# 成為初級資料分析師 | Python 與資料科學應用

*網頁資料擷取*

**郭耀仁**

# 大綱

- 網頁資料擷取的核心任務
- 擷取 JSON 格式網頁資料
- 擷取 XML 格式網頁資料
- 擷取 HTML 格式網頁資料
- 瀏覽器自動化

# 網頁資料擷取的核心任務

# 盤點核心任務

以 Python 豐富的套件、Chrome 瀏覽器外掛與開發者工具來進行兩項核心任務：

1. 請求資料 Requesting Data
2. 解析資料 Parsing Data

# HTTP

> *超文本傳輸協定(HTTP) 是一種用來傳輸超媒體文件(像是HTML文件) 的應用層協定，被設計來讓瀏覽器和伺服器進行溝通，但也可做其他用途。HTTP 遵循標準客戶端–伺服器模式，由客戶端連線以發送請求，然後等待接收回應。*

Source: https://developer.mozilla.org/zh-TW/docs/Web/HTTP (https://developer.mozilla.org/zh-TW/docs/Web/HTTP)

## HTTP 定義了一組能令給定資源，執行特定操作的請求方法（request methods），其中與網頁資料擷取的請求資料最相關的是：

- GET
- POST

## 請求資料是雙向的

- 由瀏覽器發給網頁伺服器的請求稱為 HTTP Request，HTTP Request 包含 Header 和 Body
- 由網頁伺服器回應瀏覽器的請求稱為 HTTP Response，HTTP Response 也包含 Header 和 Body
- Request Header 中的 Request Method 表示瀏覽器希望網頁伺服器做些什麼事
- Response Header 中的 Status Code 表示網頁伺服器告訴瀏覽器事情辦好沒

## 請求資料需要使用的工具

- Chrome 開發者工具
- Quick JavaScript Switcher (https://chrome.google.com/webstore/detail/quick-javascript-switcher/geddoclleiomckbhadiaipdggiiccfje)
- `requests` 套件

## Chrome 開發者工具

> *Chrome 開發者工具是一套內建於 Google Chrome 中的 Web 開發和測試工具。*

Source: https://developers.google.com/web/tools/chrome-devtools/?hl=zh-TW (https://developers.google.com/web/tools/chrome-devtools/?hl=zh-TW)

