勞動部產業新尖兵計畫

人工智慧金融應用與實務培訓班

課程模組： AI 金融科技課程 - 網路爬蟲技術

4. 網路爬蟲實做

葉建華 (Yeh, Jian-hua)

tdi.jhyeh@tdi.edu.tw
au4290@gmail.com

講次內容

- 使用 webbrowser 模組
- 使用 requests 模組
- 使用 urllib 模組
- Cookie？ 餅乾?
- 代理伺服器

使用 webbrowser 模組

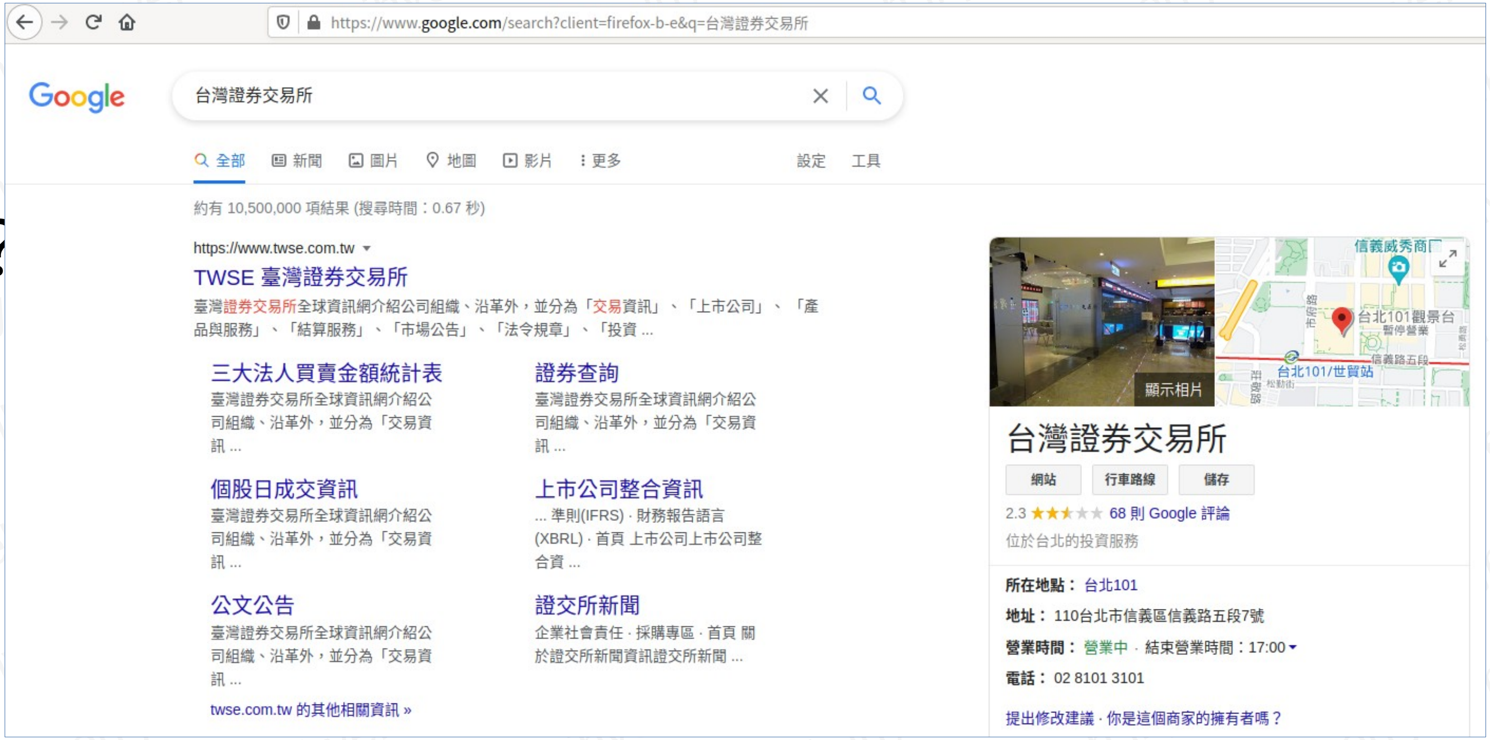
- Python 內建模組
- 轉介使用系統瀏覽器
- open() 方法帶入網址即可

使用 webbrowser 模組

- import webbrowser
- 查一下台灣證券交易所官網在哪吧!
  - 在哪?

使用 webbrowser 模組

- import webbrowser
- 查一下
  - 在哪?



https://www.google.com/search?client=firefox-b-e&q=台灣證券交易所

台灣證券交易所

約有 10,500,000 項結果 (搜尋時間：0.67 秒)

https://www.twse.com.tw

TWSE 臺灣證券交易所

臺灣證券交易所全球資訊網介紹公司組織、沿革外，並分為「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」、「投資 ...

三大法人買賣金額統計表
臺灣證券交易所全球資訊網介紹公司組織、沿革外，並分為「交易資訊 ...

證券查詢
臺灣證券交易所全球資訊網介紹公司組織、沿革外，並分為「交易資訊 ...

個股日成交資訊
臺灣證券交易所全球資訊網介紹公司組織、沿革外，並分為「交易資訊 ...

上市公司整合資訊
... 準則(IFRS) · 財務報告語言 (XBRL) · 首頁 上市公司上市公司整合資 ...

公文公告
臺灣證券交易所全球資訊網介紹公司組織、沿革外，並分為「交易資訊 ...

證交所新聞
企業社會責任 · 採購專區 · 首頁 關於證交所新聞資訊證交所新聞 ...

twse.com.tw 的其他相關資訊 »

台灣證券交易所

2.3 ★★★★★ 68 則 Google 評論

位於台北的投資服務

所在地點： 台北101

地址： 110台北市信義區信義路五段7號

營業時間： 營業中 · 結束營業時間：17:00

電話： 02 8101 3101

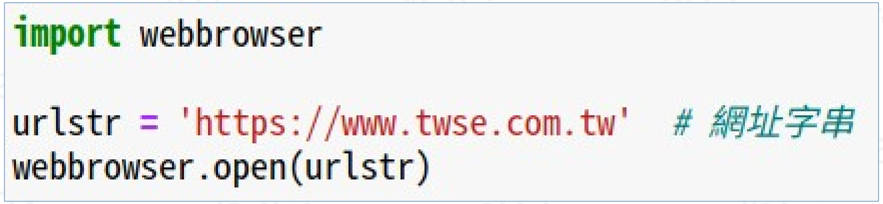
提出修改建議 · 你是這個商家的擁有者嗎？

台灣證券交易所官網

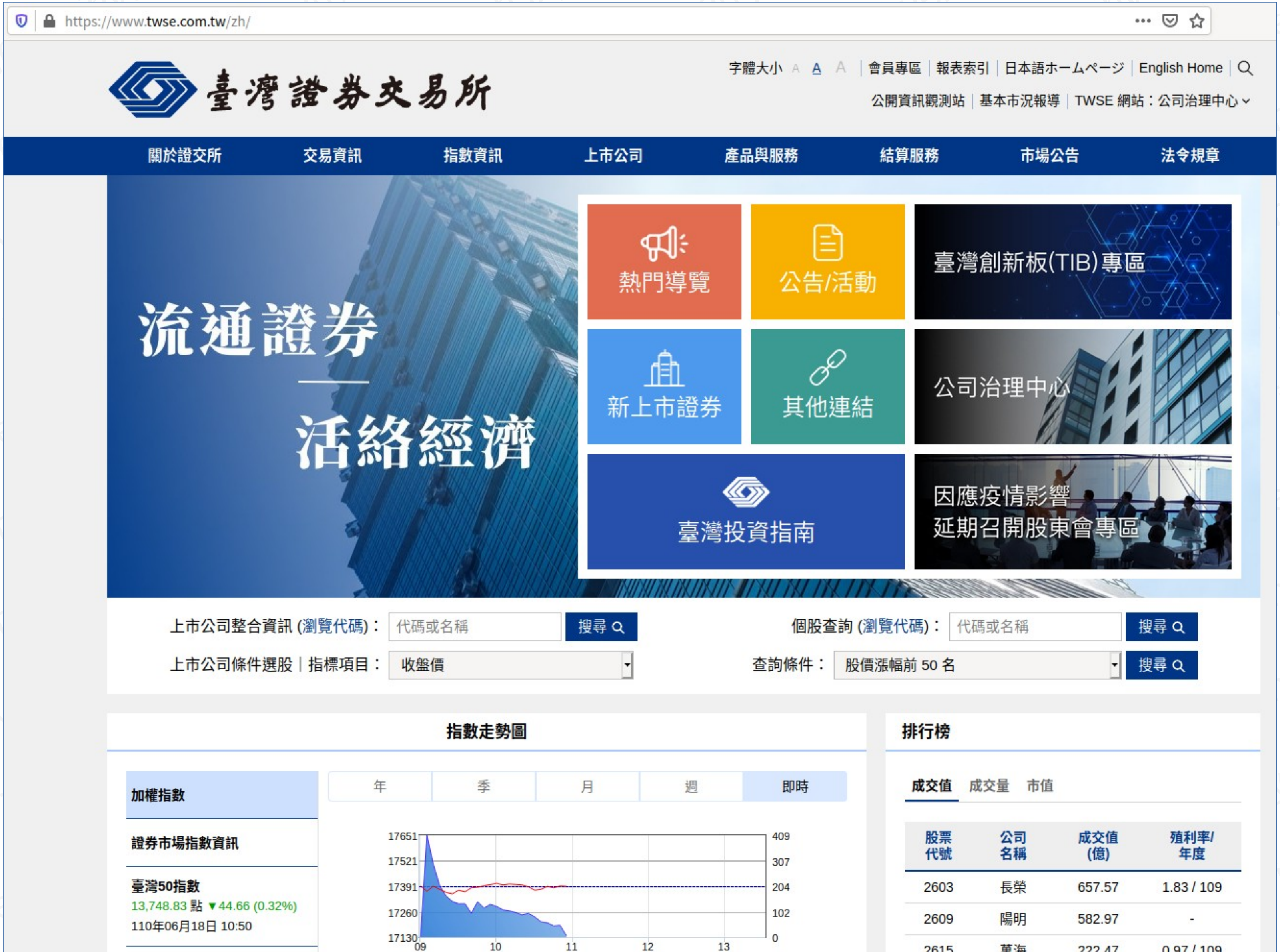
- https://www.twse.com.tw
- 是 https 哦！ （還記得嗎？）
- .com 代表商業公司單位
- .tw 代表台灣地區的機構
- 用 webbrowser.open() 來開啟吧！

台灣證券交易所官網

- https://www.twse.com.tw
- 是 https 哦！ （還記得嗎？）
- .com 代表商業公司單位
- .tw 代表台灣地區的機構
- 用 webbrowser.open() 來開啟吧！



```
import webbrowser

urlstr = 'https://www.twse.com.tw'  # 網址字串
webbrowser.open(urlstr)
```



https://www.twse.com.tw/zh/

臺灣證券交易所

上市公司整合資訊 (瀏覽代碼)： 代碼或名稱 搜尋

個股查詢 (瀏覽代碼)： 代碼或名稱 搜尋

上市公司條件選股 | 指標項目： 收盤價

查詢條件： 股價漲幅前 50 名 搜尋

指數走勢圖

加權指數

證券市場指數資訊

臺灣50指數
13,748.83 點 ▼44.66 (0.32%)
110年06月18日 10:50

排行榜

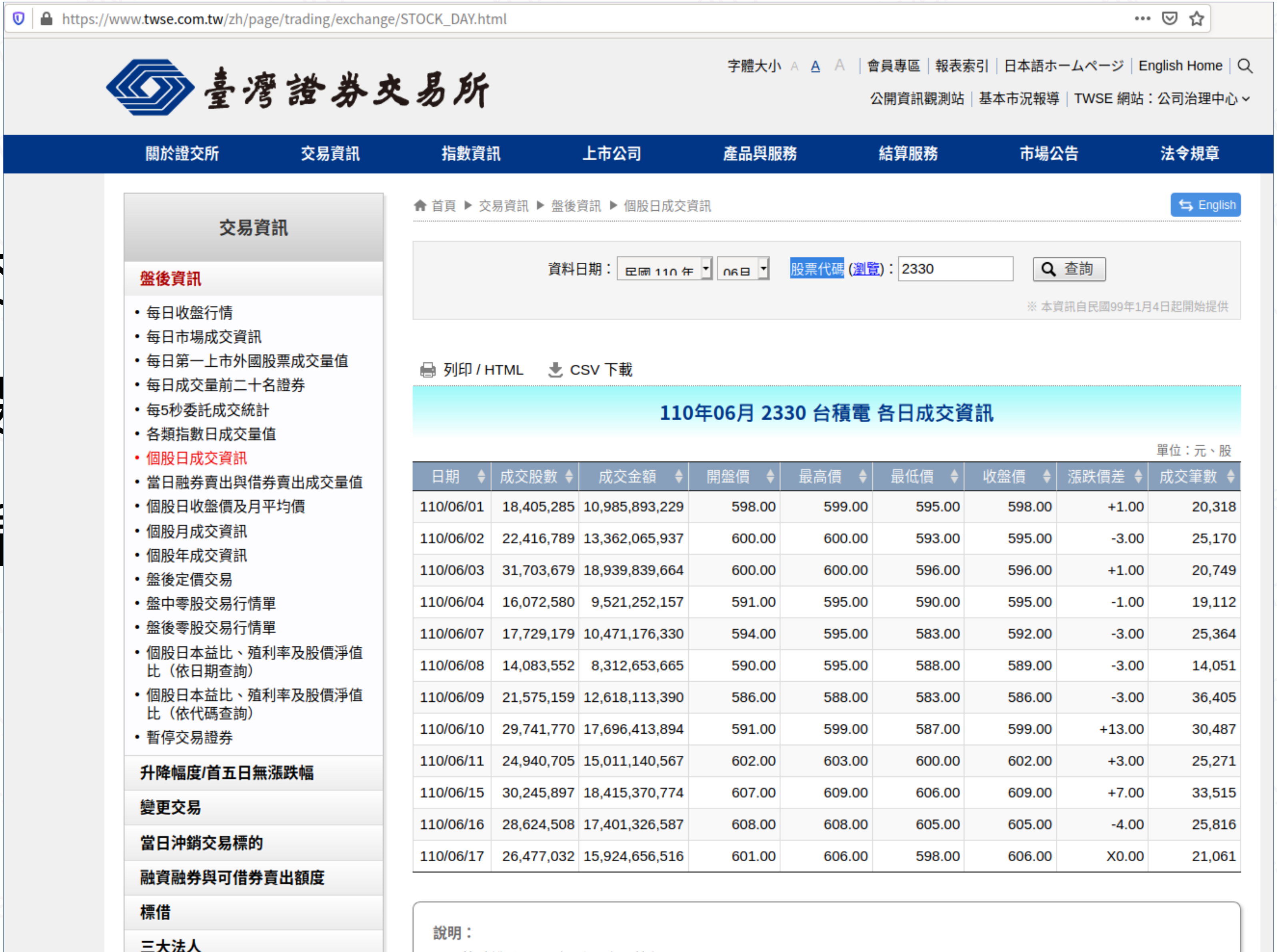
成交值 成交量 市值

股票代號	公司名稱	成交值(億)	殖利率/年度
2603	長榮	657.57	1.83 / 109
2609	陽明	582.97	-
2615	萬海	222.47	0.97 / 109

瀏覽一下 TWSE...

