetime import date, datetime, timedelta
currenttime = datetime.now()
for i in range(1,3):
    dt = currenttime - timedelta(days = i)
    print(dt)
    print(dt.strftime('%Y/%m/%d'))
```

但是必須要民國時間

產生民國日期

```
from datetime import date,datetime, timedelta
currenttime = datetime.now()
for i in range(1,3):
    dt = currenttime - timedelta(days = i)
    year = int(dt.strftime('%Y')) - 1911
    monthdate = dt.strftime('%m/%d')
    print('{} / {}'.format(year, monthdate))
```

增加日期轉換函式

```
def getTWDate(dt):  
    year = int(dt.strftime('%Y')) - 1911  
    monthdate = dt.strftime('%m/%d')  
    ymd = '{}/{}'.format(year, monthdate)  
    return ymd
```


修改原本函式

```
def getTradingVolume(dt):  
    payload['qdate'] = getTWDate(dt)  
    res = requests.post('http://www.twse.com.tw/ch/trading/fund/T86/T86.php',  
data=payload)  
    soup = bs(res.text, 'html5lib')  
    tbl = soup.select('#tbl-sortable')[0]  
    dfs = pandas.read_html(tbl.prettify('utf-8'), encoding='utf-8')  
    stockdf = dfs[0]  
    stockdf['ymd'] = dt  
    return stockdf
```

批次執行30天的資料

```
dfs = []  
currenttime = datetime.now()  
for i in range(1,30):  
    dt = currenttime.date() - timedelta(days = i)  
    dfs.append(getTradingVolume(dt))
```

合併所有的Data Frame

```
stockdf = pandas.concat(dfs, ignore_index=True)  
len(stockdf)
```

| | 證券代號 | 證券名稱 | 外資買進股數 | 外資賣出股數 | 外資買賣超股數 | 投信買進股數 | 投信賣出股數 | 投信買賣超股數 | 自營商買賣超股數 | 自營商買進股數(自行買賣) | 自營商賣出股數(自行買賣) | 自營商買賣超股數(自行買賣) | 自營商買進股數(避險) | 自營商賣出股數(避險) | 自買股數(避險) |
|---|------|------|---------|---------|---------|--------|--------|---------|----------|---------------|---------------|----------------|-------------|-------------|----------|
| 0 | 3474 | 華亞科 | 5865836 | 3726000 | 2139836 | 0 | 0 | 0 | -4000 | 0 | 0 | 0 | 0 | 4000 | -4 |
| 1 | 2408 | 南亞科 | 2755000 | 1194000 | 1561000 | 0 | 0 | 0 | 104000 | 1000 | 0 | 1000 | 106000 | 3000 | 10 |
| 2 | 2449 | 京元 | 6490000 | 5054000 | 1436000 | 144000 | 443000 | -299000 | 131000 | 42000 | 77000 | -35000 | 509000 | 343000 | 16 |

將資料儲存至Excel

stockdf.to_excel('stock.xlsx')

stock.xlsx - Excel

檔案常用插入版面配置公式資料校閱檢視小組

新經明體11

B I U

資料儲存 (SQLITE)

課前預備

■ 安裝SQLite

- 請至官網<http://sqlite.org/download.html> 下載適合自己作業系統的版本安裝

■ 安裝SQLite Manager

- 打開Firefox 後至下列網址
<https://addons.mozilla.org/en-US/firefox/addon/sqlite-manager/> 下載適合自己作業系統的版本安裝

課前知識

■ 特性

- self-contained
- serverless
- zero-configuration
- Transactional

■ ACID 資料庫

■ 支援 SQL 92 語法

■ 開源

使用Pandas 將資料塞進資料庫