Imgur

# 進行網頁資料擷取時，會高度仰賴 Chrome 開發者工具中的 Network 頁籤

使用 Network 頁籤瞭解請求和下載的檔案

# 點選 Network 之後重新整理網頁觀察

Imgur

# 通常我們需要擷取的資料會被歸類在這兩個大類檔案中

- XHR(XMLHttpRequest)
- Doc

# 可以使用 Quick JavaScript Switcher 協助判斷

*快速地開啟、關閉 JavaScript*

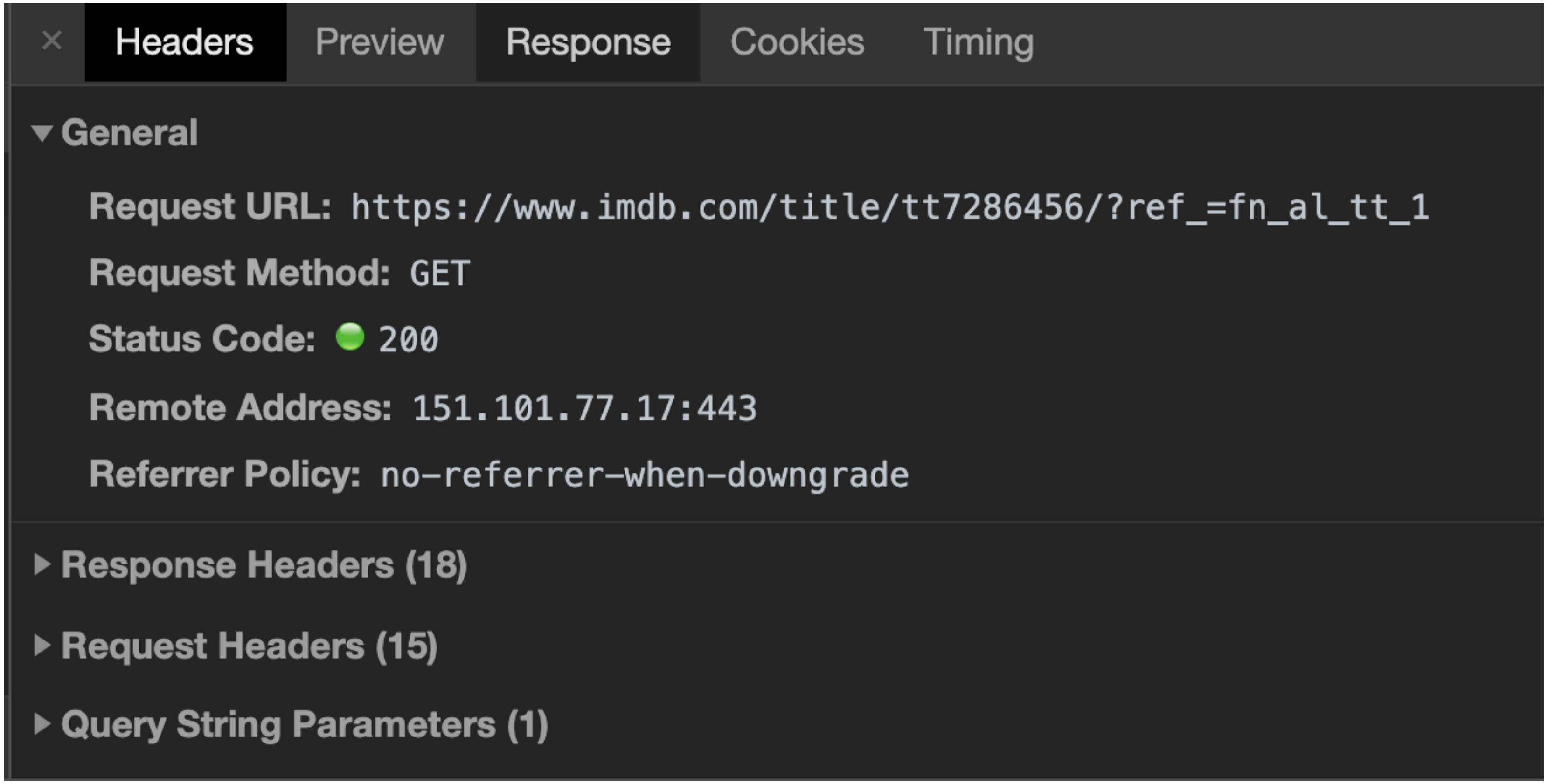
Source: https://chrome.google.com/webstore/detail/quick-javascript-switcher/geddoclleiomckbhadiaipdggiiccfje (https://chrome.google.com/webstore/detail/quick-javascript-switcher/geddoclleiomckbhadiaipdggiiccfje)

## 示範 Quick JavaScript Switcher 功能

- https://www.imdb.com/title/tt7286456/ (https://www.imdb.com/title/tt7286456/)
- https://ecshweb.pchome.com.tw/search/v3.3/ (https://ecshweb.pchome.com.tw/search/v3.3/)

# 找到資料之後即可檢視細節

- Headers
    - General
    - Response Headers
    - Request Headers
    - Query String Parameters(if any)
    - Form Data(if any, for POST)
- Preview
- Response
- Cookies



▼ General

Request URL: https://www.imdb.com/title/tt7286456/?ref_=fn_al_tt_1

Request Method: GET

Status Code: ● 200

Remote Address: 151.101.77.17:443

Referrer Policy: no-referrer-when-downgrade

▶ Response Headers (18)