- 交易資訊 -> 個股日成交資訊
- 股票代碼輸入 2330，按「查詢」鈕
- 看到了什麼?

- 交
- 股
- 看



https://www.twse.com.tw/zh/page/trading/exchange/STOCK_DAY.html

臺灣證券交易所

字體大小 A A A | 會員專區 | 報表索引 | 日本語ホームページ | English Home

公開資訊觀測站 | 基本市況報導 | TWSE 網站：公司治理中心

關於證交所 交易資訊 指數資訊 上市公司 產品與服務 結算服務 市場公告 法令規章

首頁 ▶ 交易資訊 ▶ 盤後資訊 ▶ 個股日成交資訊

English

交易資訊

盤後資訊
- 每日收盤行情
- 每日市場成交資訊
- 每日第一上市外國股票成交量值
- 每日成交量前二十名證券
- 每5秒委託成交統計
- 各類指數日成交量值
- 個股日成交資訊
- 當日融券賣出與借券賣出成交量值
- 個股日收盤價及月平均價
- 個股月成交資訊
- 個股年成交資訊
- 盤後定價交易
- 盤中零股交易行情單
- 盤後零股交易行情單
- 個股日本益比、殖利率及股價淨值比（依日期查詢）
- 個股日本益比、殖利率及股價淨值比（依代碼查詢）
- 暫停交易證券

升降幅度/首五日無漲跌幅

變更交易

當日沖銷交易標的

融資融券與可借券賣出額度

標借

三大法人

資料日期：民國 110 年 06月 股票代碼 (瀏覽)：2330 查詢

※ 本資訊自民國99年1月4日起開始提供

列印 / HTML CSV 下載

110年06月 2330 台積電 各日成交資訊

單位：元、股

日期	成交股數	成交金額	開盤價	最高價	最低價	收盤價	漲跌價差	成交筆數
110/06/01	18,405,285	10,985,893,229	598.00	599.00	595.00	598.00	+1.00	20,318
110/06/02	22,416,789	13,362,065,937	600.00	600.00	593.00	595.00	-3.00	25,170
110/06/03	31,703,679	18,939,839,664	600.00	600.00	596.00	596.00	+1.00	20,749
110/06/04	16,072,580	9,521,252,157	591.00	595.00	590.00	595.00	-1.00	19,112
110/06/07	17,729,179	10,471,176,330	594.00	595.00	583.00	592.00	-3.00	25,364
110/06/08	14,083,552	8,312,653,665	590.00	595.00	588.00	589.00	-3.00	14,051
110/06/09	21,575,159	12,618,113,390	586.00	588.00	583.00	586.00	-3.00	36,405
110/06/10	29,741,770	17,696,413,894	591.00	599.00	587.00	599.00	+13.00	30,487
110/06/11	24,940,705	15,011,140,567	602.00	603.00	600.00	602.00	+3.00	25,271
110/06/15	30,245,897	18,415,370,774	607.00	609.00	606.00	609.00	+7.00	33,515
110/06/16	28,624,508	17,401,326,587	608.00	608.00	605.00	605.00	-4.00	25,816
110/06/17	26,477,032	15,924,656,516	601.00	606.00	598.00	606.00	X0.00	21,061

說明：

2330 各日成交資訊

- 玩一下「 CSV 下載」連結
  - 按右鍵，選「複製連結網址」

    （或稱 Copy Link Location ）
  - 找個記事本貼上，看到了什麼?

2330 各日成交資訊

- 玩一下「 CSV 下載」連結
  - 按右鍵，選「複製連結網址」

    （或稱 Copy Link Location ）
  - 找個記事本貼上，看到了什麼?
    - https://www.twse.com.tw/exchangeReport/STOCK_DAY?response=csv&date=20210618&stockNo=2330

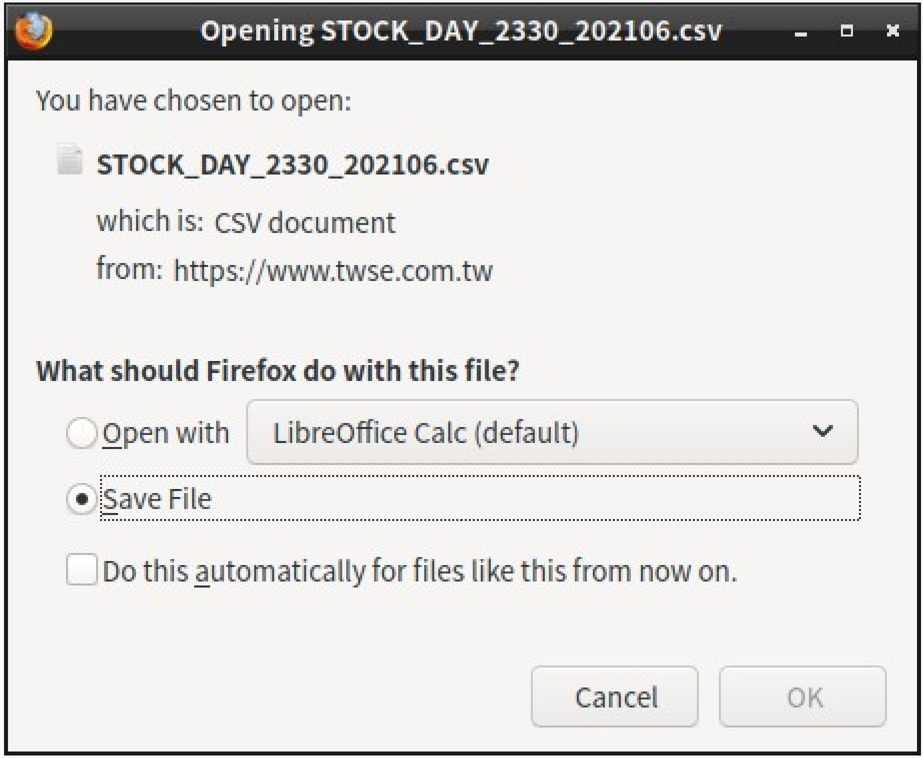
解析網址與參數

- https://www.twse.com.tw/exchangeReport/STOCK_DAY?response=csv&date=20210618&stockNo=2330
  - 網頁程式網址： https://www.twse.com.tw/exchangeReport/STOCK_DAY
  - 參數1： response，參數值csv代表輸出CSV格式
  - 參數2： date，參數值20210618代表輸出資料參考日期
  - 參數3： stockNo，參數值2330代表台積電股票代碼

把這個網址用webbrowser開啟，會發生什麼事?

解析網址與參數

- https://www.tw STOCK_DAY? response=csv 330
  - 網頁程式網址 angeReport/ STOCK_DAY
  - 參數 1： res 格式
  - 參數 2： dat 斗參考日期
  - 參數 3： sto 票代碼



Opening STOCK_DAY_2330_202106.csv

You have chosen to open:

STOCK_DAY_2330_202106.csv

which is: CSV document

from: https://www.twse.com.tw

What should Firefox do with this file?

○ Open with LibreOffice Calc (default)

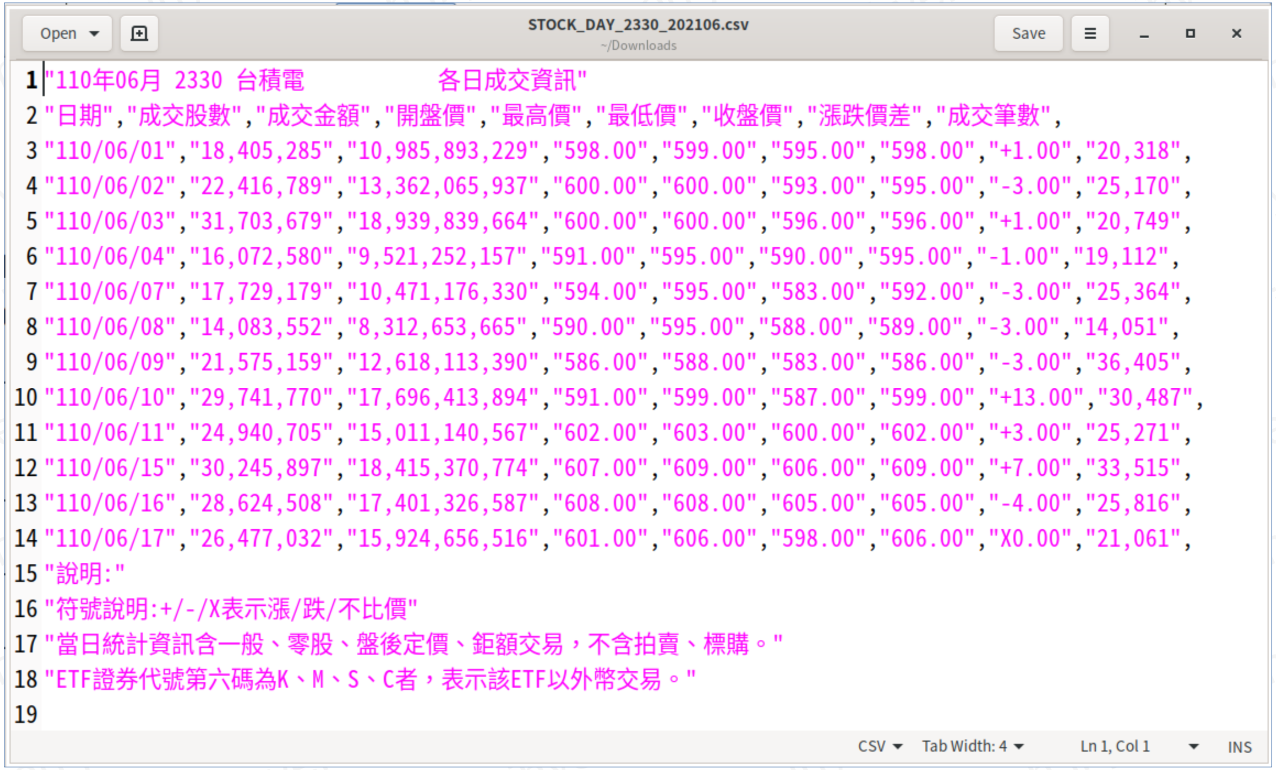
● Save File

□ Do this automatically for files like this from now on.

出現另存新檔畫面？！ 為什麼?