```
import sqlite3 as lite
with lite.connect('finance.sqlite') as db:
    stockdf.to_sql(name='trading_volume',
index=False, con=db, if_exists='replace')
```


開啟SQLite Manager



使用SQLite Manager瀏覽資料

檢視所有塞入的資料

SQLite Manager - C:\Users\User\pyfinance\finance.sqlite

Database (D) Table (T) Index (I) View (V) Trigger (T) Tools (O) Help

Directory (Select Profile Database) Go

finance.sqlite

Master Table (1)
Tables (1)
enterprise_eps
Views (0)
Indexes (0)
Triggers (0)

Structure Browse & Search Execute SQL DB Settings

TABLE enterprise_eps Search (H) Show All Add (A) Duplicate (D) Edit (E) Delete (L)

| rowid | 公司代號 | 公司名稱 | 產業別 | 基本每股盈... | 普通股每股... | 營業收入 | 營業利益 | 營業外收入... | 稅後淨利 | 民 |
|-------|------|----------|------|----------|--------------|-----------|-----------|-----------|---------|---|
| 1 | 1102 | 亞洲水泥股... | 水泥工業 | 0.4 | 新台幣 10.00... | 13931550 | 339801.0 | 1250044.0 | 1371559 | |
| 2 | 1101 | 台灣水泥股... | 水泥工業 | 0.38 | 新台幣 10.00... | 24114047 | 2026729.0 | 314060.0 | 1999624 | |
| 3 | 1104 | 環球水泥股... | 水泥工業 | 0.3 | 新台幣 10.00... | 1248072 | 30247.0 | 156012.0 | 183441 | |
| 4 | 1108 | 華僑水泥股... | 水泥工業 | 0.17 | 新台幣 10.00... | 1203671 | 98223.0 | -13612.0 | 63869 | |
| 5 | 1103 | 嘉新水泥股... | 水泥工業 | 0.13 | 新台幣 10.00... | 741189 | -149811.0 | 183613.0 | 59637 | |
| 6 | 1110 | 東南水泥股... | 水泥工業 | 0.11 | 新台幣 10.00... | 460227 | 51795.0 | 7706.0 | 61097 | |
| 7 | 1109 | 信大水泥股... | 水泥工業 | 0.09 | 新台幣 10.00... | 916242 | 35919.0 | 7715.0 | 35090 | |
| 8 | 1256 | 鮮活控股股... | 食品工業 | 2.59 | 新台幣 10.00... | 341578 | 57968.0 | 7039.0 | 45783 | |
| 9 | 1232 | 大統益股份... | 食品工業 | 1.32 | 新台幣 10.00... | 4882485 | 210352.0 | 49243.0 | 212608 | |
| 10 | 1227 | 佳格食品股... | 食品工業 | 1.23 | 新台幣 10.00... | 5076330 | 749193.0 | 93050.0 | 699669 | |
| 11 | 1231 | 聯華食品工... | 食品工業 | 0.91 | 新台幣 10.00... | 1482501 | 117267.0 | 16905.0 | 111496 | |
| 12 | 1233 | 天仁茶業股... | 食品工業 | 0.76 | 新台幣 10.00... | 532755 | 78968.0 | 4005.0 | 68672 | |
| 13 | 1216 | 統一企業股... | 食品工業 | 0.75 | 新台幣 10.00... | 104634790 | 5759096.0 | 1223048.0 | 5739800 | |
| 14 | 1702 | 南僑化學工... | 食品工業 | 0.6 | 新台幣 10.00... | 2917694 | 234161.0 | -12219.0 | 148329 | |
| 15 | 1236 | 宏亞食品股... | 食品工業 | 0.58 | 新台幣 10.00... | 726305 | 74045.0 | 624.0 | 62845 | |
| 16 | 1201 | 味全食品工... | 食品工業 | 0.58 | 新台幣 10.00... | 6700847 | 376786.0 | -25129.0 | 247114 | |
| 17 | 1210 | 大成興城企... | 食品工業 | 0.45 | 新台幣 10.00... | 21357129 | 250873.0 | 87210.0 | 291215 | |
| 18 | 1203 | 味王股份有... | 食品工業 | 0.35 | 新台幣 10.00... | 1643325 | 132636.0 | 8902.0 | 120200 | |
| 19 | 1215 | 台灣卜蜂企... | 食品工業 | 0.31 | 新台幣 10.00... | 4380233 | 97982.0 | -8828.0 | 68825 | |
| 20 | 1228 | 聯華食品股... | 食品工業 | 0.3 | 新台幣 10.00... | 1482501 | 117267.0 | 16905.0 | 111496 | |

<< < 1 to 100 of 9825 > >>

使用Pandas 下SQL 查詢資料

```
import sqlite3 as lite
with lite.connect('finance.sqlite') as db:
    df = pd.read_sql_query('SELECT count(1)
FROM trading_volume;', db)
df
```

The background features a light blue 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