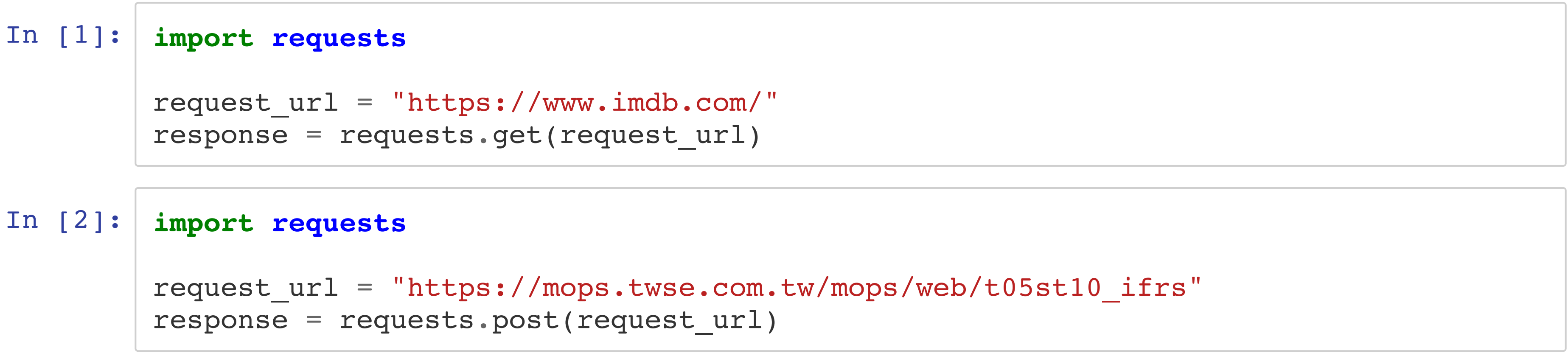
▶ Request Headers (15)

▶ Query String Parameters (1)

Imgur

## 常用的 `requests` 函數

- `requests.get()`：進行 GET 請求（下載檔案）
- `requests.post()`：進行 POST 請求（上傳資料）



In [1]:

```
import requests

request_url = "https://www.imdb.com/"
response = requests.get(request_url)
```

In [2]:

```
import requests

request_url = "https://mops.twse.com.tw/mops/web/t05st10_ifrs"
response = requests.post(request_url)
```

## 回應（Response 類別）的方法與屬性

- `response.status_code`：查看狀態碼
- `response.json()`：將回應直接轉換為 Python 的資料結構（`list` 或 `dict`）
- `response.content`：將回應轉換為 `bytes`
- `response.text`：將回應轉換為 `str`

## 檢視資料細節的 Preview 與 Response確認格式

- 如果資料是 JSON 格式：呼叫回應的 `.json()` 方法後直接以 Python 資料結構解析
- 如果資料是 XML 格式：呼叫回應的 `.content` 屬性後以 `lxml` 搭配 XPath 解析
- 如果資料是 HTML 格式：呼叫回應的 `.text` 屬性後以 `bs4` 搭配 CSS Selector 解析

# 擷取 JSON 格式網頁資料

# JSON 格式網頁資料範例

- data.nba (http://data.nba.net/prod/v1/today.json)
- PChome (https://ecshweb.pchome.com.tw/search/v3.3/all/results?q=macbook&page=1&sort=sale/dc)