解析網址與參數

- 瀏覽器無法顯示的資料格式，會跳出「另存」



```
"110年06月 2330 台積電           各日成交資訊"
"日期","成交股數","成交金額","開盤價","最高價","最低價","收盤價","漲跌價差","成交筆數",
"110/06/01","18,405,285","10,985,893,229","598.00","599.00","595.00","598.00","+1.00","20,318",
"110/06/02","22,416,789","13,362,065,937","600.00","600.00","593.00","595.00","-3.00","25,170",
"110/06/03","31,703,679","18,939,839,664","600.00","600.00","596.00","596.00","+1.00","20,749",
"110/06/04","16,072,580","9,521,252,157","591.00","595.00","590.00","595.00","-1.00","19,112",
"110/06/07","17,729,179","10,471,176,330","594.00","595.00","583.00","592.00","-3.00","25,364",
"110/06/08","14,083,552","8,312,653,665","590.00","595.00","588.00","589.00","-3.00","14,051",
"110/06/09","21,575,159","12,618,113,390","586.00","588.00","583.00","586.00","-3.00","36,405",
"110/06/10","29,741,770","17,696,413,894","591.00","599.00","587.00","599.00","+13.00","30,487",
"110/06/11","24,940,705","15,011,140,567","602.00","603.00","600.00","602.00","+3.00","25,271",
"110/06/15","30,245,897","18,415,370,774","607.00","609.00","606.00","609.00","+7.00","33,515",
"110/06/16","28,624,508","17,401,326,587","608.00","608.00","605.00","605.00","-4.00","25,816",
"110/06/17","26,477,032","15,924,656,516","601.00","606.00","598.00","606.00","X0.00","21,061",
"說明:"
"符號說明:+/-/X表示漲/跌/不比價"
"當日統計資訊含一般、零股、盤後定價、鉅額交易，不含拍賣、標購。"
"ETF證券代號第六碼為K、M、S、C者，表示該ETF以外幣交易。"

```

練習：中鋼今年一月各日成交資訊

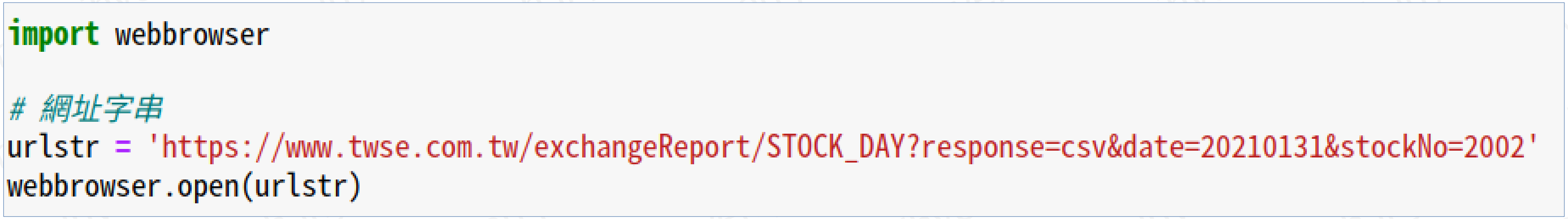
- 剛剛抓到台積電的各日成交資訊，使用的網址：
  - https://www.twse.com.tw/exchangeReport/STOCK_DAY?response=csv&date=20210618&stockNo=2330
- 抓一下中鋼的今年一月的各日成交資訊
- 怎麼做?

練習：中鋼今年一月各日成交資訊

- 中鋼代碼： 2002
- 抓今年一月整個月，我用 20210131 參數值
- 所以網址會長成什麼樣?

練習： 中鋼今年一月各日成交資訊

- https://www.twse.com.tw/exchangeReport/STOCK_DAY?response=csv&date=20210131&stockNo=2002
- 用 webbrowser.open()



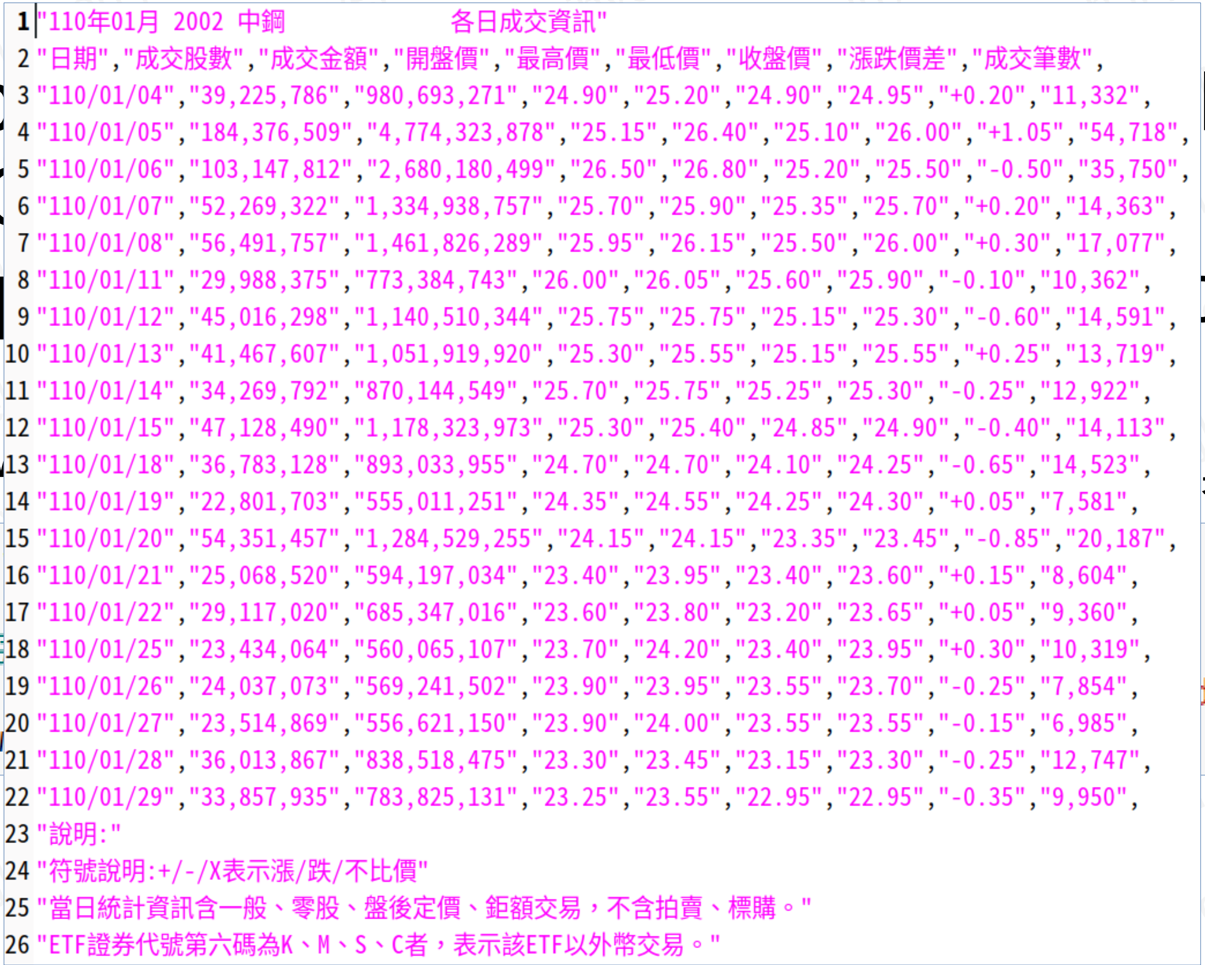
```python
import webbrowser

# 網址字串
urlstr = 'https://www.twse.com.tw/exchangeReport/STOCK_DAY?response=csv&date=20210131&stockNo=2002'
webbrowser.open(urlstr)
```

練習：中鋼今年一月各日成交資訊

- http   Report/
  STO
  res   ockNo=2002

- 用 w   果然看到一月整月的資料!

```
import

# 網址
urlstr                                        date=20210131&stockNo=2002'
webbrow
```



```
"110年01月 2002 中鋼             各日成交資訊"
"日期","成交股數","成交金額","開盤價","最高價","最低價","收盤價","漲跌價差","成交筆數",
"110/01/04","39,225,786","980,693,271","24.90","25.20","24.90","24.95","+0.20","11,332",
"110/01/05","184,376,509","4,774,323,878","25.15","26.40","25.10","26.00","+1.05","54,718",
"110/01/06","103,147,812","2,680,180,499","26.50","26.80","25.20","25.50","-0.50","35,750",
"110/01/07","52,269,322","1,334,938,757","25.70","25.90","25.35","25.70","+0.20","14,363",
"110/01/08","56,491,757","1,461,826,289","25.95","26.15","25.50","26.00","+0.30","17,077",
"110/01/11","29,988,375","773,384,743","26.00","26.05","25.60","25.90","-0.10","10,362",
"110/01/12","45,016,298","1,140,510,344","25.75","25.75","25.15","25.30","-0.60","14,591",
"110/01/13","41,467,607","1,051,919,920","25.30","25.55","25.15","25.55","+0.25","13,719",
"110/01/14","34,269,792","870,144,549","25.70","25.75","25.25","25.30","-0.25","12,922",
"110/01/15","47,128,490","1,178,323,973","25.30","25.40","24.85","24.90","-0.40","14,113",
"110/01/18","36,783,128","893,033,955","24.70","24.70","24.10","24.25","-0.65","14,523",
"110/01/19","22,801,703","555,011,251","24.35","24.55","24.25","24.30","+0.05","7,581",
"110/01/20","54,351,457","1,284,529,255","24.15","24.15","23.35","23.45","-0.85","20,187",
"110/01/21","25,068,520","594,197,034","23.40","23.95","23.40","23.60","+0.15","8,604",
"110/01/22","29,117,020","685,347,016","23.60","23.80","23.20","23.65","+0.05","9,360",
"110/01/25","23,434,064","560,065,107","23.70","24.20","23.40","23.95","+0.30","10,319",
"110/01/26","24,037,073","569,241,502","23.90","23.95","23.55","23.70","-0.25","7,854",
"110/01/27","23,514,869","556,621,150","23.90","24.00","23.55","23.55","-0.15","6,985",
"110/01/28","36,013,867","838,518,475","23.30","23.45","23.15","23.30","-0.25","12,747",
"110/01/29","33,857,935","783,825,131","23.25","23.55","22.95","22.95","-0.35","9,950",
"說明:"
"符號說明:+/-/X表示漲/跌/不比價"
"當日統計資訊含一般、零股、盤後定價、鉅額交易，不含拍賣、標購。"
"ETF證券代號第六碼為K、M、S、C者，表示該ETF以外幣交易。"
```

然後?

你都學會 CSV 處理了

不是嗎？

後續練習

- 怎麼抓 2020 年一整年的中鋼各日成交資訊?
- 怎麼抓全市場一整年的各日成交資訊?

小心，有可能會被台灣證券交易所封鎖哦!

講次內容

- 使用 webbrowser 模組
- 使用 requests 模組
- 使用 urllib 模組
- Cookie？ 餅乾?
- 代理伺服器

使用 requests 模組

- 第三方模組
  - 需要額外安裝： pip install requests

Anaconda 早就幫你裝好了！

- 功能很多！
  - 資料自動解碼，資料自動解壓縮，基本認證
  - 支援連接池、連接逾時
  - 支援代理伺服器 ... 等等

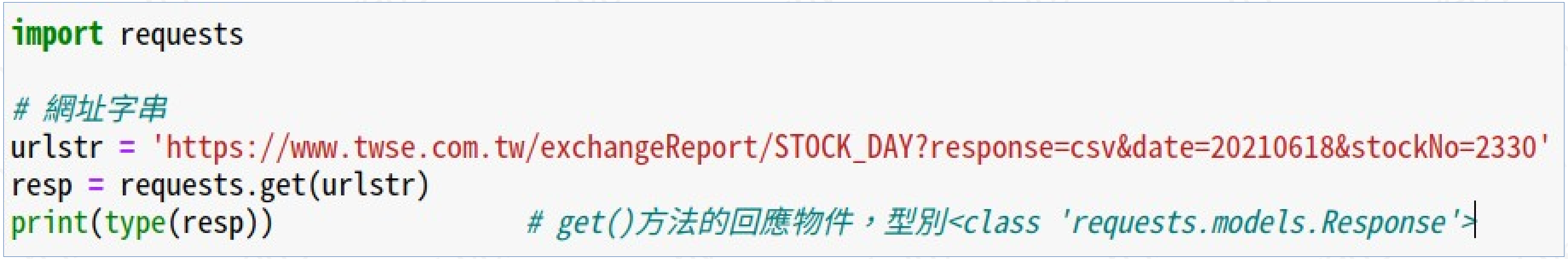
再訪 2330 各日成交資訊

- https://www.twse.com.tw/exchangeReport/STOCK_DAY?response=csv&date=20210618&stockNo=2330
  - 網頁程式網址： https://www.twse.com.tw/exchangeReport/STOCK_DAY
  - 參數 1： response，參數值 csv 代表輸出 CSV 格式
  - 參數 2： date，參數值 20210618 代表輸出資料參考日期
  - 參數 3： stockNo，參數值 2330 代表台積電股票代碼

這次用 requests 開啟，會發生什麼事?

再訪 2330 各日成交資訊

- 使用 requests.get() 方法可以傳回 HTML 內容

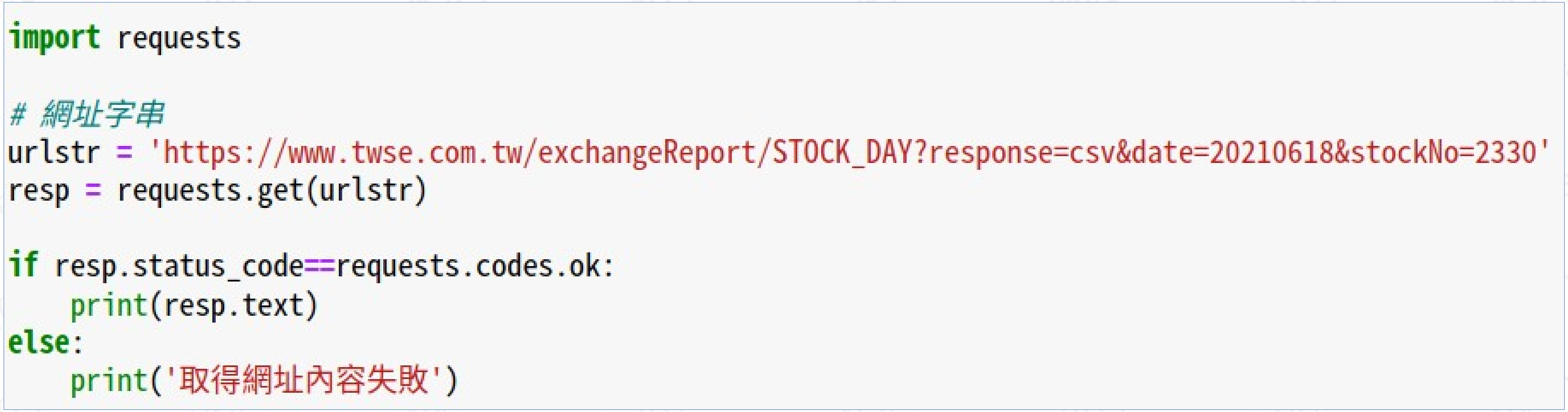


```python
import requests

# 網址字串
urlstr = 'https://www.twse.com.tw/exchangeReport/STOCK_DAY?response=csv&date=20210618&stockNo=2330'
resp = requests.get(urlstr)
print(type(resp))              # get()方法的回應物件，型別<class 'requests.models.Response'>
```

- 回應的物件是 Response
  - 屬性 status_code： HTTP 協定的傳回碼，代表請求狀態
  - 屬性 text： HTML 內容

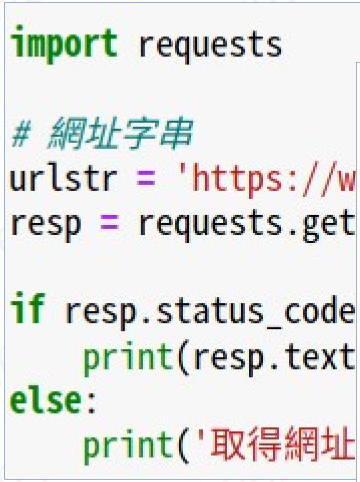
requests.codes.ok 代表成功!

再訪 2330 各日成交資訊

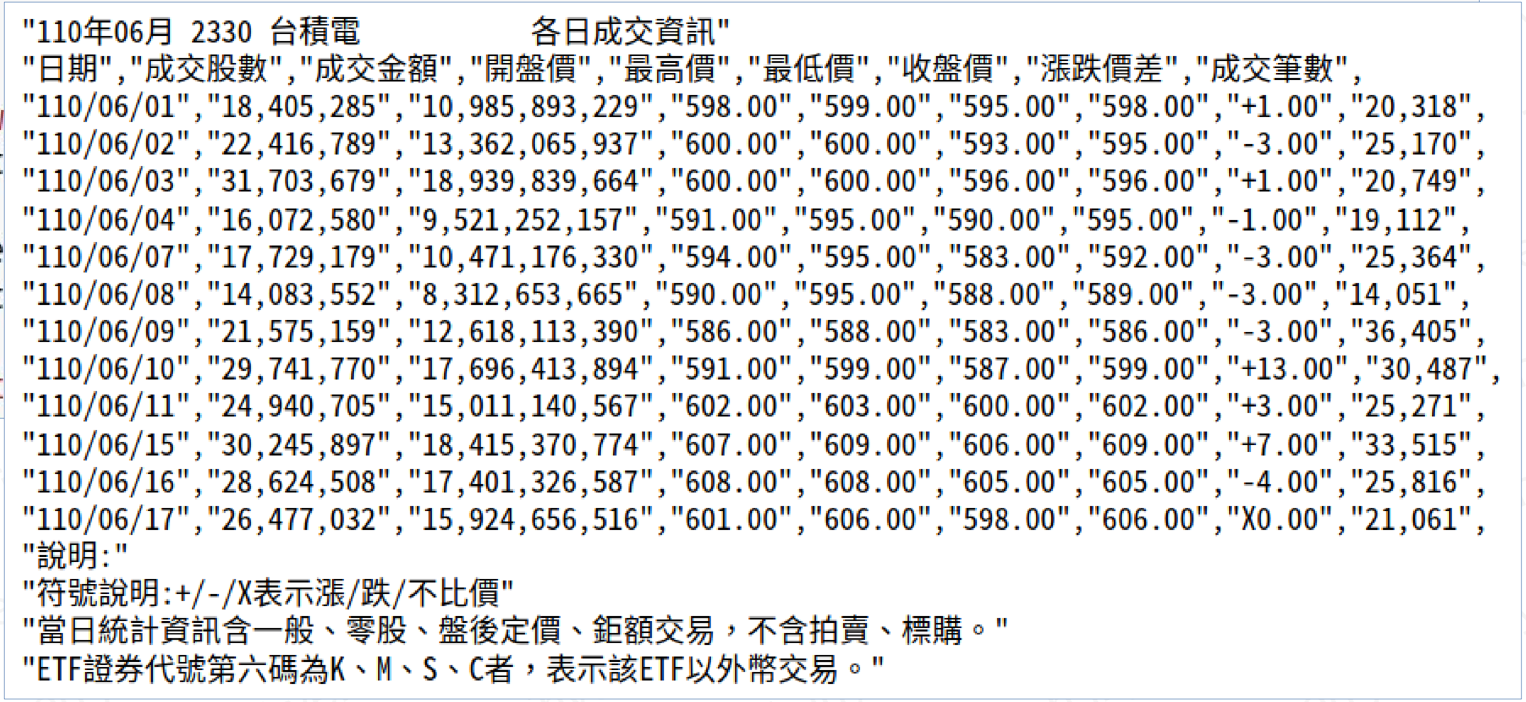


```python
import requests

# 網址字串
urlstr = 'https://www.twse.com.tw/exchangeReport/STOCK_DAY?response=csv&date=20210618&stockNo=2330'
resp = requests.get(urlstr)

if resp.status_code==requests.codes.ok:
    print(resp.text)
else:
    print('取得網址內容失敗')
```

再訪 2330 各日成交資訊



```
import requests

# 網址字串
urlstr = 'https://w
resp = requests.get

if resp.status_code
    print(resp.text
else:
    print('取得網址
```



```
"110年06月 2330 台積電           各日成交資訊"
"日期","成交股數","成交金額","開盤價","最高價","最低價","收盤價","漲跌價差","成交筆數",
"110/06/01","18,405,285","10,985,893,229","598.00","599.00","595.00","598.00","+1.00","20,318",
"110/06/02","22,416,789","13,362,065,937","600.00","600.00","593.00","595.00","-3.00","25,170",
"110/06/03","31,703,679","18,939,839,664","600.00","600.00","596.00","596.00","+1.00","20,749",
"110/06/04","16,072,580","9,521,252,157","591.00","595.00","590.00","595.00","-1.00","19,112",
"110/06/07","17,729,179","10,471,176,330","594.00","595.00","583.00","592.00","-3.00","25,364",
"110/06/08","14,083,552","8,312,653,665","590.00","595.00","588.00","589.00","-3.00","14,051",
"110/06/09","21,575,159","12,618,113,390","586.00","588.00","583.00","586.00","-3.00","36,405",
"110/06/10","29,741,770","17,696,413,894","591.00","599.00","587.00","599.00","+13.00","30,487",
"110/06/11","24,940,705","15,011,140,567","602.00","603.00","600.00","602.00","+3.00","25,271",
"110/06/15","30,245,897","18,415,370,774","607.00","609.00","606.00","609.00","+7.00","33,515",
"110/06/16","28,624,508","17,401,326,587","608.00","608.00","605.00","605.00","-4.00","25,816",
"110/06/17","26,477,032","15,924,656,516","601.00","606.00","598.00","606.00","X0.00","21,061",
"說明:"
"符號說明:+/-/X表示漲/跌/不比價"
"當日統計資訊含一般、零股、盤後定價、鉅額交易，不含拍賣、標購。"
"ETF證券代號第六碼為K、M、S、C者，表示該ETF以外幣交易。"
```

網頁有可能下載異常

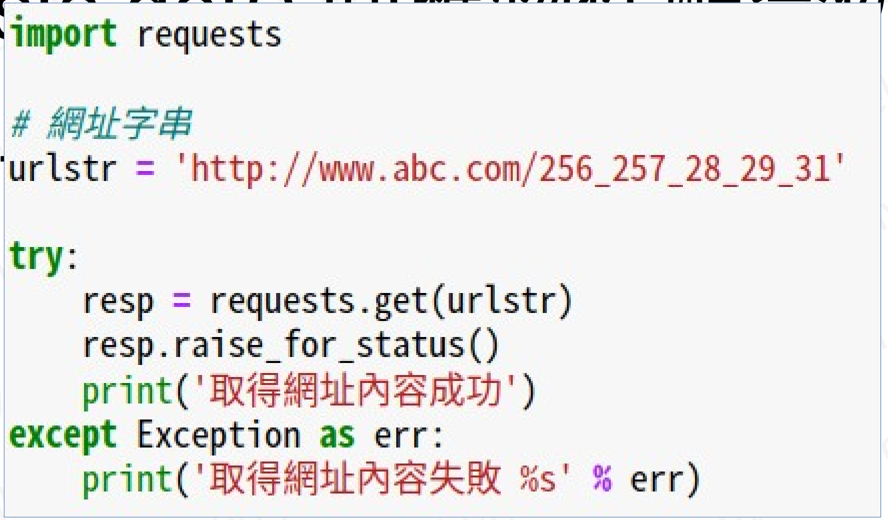
- 下載異常狀況不一定都是網址錯誤!
  - 網路連線異常（網路服務有問題）
  - 頻繁存取遭擋（伺服端反爬蟲機制）
  - 不符存取規範（伺服端反爬蟲機制）

網頁下載異常處理

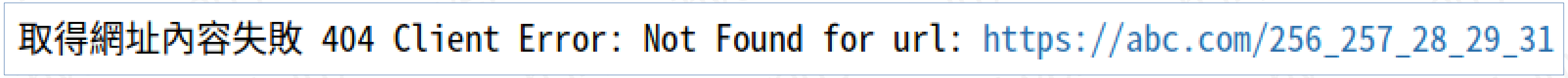
- 使用 requests.get() 回應物件配合例外機制
- raise_for_status() 方法

網頁下載異常處理

- 使用 requests.get() 回應物件配合例外機制
- raise_for_s



```
import requests

# 網址字串
urlstr = 'http://www.abc.com/256_257_28_29_31'

try:
    resp = requests.get(urlstr)
    resp.raise_for_status()
    print('取得網址內容成功')
except Exception as err:
    print('取得網址內容失敗 %s' % err)
```



```
取得網址內容失敗 404 Client Error: Not Found for url: https://abc.com/256_257_28_29_31
```

網頁伺服器會擋？！

- 網頁伺服器採取反爬蟲動作的理由
  - 基於安全理由，拒絕頻繁存取
  - 不想增加伺服器的網路流量（負擔）
  - 希望有規範的存取 (robots.txt)

HTTP 協定錯誤碼

- 常見的錯誤碼

錯誤碼	代碼意義
400	request 不正確
401	request 不被授權
402	request 完成必須有所回應
403	禁止使用這項資源
404	request 的資源不存在
405	不允許 request 資源的方法
406	不接受此客戶端

有時候 request 錯誤是因為 ...

你不是「真的」瀏覽器！

被擋了啦！

如何偽裝成瀏覽器

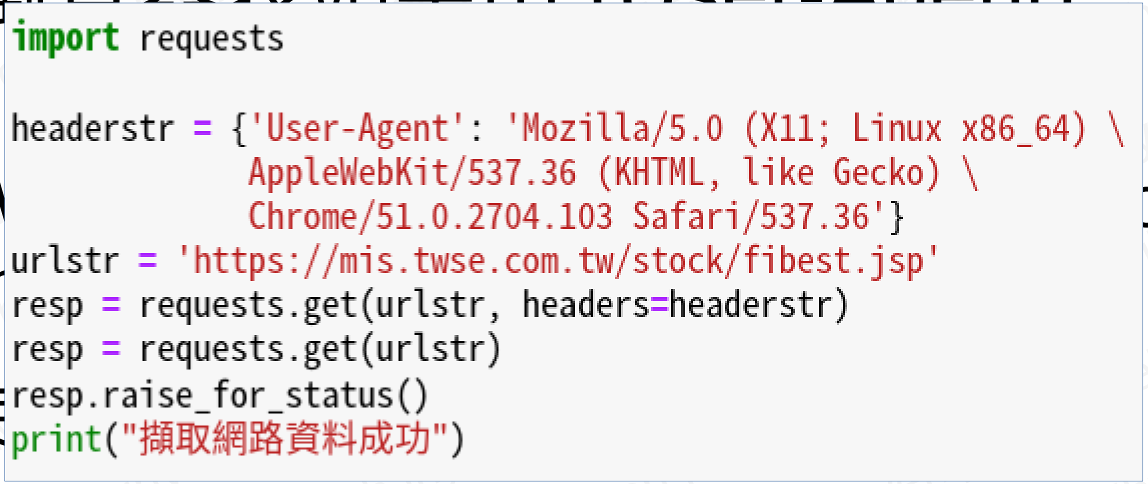
- 使用 requests.get() 時，你就已經是網路爬蟲了
  - 所以你不是瀏覽器！別裝了！
- 真的要裝？那要做功課
  - 最簡單的偽裝法：透過 HTTP request header
  - 表頭 (header) 偽裝法：'User-Agent'

HTTP 表頭偽裝

- 常見的瀏覽器表頭字串 (User-Agent)
  - 'Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/51.0.2704.103 Safari/537.36'
  - 如何使用：用 'User-Agent' 當 key，上述字串當 value，傳入 requests.get() 的 headers 參數

HTTP 表頭偽裝

- 常見的瀏覽器表頭字串 (User-Agent)
  - 'Mozill
    AppleW o)
    Chrom
  - 如何使 字串當 value， 傳入 requests.get() 的 headers 參數