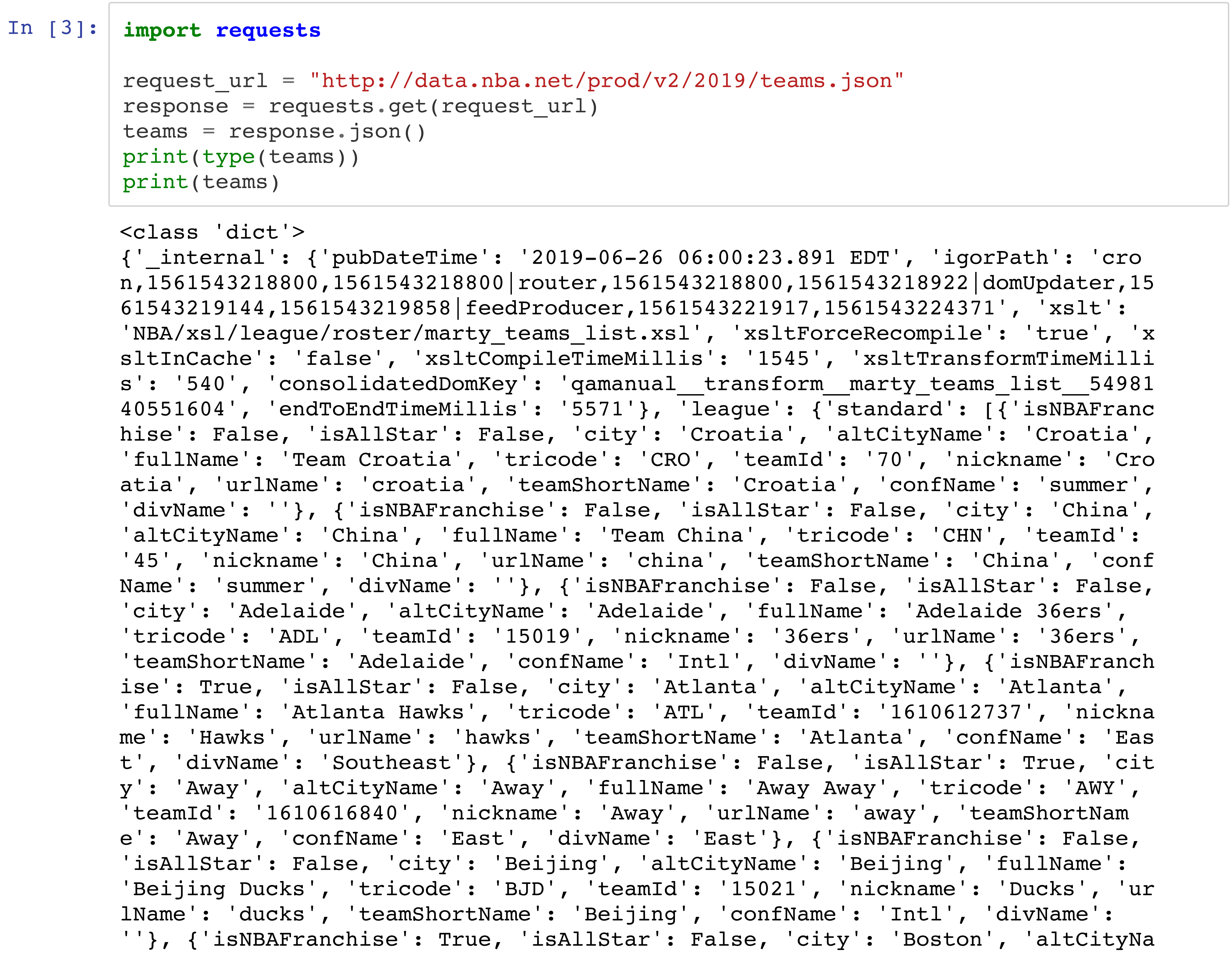
## 幫助瀏覽 JSON 資料的 Chrome 外掛

JSON View
(https://chrome.google.com/webstore/detail/jsonview/chklaanhfefbnpoihckbnefhakgolnmc)