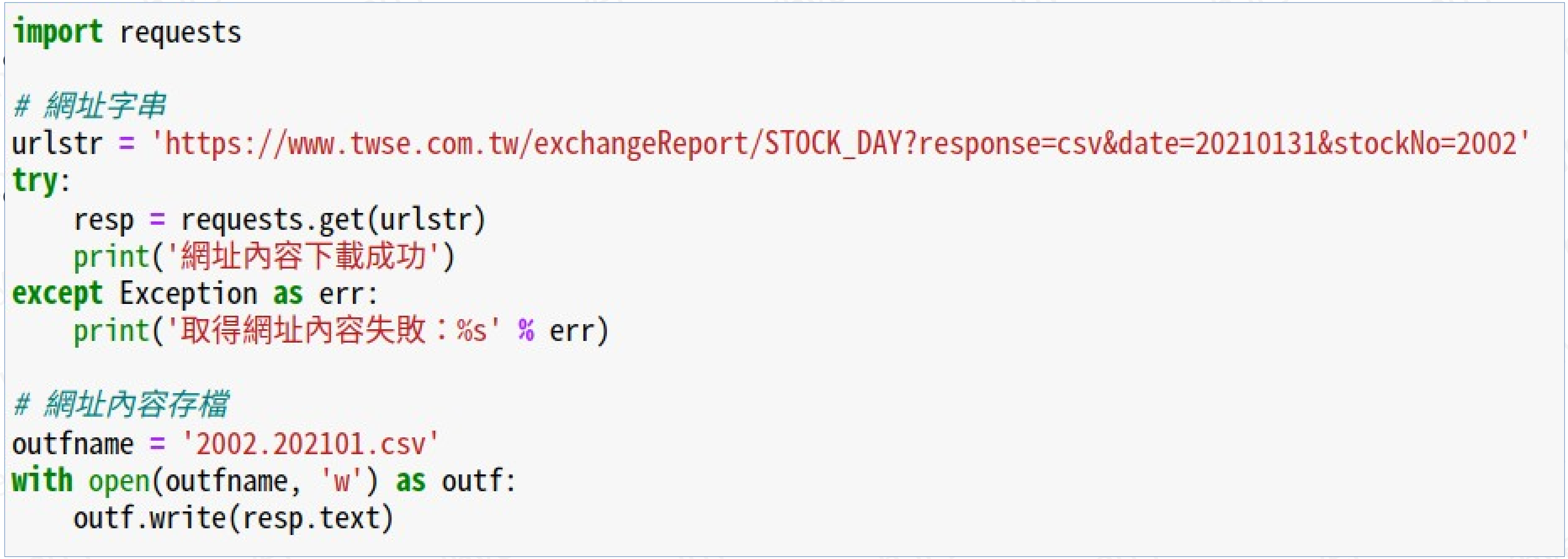
```python
import requests

headerstr = {'User-Agent': 'Mozilla/5.0 (X11; Linux x86_64) \
             AppleWebKit/537.36 (KHTML, like Gecko) \
             Chrome/51.0.2704.103 Safari/537.36'}
urlstr = 'https://mis.twse.com.tw/stock/fibest.jsp'
resp = requests.get(urlstr, headers=headerstr)
resp = requests.get(urlstr)
resp.raise_for_status()
print("擷取網路資料成功")
```

robots.txt?

- 儲存在網站根目錄的檔案
- 並非正式規範（但是大家已經用習慣了）
- 內容
  - User-Agent： 描述針對的 robot
  - Disallow： 不允許存取的範疇

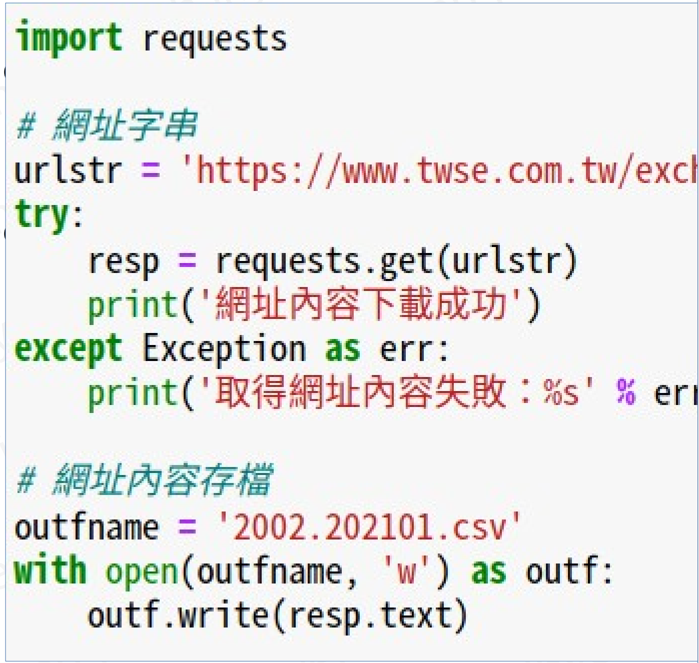
```
# 允許所有網路爬蟲，沒有限制存取
User-Agent: *
Disallow:

# 允許特定網路爬蟲，沒有限制存取
User-Agent: Applebot
Disallow:

# 攔截所有網路爬蟲，限制存取任何資料
User-Agent: *
Disallow: /

# 攔截所有網路爬蟲，限制存取特定目錄
User-Agent: *
Disallow: /private/

# 攔截所有網路爬蟲，限制存取特定檔案
User-Agent: *
Disallow: /*.jsp$
```

儲存下載內容：文字式

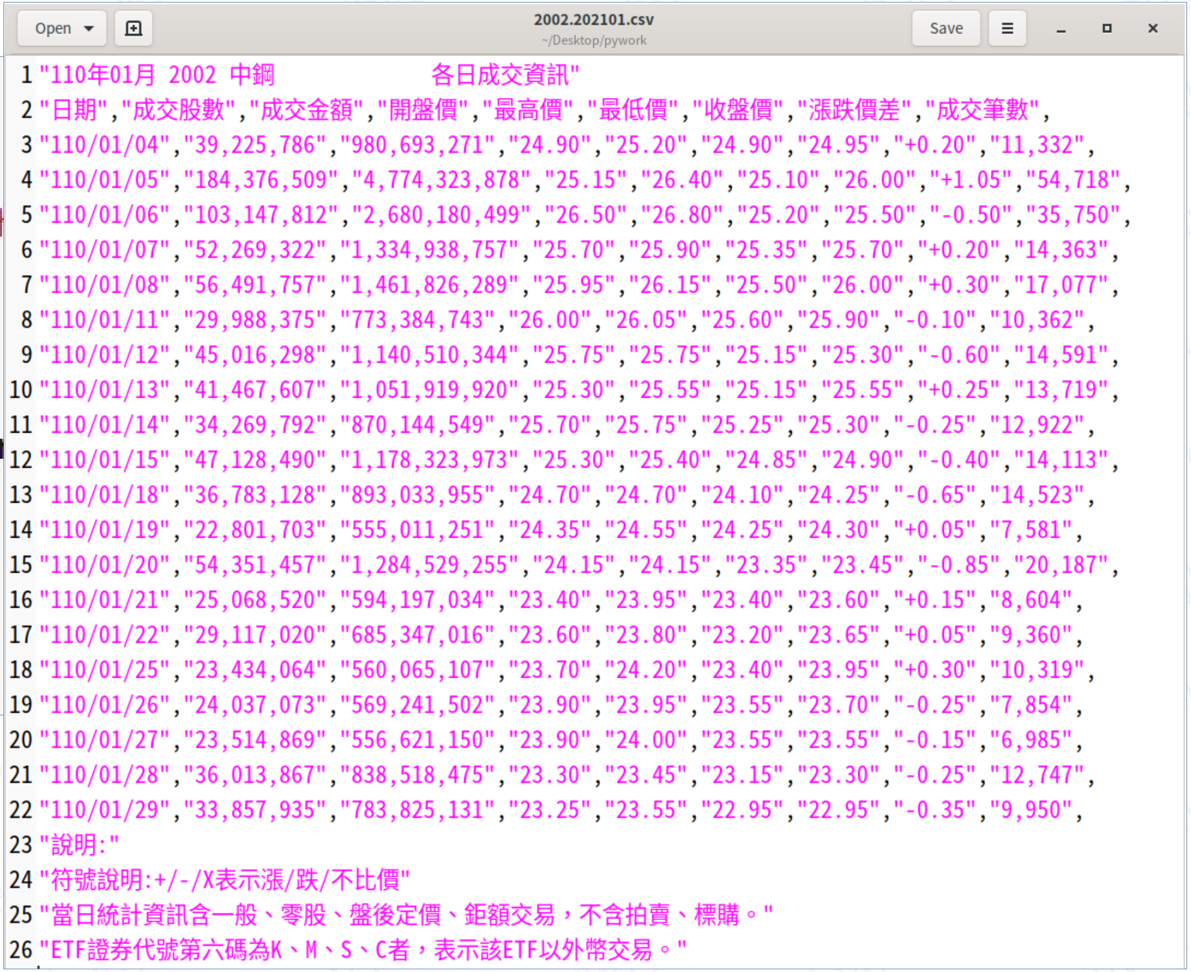


```
import requests

# 網址字串
urlstr = 'https://www.twse.com.tw/exchangeReport/STOCK_DAY?response=csv&date=20210131&stockNo=2002'
try:
    resp = requests.get(urlstr)
    print('網址內容下載成功')
except Exception as err:
    print('取得網址內容失敗：%s' % err)

# 網址內容存檔
outfname = '2002.202101.csv'
with open(outfname, 'w') as outf:
    outf.write(resp.text)
```

儲存下載內容：文字式

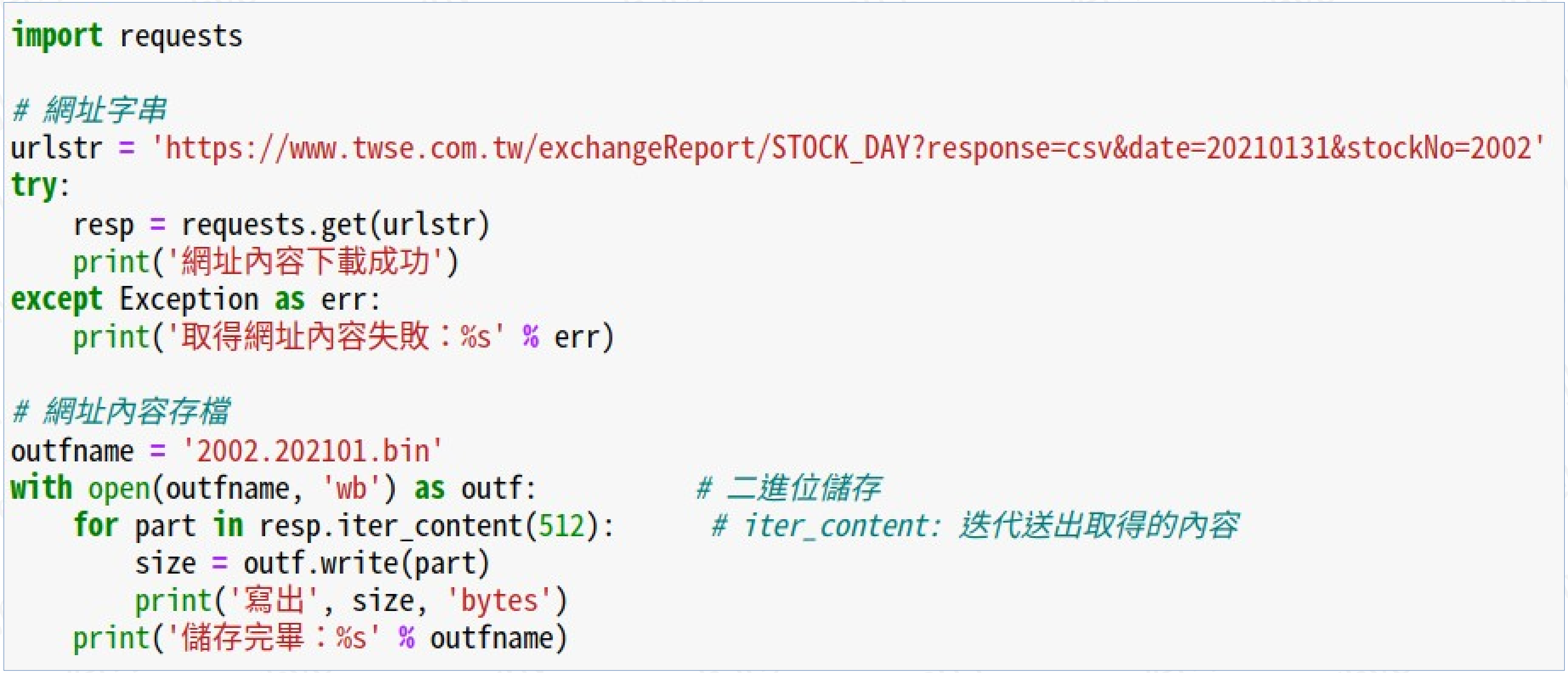


```
import requests

# 網址字串
urlstr = 'https://www.twse.com.tw/exch
try:
    resp = requests.get(urlstr)
    print('網址內容下載成功')
except Exception as err:
    print('取得網址內容失敗：%s' % err

# 網址內容存檔
outfname = '2002.202101.csv'
with open(outfname, 'w') as outf:
    outf.write(resp.text)
```



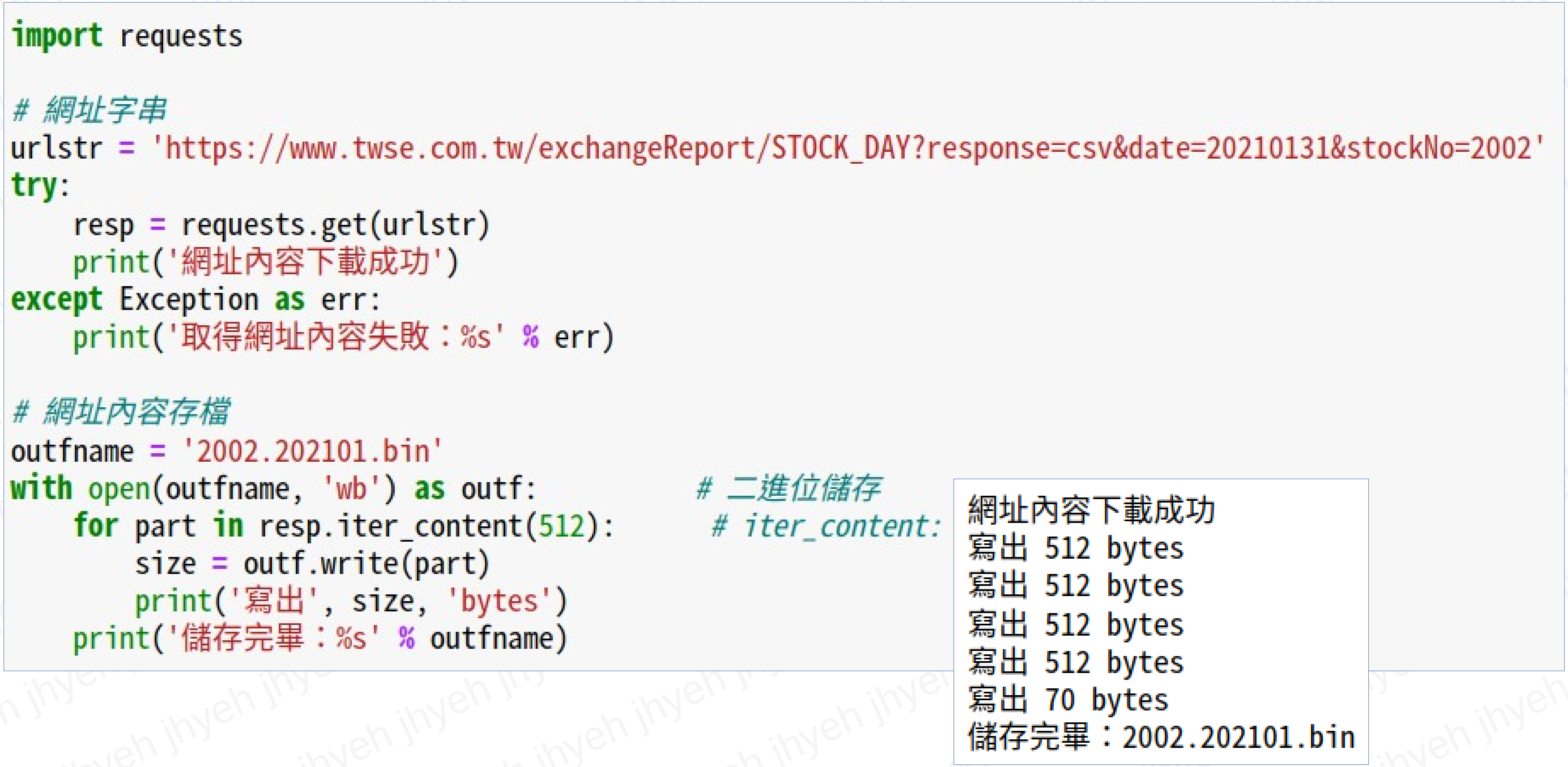
```
"110年01月 2002 中鋼                各日成交資訊"
"日期","成交股數","成交金額","開盤價","最高價","最低價","收盤價","漲跌價差","成交筆數",
"110/01/04","39,225,786","980,693,271","24.90","25.20","24.90","24.95","+0.20","11,332",
"110/01/05","184,376,509","4,774,323,878","25.15","26.40","25.10","26.00","+1.05","54,718",
"110/01/06","103,147,812","2,680,180,499","26.50","26.80","25.20","25.50","-0.50","35,750",
"110/01/07","52,269,322","1,334,938,757","25.70","25.90","25.35","25.70","+0.20","14,363",
"110/01/08","56,491,757","1,461,826,289","25.95","26.15","25.50","26.00","+0.30","17,077",
"110/01/11","29,988,375","773,384,743","26.00","26.05","25.60","25.90","-0.10","10,362",
"110/01/12","45,016,298","1,140,510,344","25.75","25.75","25.15","25.30","-0.60","14,591",
"110/01/13","41,467,607","1,051,919,920","25.30","25.55","25.15","25.55","+0.25","13,719",
"110/01/14","34,269,792","870,144,549","25.70","25.75","25.25","25.30","-0.25","12,922",
"110/01/15","47,128,490","1,178,323,973","25.30","25.40","24.85","24.90","-0.40","14,113",
"110/01/18","36,783,128","893,033,955","24.70","24.70","24.10","24.25","-0.65","14,523",
"110/01/19","22,801,703","555,011,251","24.35","24.55","24.25","24.30","+0.05","7,581",
"110/01/20","54,351,457","1,284,529,255","24.15","24.15","23.35","23.45","-0.85","20,187",
"110/01/21","25,068,520","594,197,034","23.40","23.95","23.40","23.60","+0.15","8,604",
"110/01/22","29,117,020","685,347,016","23.60","23.80","23.20","23.65","+0.05","9,360",
"110/01/25","23,434,064","560,065,107","23.70","24.20","23.40","23.95","+0.30","10,319",
"110/01/26","24,037,073","569,241,502","23.90","23.95","23.55","23.70","-0.25","7,854",
"110/01/27","23,514,869","556,621,150","23.90","24.00","23.55","23.55","-0.15","6,985",
"110/01/28","36,013,867","838,518,475","23.30","23.45","23.15","23.30","-0.25","12,747",
"110/01/29","33,857,935","783,825,131","23.25","23.55","22.95","22.95","-0.35","9,950",
"說明:"
"符號說明:+/-/X表示漲/跌/不比價"
"當日統計資訊含一般、零股、盤後定價、鉅額交易，不含拍賣、標購。"
"ETF證券代號第六碼為K、M、S、C者，表示該ETF以外幣交易。"
```

儲存下載內容：二進位



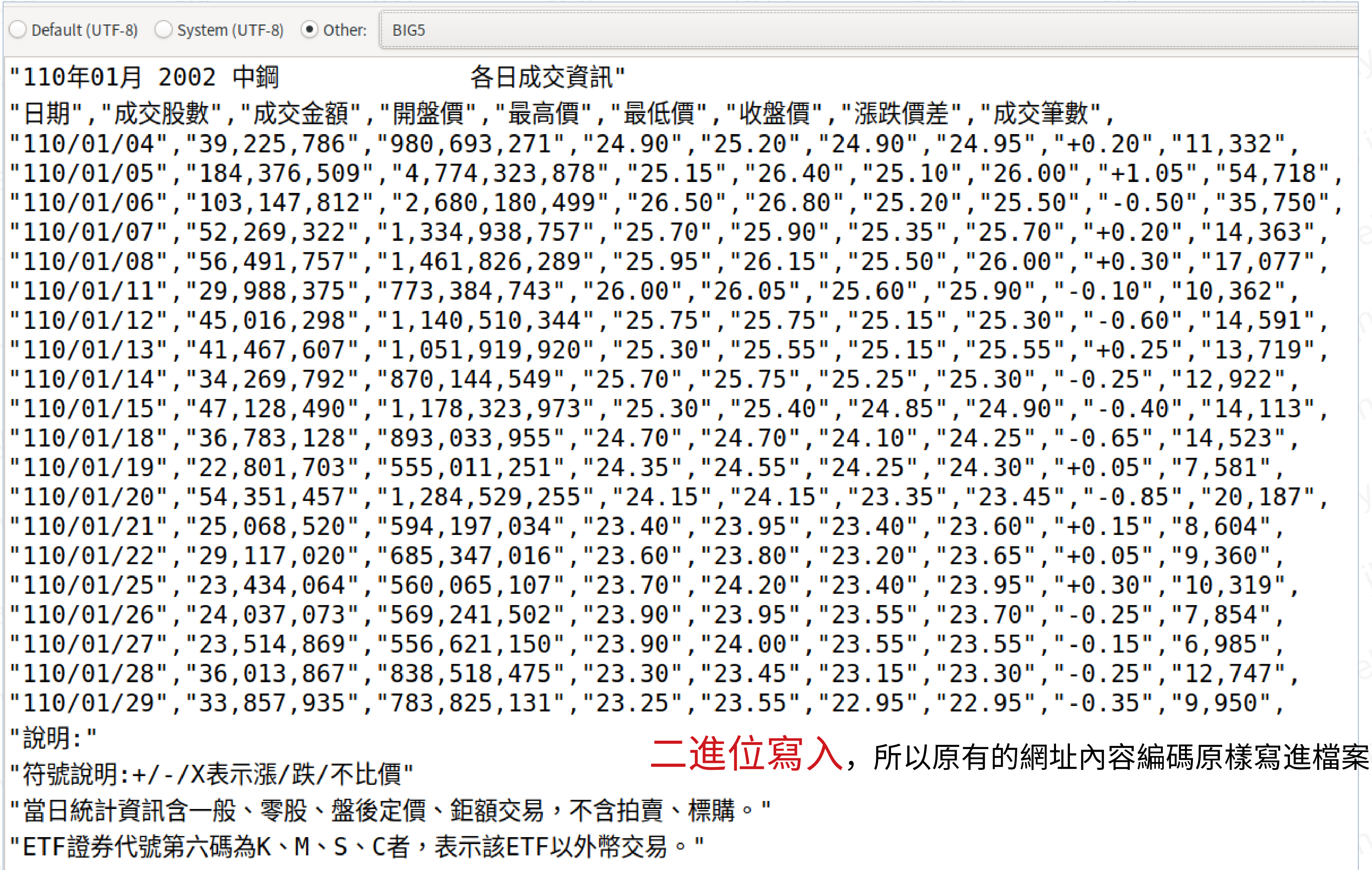
```python
import requests

# 網址字串
urlstr = 'https://www.twse.com.tw/exchangeReport/STOCK_DAY?response=csv&date=20210131&stockNo=2002'
try:
    resp = requests.get(urlstr)
    print('網址內容下載成功')
except Exception as err:
    print('取得網址內容失敗：%s' % err)

# 網址內容存檔
outfname = '2002.202101.bin'
with open(outfname, 'wb') as outf:          # 二進位儲存
    for part in resp.iter_content(512):      # iter_content: 迭代送出取得的內容
        size = outf.write(part)
        print('寫出', size, 'bytes')
    print('儲存完畢：%s' % outfname)
```

儲存下載內容：二進位



```
import requests

# 網址字串
urlstr = 'https://www.twse.com.tw/exchangeReport/STOCK_DAY?response=csv&date=20210131&stockNo=2002'
try:
    resp = requests.get(urlstr)
    print('網址內容下載成功')
except Exception as err:
    print('取得網址內容失敗：%s' % err)

# 網址內容存檔
outfname = '2002.202101.bin'
with open(outfname, 'wb') as outf:          # 二進位儲存
    for part in resp.iter_content(512):      # iter_content:
        size = outf.write(part)
        print('寫出', size, 'bytes')
    print('儲存完畢：%s' % outfname)
```

```
網址內容下載成功
寫出 512 bytes
寫出 512 bytes
寫出 512 bytes
寫出 512 bytes
寫出 70 bytes
儲存完畢：2002.202101.bin
```



○ Default (UTF-8) ○ System (UTF-8) ● Other: BIG5

```
"110年01月 2002 中鋼           各日成交資訊"
"日期","成交股數","成交金額","開盤價","最高價","最低價","收盤價","漲跌價差","成交筆數",
"110/01/04","39,225,786","980,693,271","24.90","25.20","24.90","24.95","+0.20","11,332",
"110/01/05","184,376,509","4,774,323,878","25.15","26.40","25.10","26.00","+1.05","54,718",
"110/01/06","103,147,812","2,680,180,499","26.50","26.80","25.20","25.50","-0.50","35,750",
"110/01/07","52,269,322","1,334,938,757","25.70","25.90","25.35","25.70","+0.20","14,363",
"110/01/08","56,491,757","1,461,826,289","25.95","26.15","25.50","26.00","+0.30","17,077",
"110/01/11","29,988,375","773,384,743","26.00","26.05","25.60","25.90","-0.10","10,362",
"110/01/12","45,016,298","1,140,510,344","25.75","25.75","25.15","25.30","-0.60","14,591",
"110/01/13","41,467,607","1,051,919,920","25.30","25.55","25.15","25.55","+0.25","13,719",
"110/01/14","34,269,792","870,144,549","25.70","25.75","25.25","25.30","-0.25","12,922",
"110/01/15","47,128,490","1,178,323,973","25.30","25.40","24.85","24.90","-0.40","14,113",
"110/01/18","36,783,128","893,033,955","24.70","24.70","24.10","24.25","-0.65","14,523",
"110/01/19","22,801,703","555,011,251","24.35","24.55","24.25","24.30","+0.05","7,581",
"110/01/20","54,351,457","1,284,529,255","24.15","24.15","23.35","23.45","-0.85","20,187",
"110/01/21","25,068,520","594,197,034","23.40","23.95","23.40","23.60","+0.15","8,604",
"110/01/22","29,117,020","685,347,016","23.60","23.80","23.20","23.65","+0.05","9,360",
"110/01/25","23,434,064","560,065,107","23.70","24.20","23.40","23.95","+0.30","10,319",
"110/01/26","24,037,073","569,241,502","23.90","23.95","23.55","23.70","-0.25","7,854",
"110/01/27","23,514,869","556,621,150","23.90","24.00","23.55","23.55","-0.15","6,985",
"110/01/28","36,013,867","838,518,475","23.30","23.45","23.15","23.30","-0.25","12,747",
"110/01/29","33,857,935","783,825,131","23.25","23.55","22.95","22.95","-0.35","9,950",
"說明:"
"符號說明:+/-/X表示漲/跌/不比價"
"當日統計資訊含一般、零股、盤後定價、鉅額交易，不含拍賣、標購。"
"ETF證券代號第六碼為K、M、S、C者，表示該ETF以外幣交易。"
```