## 擷取 JSON 格式網頁資料步驟

- 使用 `requests` 請求資料
- 呼叫回應的 `.json()` 方法，例如 `response.json()`
- 視需求進行摘要

以 http://data.nba.net/prod/v2/2019/teams.json (http://data.nba.net/prod/v2/2019/teams.json) 示範



```
In [3]: import requests

        request_url = "http://data.nba.net/prod/v2/2019/teams.json"
        response = requests.get(request_url)
        teams = response.json()
        print(type(teams))
        print(teams)
```

```
<class 'dict'>
{'_internal': {'pubDateTime': '2019-06-26 06:00:23.891 EDT', 'igorPath': 'cro
n,1561543218800,1561543218800|router,1561543218800,1561543218922|domUpdater,15
61543219144,1561543219858|feedProducer,1561543221917,1561543224371', 'xslt':
'NBA/xsl/league/roster/marty_teams_list.xsl', 'xsltForceRecompile': 'true', 'x
sltInCache': 'false', 'xsltCompileTimeMillis': '1545', 'xsltTransformTimeMilli
s': '540', 'consolidatedDomKey': 'qamanual__transform__marty_teams_list__54981
40551604', 'endToEndTimeMillis': '5571'}, 'league': {'standard': [{'isNBAFranc
hise': False, 'isAllStar': False, 'city': 'Croatia', 'altCityName': 'Croatia',
'fullName': 'Team Croatia', 'tricode': 'CRO', 'teamId': '70', 'nickname': 'Cro
atia', 'urlName': 'croatia', 'teamShortName': 'Croatia', 'confName': 'summer',
'divName': ''}, {'isNBAFranchise': False, 'isAllStar': False, 'city': 'China',
'altCityName': 'China', 'fullName': 'Team China', 'tricode': 'CHN', 'teamId':
'45', 'nickname': 'China', 'urlName': 'china', 'teamShortName': 'China', 'conf
Name': 'summer', 'divName': ''}, {'isNBAFranchise': False, 'isAllStar': False,
'city': 'Adelaide', 'altCityName': 'Adelaide', 'fullName': 'Adelaide 36ers',
'tricode': 'ADL', 'teamId': '15019', 'nickname': '36ers', 'urlName': '36ers',
'teamShortName': 'Adelaide', 'confName': 'Intl', 'divName': ''}, {'isNBAFranch
ise': True, 'isAllStar': False, 'city': 'Atlanta', 'altCityName': 'Atlanta',
'fullName': 'Atlanta Hawks', 'tricode': 'ATL', 'teamId': '1610612737', 'nickna
me': 'Hawks', 'urlName': 'hawks', 'teamShortName': 'Atlanta', 'confName': 'Eas
t', 'divName': 'Southeast'}, {'isNBAFranchise': False, 'isAllStar': True, 'cit
y': 'Away', 'altCityName': 'Away', 'fullName': 'Away Away', 'tricode': 'AWY',
'teamId': '1610616840', 'nickname': 'Away', 'urlName': 'away', 'teamShortNam
e': 'Away', 'confName': 'East', 'divName': 'East'}, {'isNBAFranchise': False,
'isAllStar': False, 'city': 'Beijing', 'altCityName': 'Beijing', 'fullName':
'Beijing Ducks', 'tricode': 'BJD', 'teamId': '15021', 'nickname': 'Ducks', 'ur
lName': 'ducks', 'teamShortName': 'Beijing', 'confName': 'Intl', 'divName':
''}, {'isNBAFranchise': True, 'isAllStar': False, 'city': 'Boston', 'altCityNa
```

```
me': 'Boston', 'fullName': 'Boston Celtics', 'tricode': 'BOS', 'teamId': '1610
612738', 'nickname': 'Celtics', 'urlName': 'celtics', 'teamShortName': 'Bosto
n', 'confName': 'East', 'divName': 'Atlantic'}, {'isNBAFranchise': True, 'isAl
lStar': False, 'city': 'Brooklyn', 'altCityName': 'Brooklyn', 'fullName': 'Bro
oklyn Nets', 'tricode': 'BKN', 'teamId': '1610612751', 'nickname': 'Nets', 'ur
lName': 'nets', 'teamShortName': 'Brooklyn', 'confName': 'East', 'divName': 'A
tlantic'}, {'isNBAFranchise': True, 'isAllStar': False, 'city': 'Charlotte',
 'altCityName': 'Charlotte', 'fullName': 'Charlotte Hornets', 'tricode': 'CHA',
 'teamId': '1610612766', 'nickname': 'Hornets', 'urlName': 'hornets', 'teamShor
tName': 'Charlotte', 'confName': 'East', 'divName': 'Southeast'}, {'isNBAFranc
hise': False, 'isAllStar': False, 'city': 'Buenos Aires', 'altCityName': 'Buen
os Aires', 'fullName': 'San Lorenzo de Almagro', 'tricode': 'SLA', 'teamId':
 '12330', 'nickname': 'San Lorenzo', 'urlName': 'san_lorenzo', 'teamShortName':
 'San Lorenzo', 'confName': 'Intl', 'divName': ''}, {'isNBAFranchise': True, 'i
sAllStar': False, 'city': 'Chicago', 'altCityName': 'Chicago', 'fullName': 'Ch
icago Bulls', 'tricode': 'CHI', 'teamId': '1610612741', 'nickname': 'Bulls',
 'urlName': 'bulls', 'teamShortName': 'Chicago', 'confName': 'East', 'divName':
 'Central'}, {'isNBAFranchise': True, 'isAllStar': False, 'city': 'Cleveland',
 'altCityName': 'Cleveland', 'fullName': 'Cleveland Cavaliers', 'tricode': 'CL
E', 'teamId': '1610612739', 'nickname': 'Cavaliers', 'urlName': 'cavaliers',
 'teamShortName': 'Cleveland', 'confName': 'East', 'divName': 'Central'}, {'isN
BAFranchise': True, 'isAllStar': False, 'city': 'Dallas', 'altCityName': 'Dall
as', 'fullName': 'Dallas Mavericks', 'tricode': 'DAL', 'teamId': '1610612742',
 'nickname': 'Mavericks', 'urlName': 'mavericks', 'teamShortName': 'Dallas', 'c
onfName': 'West', 'divName': 'Southwest'}, {'isNBAFranchise': True, 'isAllSta
r': False, 'city': 'Denver', 'altCityName': 'Denver', 'fullName': 'Denver Nugg
ets', 'tricode': 'DEN', 'teamId': '1610612743', 'nickname': 'Nuggets', 'urlNam
e': 'nuggets', 'teamShortName': 'Denver', 'confName': 'West', 'divName': 'Nort
hwest'}, {'isNBAFranchise': True, 'isAllStar': False, 'city': 'Detroit', 'altC
ityName': 'Detroit', 'fullName': 'Detroit Pistons', 'tricode': 'DET', 'teamI
d': '1610612765', 'nickname': 'Pistons', 'urlName': 'pistons', 'teamShortNam
e': 'Detroit', 'confName': 'East', 'divName': 'Central'}, {'isNBAFranchise': F
alse, 'isAllStar': False, 'city': 'Franca', 'altCityName': 'Franca', 'fullNam
e': 'SESI/Franca', 'tricode': 'FRA', 'teamId': '12332', 'nickname': 'Franca',
 'urlName': 'franca', 'teamShortName': 'Franca', 'confName': 'Intl', 'divName':
 ''}, {'isNBAFranchise': True, 'isAllStar': False, 'city': 'Golden State', 'alt
CityName': 'Golden State', 'fullName': 'Golden State Warriors', 'tricode': 'GS
W', 'teamId': '1610612744', 'nickname': 'Warriors', 'urlName': 'warriors', 'te
```

```