二進位寫入，所以原有的網址內容編碼原樣寫進檔案了！

講次內容

- 使用 webbrowser 模組
- 使用 requests 模組
- 使用 urllib 模組
- Cookie？ 餅乾?
- 代理伺服器

urllib 模組

- Python 內建模組
  - 2.x 版分 urllib, urllib2
  - 3.x 版只有 urllib
- 子模組
  - request：透過網址下載內容
  - parse：解析網址
  - error：傳回 request 發生錯誤或異常的原因
  - robotparser：用來解析 robots.txt

urllib 模組

- Python 內建模組
  - 2.x 版分 urllib, urllib2
  - 3.x 版只有 urllib
- 子模組
  - request： 透過網址下載內容
  - parse： 解析網址
  - error： 傳回 request 發生錯誤或異常的原因
  - robotparser： 用來解析 robots.txt

本課程以 requests 模組教學為主，此部份省略

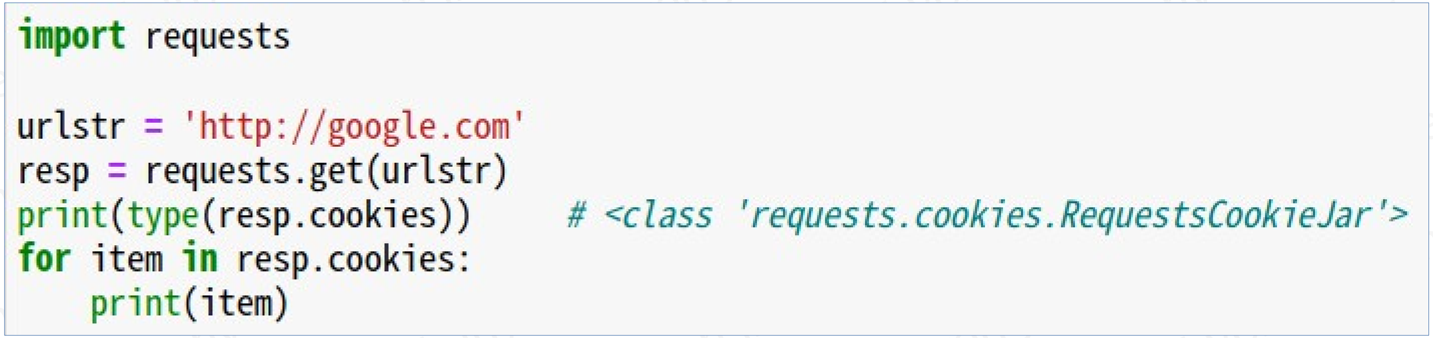
講次內容

- 使用 webbrowser 模組
- 使用 requests 模組
- 使用 urllib 模組
- Cookie？ 餅乾?
- 代理伺服器

Cookie

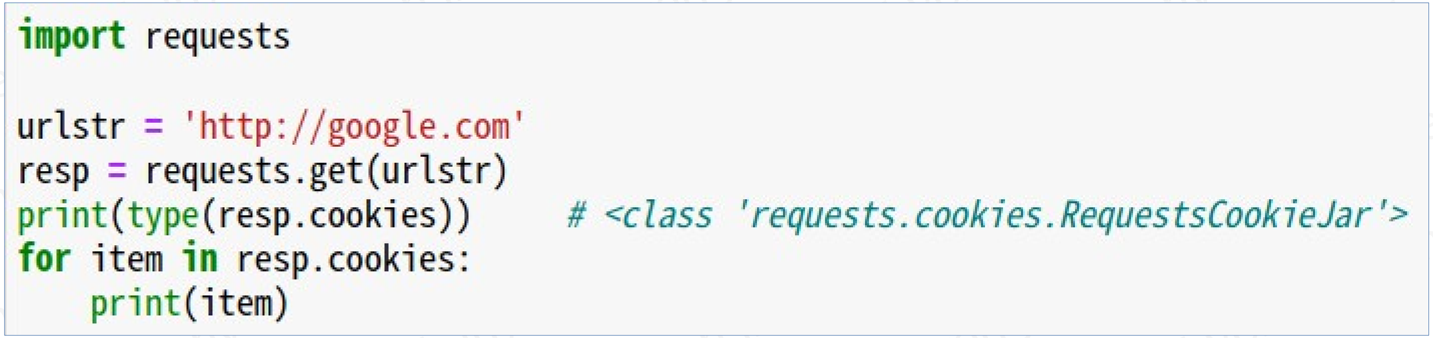
- 網站辨識使用者身份所設定的資訊
- 鍵值 - 資料值配對儲存
- 儲存在使用者電腦中

看一下 Google 的 Cookie

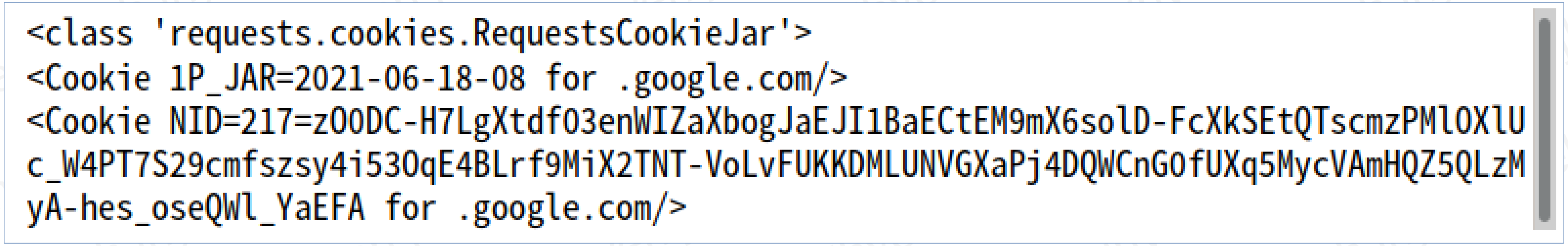


```python
import requests

urlstr = 'http://google.com'
resp = requests.get(urlstr)
print(type(resp.cookies))     # <class 'requests.cookies.RequestsCookieJar'>
for item in resp.cookies:
    print(item)
```

看一下 Google 的 Cookie



```
import requests

urlstr = 'http://google.com'
resp = requests.get(urlstr)
print(type(resp.cookies))     # <class 'requests.cookies.RequestsCookieJar'>
for item in resp.cookies:
    print(item)
```

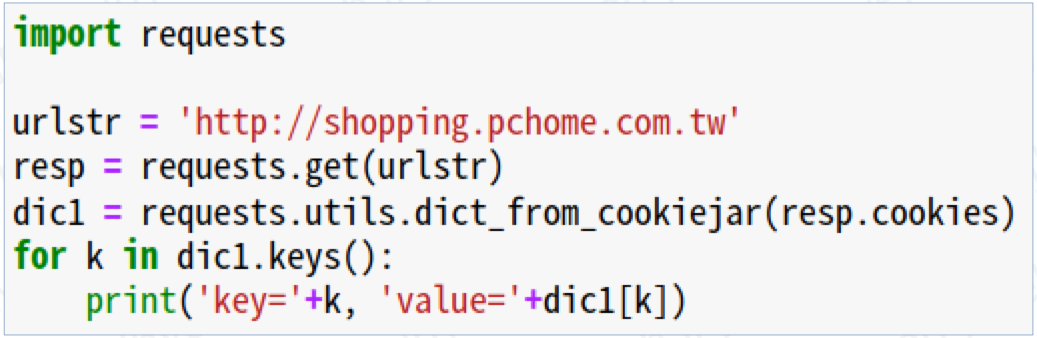


```
<class 'requests.cookies.RequestsCookieJar'>
<Cookie 1P_JAR=2021-06-18-08 for .google.com/>
<Cookie NID=217=zO0DC-H7LgXtdf03enWIZaXbogJaEJI1BaECtEM9mX6solD-FcXkSEtQTscmzPMlOXlU
c_W4PT7S29cmfszsy4i53OqE4BLrf9MiX2TNT-VoLvFUKKDMLUNVGXaPj4DQWCnGOfUXq5MycVAmHQZ5QLzM
yA-hes_oseQWl_YaEFA for .google.com/>
```

不好懂 ...

看一下 PCHome 的 Cookie

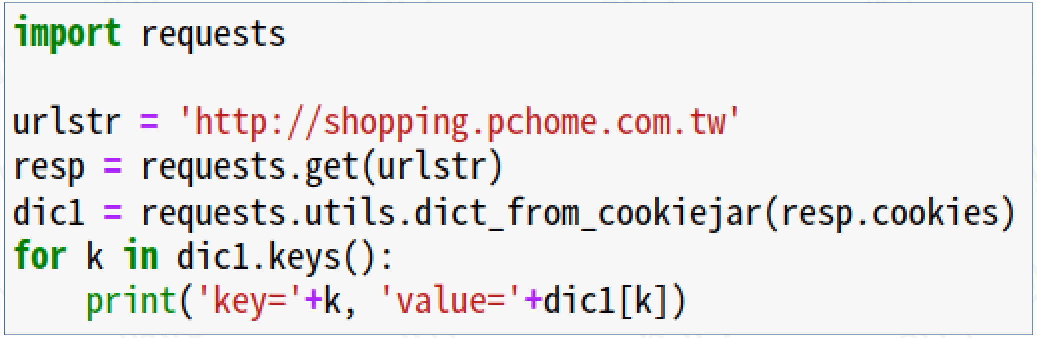
- 這次我們想辦法用 dict 來取出 cookie
  - 使用 dict_from_cookiejar() 方法



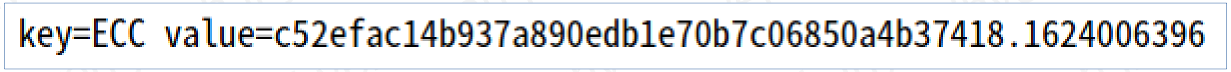
```python
import requests

urlstr = 'http://shopping.pchome.com.tw'
resp = requests.get(urlstr)
dic1 = requests.utils.dict_from_cookiejar(resp.cookies)
for k in dic1.keys():
    print('key='+k, 'value='+dic1[k])
```

看一下 PCHome 的 Cookie

- 這次我們想辦法用 dict 來取出 cookie
  - 使用 dict_from_cookiejar() 方法



```
import requests

urlstr = 'http://shopping.pchome.com.tw'
resp = requests.get(urlstr)
dic1 = requests.utils.dict_from_cookiejar(resp.cookies)
for k in dic1.keys():
    print('key='+k, 'value='+dic1[k])
```



```
key=ECC value=c52efac14b937a890edb1e70b7c06850a4b37418.1624006396
```

原來可以轉 dict 結構！！

講次內容

- 使用 webbrowser 模組
- 使用 requests 模組
- 使用 urllib 模組
- Cookie？ 餅乾?
- 代理伺服器

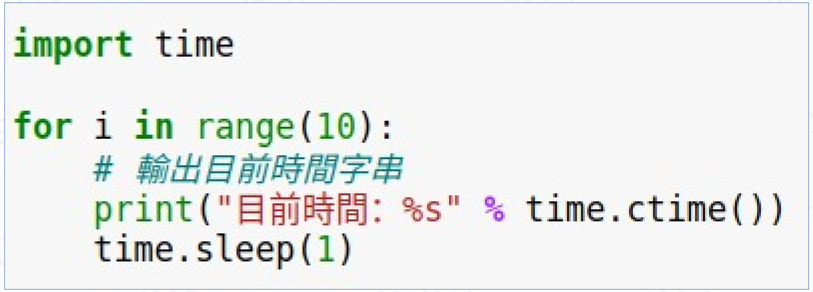
還記得爬蟲為什麼會被擋嗎?