amShortName': 'Golden State', 'confName': 'West', 'divName': 'Pacific'}, {'isN
BAFranchise': False, 'isAllStar': False, 'city': 'Guangzhou', 'altCityName':
'Guangzhou', 'fullName': 'Guangzhou Long-Lions', 'tricode': 'GUA', 'teamId':
'15018', 'nickname': 'Long-Lions', 'urlName': 'long-lions', 'teamShortName':
'Guangzhou', 'confName': 'Intl', 'divName': ''}, {'isNBAFranchise': False, 'is
AllStar': False, 'city': 'Haifa', 'altCityName': 'Haifa', 'fullName': 'Haifa M
accabi Haifa', 'tricode': 'MAC', 'teamId': '93', 'nickname': 'Maccabi Haifa',
'urlName': 'maccabi_haifa', 'teamShortName': 'Maccabi Haifa', 'confName': 'Int
l', 'divName': ''}, {'isNBAFranchise': False, 'isAllStar': True, 'city': 'Hom
e', 'altCityName': 'Home', 'fullName': 'Home Home', 'tricode': 'HME', 'teamI
d': '1610616839', 'nickname': 'Home', 'urlName': 'home', 'teamShortName': 'Hom
e', 'confName': 'East', 'divName': 'East'}, {'isNBAFranchise': True, 'isAllSta
r': False, 'city': 'Houston', 'altCityName': 'Houston', 'fullName': 'Houston R
ockets', 'tricode': 'HOU', 'teamId': '1610612745', 'nickname': 'Rockets', 'url
Name': 'rockets', 'teamShortName': 'Houston', 'confName': 'West', 'divName':
'Southwest'}, {'isNBAFranchise': True, 'isAllStar': False, 'city': 'Indiana',
'altCityName': 'Indiana', 'fullName': 'Indiana Pacers', 'tricode': 'IND', 'tea
mId': '1610612754', 'nickname': 'Pacers', 'urlName': 'pacers', 'teamShortNam
e': 'Indiana', 'confName': 'East', 'divName': 'Central'}, {'isNBAFranchise': T
rue, 'isAllStar': False, 'city': 'LA', 'altCityName': 'LA Clippers', 'fullNam
e': 'LA Clippers', 'tricode': 'LAC', 'teamId': '1610612746', 'nickname': 'Clip
pers', 'urlName': 'clippers', 'teamShortName': 'LA Clippers', 'confName': 'Wes
t', 'divName': 'Pacific'}, {'isNBAFranchise': True, 'isAllStar': False, 'cit
y': 'Los Angeles', 'altCityName': 'Los Angeles Lakers', 'fullName': 'Los Angel
es Lakers', 'tricode': 'LAL', 'teamId': '1610612747', 'nickname': 'Lakers', 'u
rlName': 'lakers', 'teamShortName': 'L.A. Lakers', 'confName': 'West', 'divNam
e': 'Pacific'}, {'isNBAFranchise': False, 'isAllStar': False, 'city': 'Melbour
ne', 'altCityName': 'Melbourne', 'fullName': 'Melbourne United', 'tricode': 'M
EL', 'teamId': '15016', 'nickname': 'United', 'urlName': 'united', 'teamShortN
ame': 'Melbourne', 'confName': 'Intl', 'divName': ''}, {'isNBAFranchise': Tru
e, 'isAllStar': False, 'city': 'Memphis', 'altCityName': 'Memphis', 'fullNam
e': 'Memphis Grizzlies', 'tricode': 'MEM', 'teamId': '1610612763', 'nickname':
'Grizzlies', 'urlName': 'grizzlies', 'teamShortName': 'Memphis', 'confName':
'West', 'divName': 'Southwest'}, {'isNBAFranchise': True, 'isAllStar': False,
'city': 'Miami', 'altCityName': 'Miami', 'fullName': 'Miami Heat', 'tricode':
'MIA', 'teamId': '1610612748', 'nickname': 'Heat', 'urlName': 'heat', 'teamSho