- 網頁伺服器採取反爬蟲動作的理由
  - 基於安全理由，拒絕頻繁存取
  - 不想增加伺服器的網路流量（負擔）
  - 希望有規範的存取 (robots.txt)

避免頻繁存取

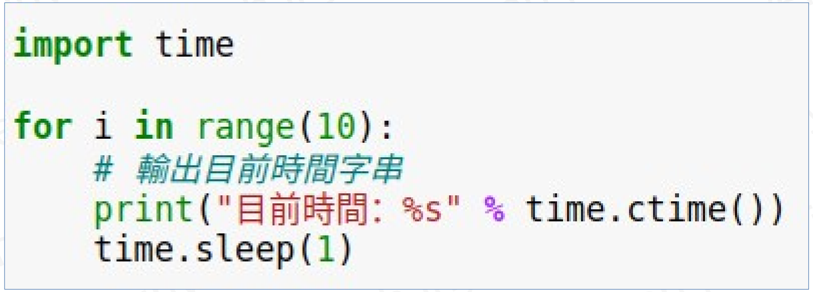
- Be nice to others!
- 大量存取之間，「睡」一下
- import time
- 使用 time.sleep() 方法
  - 參數：睡覺秒數

sleep() 方法測試

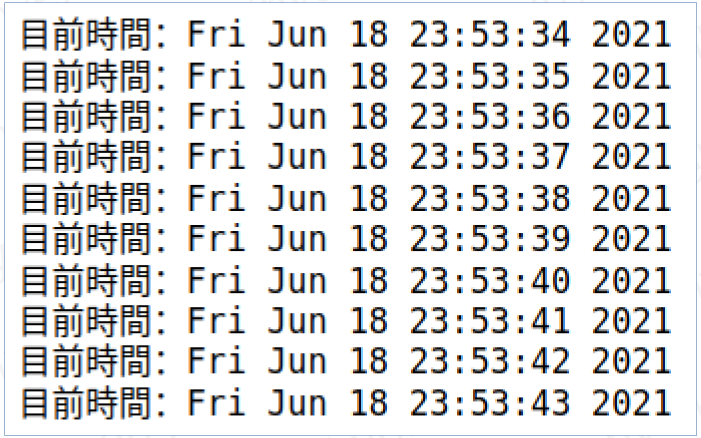


```
import time

for i in range(10):
    # 輸出目前時間字串
    print("目前時間: %s" % time.ctime())
    time.sleep(1)
```

sleep() 方法測試



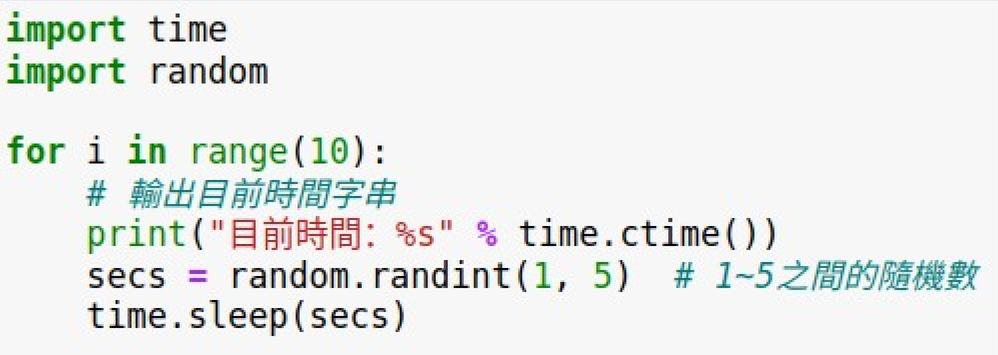
```python
import time

for i in range(10):
    # 輸出目前時間字串
    print("目前時間: %s" % time.ctime())
    time.sleep(1)
```



```
目前時間: Fri Jun 18 23:53:34 2021
目前時間: Fri Jun 18 23:53:35 2021
目前時間: Fri Jun 18 23:53:36 2021
目前時間: Fri Jun 18 23:53:37 2021
目前時間: Fri Jun 18 23:53:38 2021
目前時間: Fri Jun 18 23:53:39 2021
目前時間: Fri Jun 18 23:53:40 2021
目前時間: Fri Jun 18 23:53:41 2021
目前時間: Fri Jun 18 23:53:42 2021
目前時間: Fri Jun 18 23:53:43 2021
```

隨機延遲

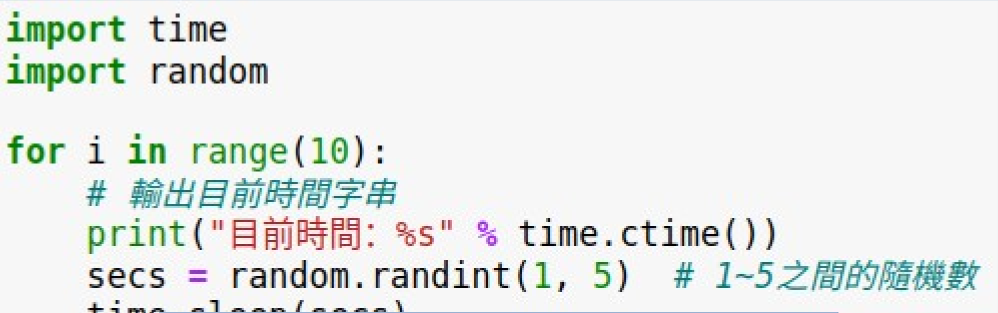
- 如果想隨機睡幾秒，看起來更「自然」一些呢?



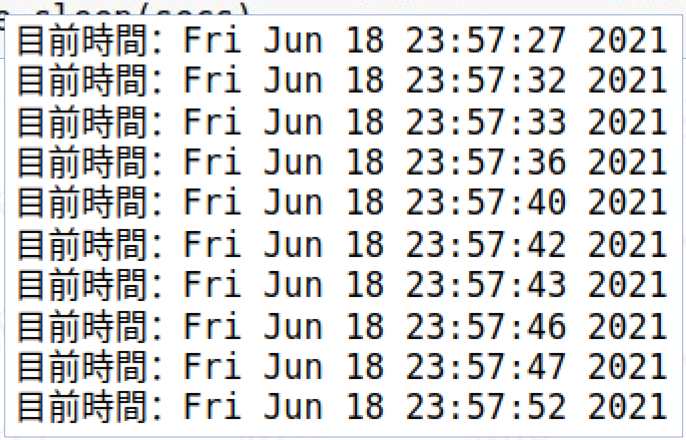
```
import time
import random

for i in range(10):
    # 輸出目前時間字串
    print("目前時間: %s" % time.ctime())
    secs = random.randint(1, 5)  # 1~5之間的隨機數
    time.sleep(secs)
```

隨機延遲

- 如果想隨機睡幾秒，看起來更「自然」一些呢?



```
import time
import random

for i in range(10):
    # 輸出目前時間字串
    print("目前時間: %s" % time.ctime())
    secs = random.randint(1, 5)  # 1~5之間的隨機數
    time.sleep(secs)
```



```
目前時間: Fri Jun 18 23:57:27 2021
目前時間: Fri Jun 18 23:57:32 2021
目前時間: Fri Jun 18 23:57:33 2021
目前時間: Fri Jun 18 23:57:36 2021
目前時間: Fri Jun 18 23:57:40 2021
目前時間: Fri Jun 18 23:57:42 2021
目前時間: Fri Jun 18 23:57:43 2021
目前時間: Fri Jun 18 23:57:46 2021
目前時間: Fri Jun 18 23:57:47 2021
目前時間: Fri Jun 18 23:57:52 2021
```

如果睡了還是避不了?

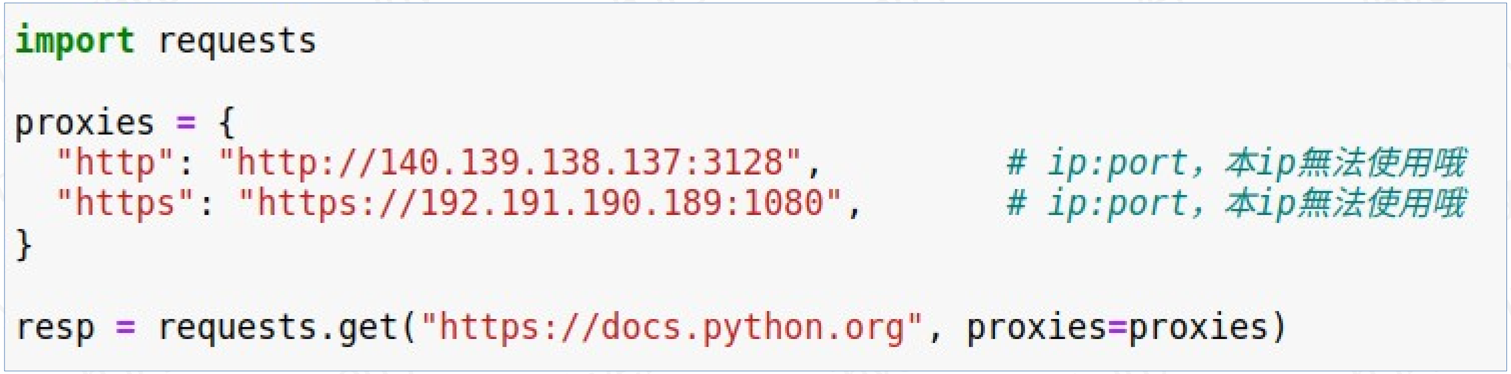
- 避不了「頻繁」這件事!
- 透過代理伺服器
  - 代理伺服器是幫忙做請求轉送
  - 可以隱藏自己的電腦 IP 位址

使用代理伺服器

- requests.get() 可增加 proxies 參數
  - 指定代理伺服器
  - 辭典型別的 proxy 伺服器資訊

使用代理伺服器

- requests.get() 可增加 proxies 參數
  - 指定代理伺服器
  - 辭典型別的 proxy 伺服器資訊

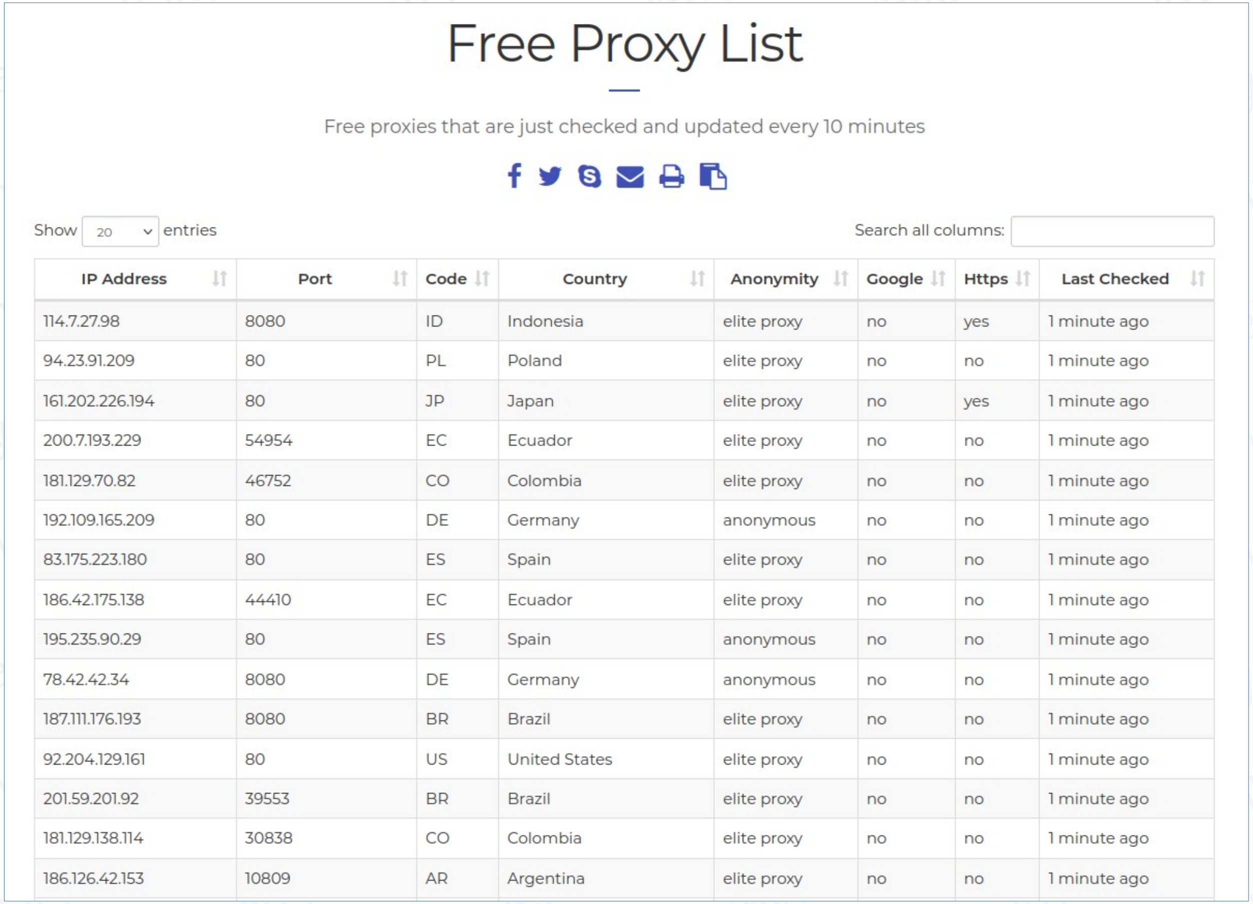


```
import requests

proxies = {
  "http": "http://140.139.138.137:3128",         # ip:port, 本ip無法使用哦
  "https": "https://192.191.190.189:1080",       # ip:port, 本ip無法使用哦
}

resp = requests.get("https://docs.python.org", proxies=proxies)
```

Hint: Free Proxy List

- FYI…
- 好用嗎?

  自己用用看 …



Free Proxy List

Free proxies that are just checked and updated every 10 minutes

IP Address	Port	Code	Country	Anonymity	Google	Https	Last Checked
114.7.27.98	8080	ID	Indonesia	elite proxy	no	yes	1 minute ago
94.23.91.209	80	PL	Poland	elite proxy	no	no	1 minute ago
161.202.226.194	80	JP	Japan	elite proxy	no	yes	1 minute ago
200.7.193.229	54954	EC	Ecuador	elite proxy	no	no	1 minute ago
181.129.70.82	46752	CO	Colombia	elite proxy	no	no	1 minute ago
192.109.165.209	80	DE	Germany	anonymous	no	no	1 minute ago
83.175.223.180	80	ES	Spain	elite proxy	no	no	1 minute ago
186.42.175.138	44410	EC	Ecuador	elite proxy	no	no	1 minute ago
195.235.90.29	80	ES	Spain	anonymous	no	no	1 minute ago
78.42.42.34	8080	DE	Germany	anonymous	no	no	1 minute ago
187.111.176.193	8080	BR	Brazil	elite proxy	no	no	1 minute ago
92.204.129.161	80	US	United States	elite proxy	no	no	1 minute ago
201.59.201.92	39553	BR	Brazil	elite proxy	no	no	1 minute ago
181.129.138.114	30838	CO	Colombia	elite proxy	no	no	1 minute ago
186.126.42.153	10809	AR	Argentina	elite proxy	no	no	1 minute ago

這個講次中，你應該學到了 ...

- 如何使用 webbrowser 模組開啟網頁
- 如何使用 requests 模組製作爬蟲
- Cookie 的原理與資料格式
- 代理伺服器！ （噓 ... ）