rtName': 'Miami', 'confName': 'East', 'divName': 'Southeast'}, {'isNBAFranchis
e': True, 'isAllStar': False, 'city': 'Milwaukee', 'altCityName': 'Milwaukee',
'fullName': 'Milwaukee Bucks', 'tricode': 'MIL', 'teamId': '1610612749', 'nick
name': 'Bucks', 'urlName': 'bucks', 'teamShortName': 'Milwaukee', 'confName':
'East', 'divName': 'Central'}, {'isNBAFranchise': True, 'isAllStar': False, 'c
ity': 'Minnesota', 'altCityName': 'Minnesota', 'fullName': 'Minnesota Timberwo
lves', 'tricode': 'MIN', 'teamId': '1610612750', 'nickname': 'Timberwolves',
'urlName': 'timberwolves', 'teamShortName': 'Minnesota', 'confName': 'West',
'divName': 'Northwest'}, {'isNBAFranchise': True, 'isAllStar': False, 'city':
'New Orleans', 'altCityName': 'New Orleans', 'fullName': 'New Orleans Pelican
s', 'tricode': 'NOP', 'teamId': '1610612740', 'nickname': 'Pelicans', 'urlNam
e': 'pelicans', 'teamShortName': 'New Orleans', 'confName': 'West', 'divName':
'Southwest'}, {'isNBAFranchise': True, 'isAllStar': False, 'city': 'New York',
'altCityName': 'New York', 'fullName': 'New York Knicks', 'tricode': 'NYK', 't
eamId': '1610612752', 'nickname': 'Knicks', 'urlName': 'knicks', 'teamShortNam
e': 'New York', 'confName': 'East', 'divName': 'Atlantic'}, {'isNBAFranchise':
False, 'isAllStar': False, 'city': 'New Zealand', 'altCityName': 'New Zealan
d', 'fullName': 'New Zealand Breakers', 'tricode': 'NZB', 'teamId': '15020',
'nickname': 'Breakers', 'urlName': 'breakers', 'teamShortName': 'New Zealand',
'confName': 'Intl', 'divName': ''}, {'isNBAFranchise': True, 'isAllStar': Fals
e, 'city': 'Oklahoma City', 'altCityName': 'Oklahoma City', 'fullName': 'Oklah
oma City Thunder', 'tricode': 'OKC', 'teamId': '1610612760', 'nickname': 'Thun
der', 'urlName': 'thunder', 'teamShortName': 'Oklahoma City', 'confName': 'Wes
t', 'divName': 'Northwest'}, {'isNBAFranchise': True, 'isAllStar': False, 'cit
y': 'Orlando', 'altCityName': 'Orlando', 'fullName': 'Orlando Magic', 'tricod
e': 'ORL', 'teamId': '1610612753', 'nickname': 'Magic', 'urlName': 'magic', 't
eamShortName': 'Orlando', 'confName': 'East', 'divName': 'Southeast'}, {'isNBA
Franchise': False, 'isAllStar': False, 'city': 'Perth', 'altCityName': 'Pert
h', 'fullName': 'Perth Wildcats', 'tricode': 'PER', 'teamId': '104', 'nicknam
e': 'Wildcats', 'urlName': 'wildcats', 'teamShortName': 'Perth', 'confName':
'Intl', 'divName': ''}, {'isNBAFranchise': True, 'isAllStar': False, 'city':
'Philadelphia', 'altCityName': 'Philadelphia', 'fullName': 'Philadelphia 76er
s', 'tricode': 'PHI', 'teamId': '1610612755', 'nickname': '76ers', 'urlName':
'sixers', 'teamShortName': 'Philadelphia', 'confName': 'East', 'divName': 'Atl
antic'}, {'isNBAFranchise': True, 'isAllStar': False, 'city': 'Phoenix', 'altC
ityName': 'Phoenix', 'fullName': 'Phoenix Suns', 'tricode': 'PHX', 'teamId':
'1610612756', 'nickname': 'Suns', 'urlName': 'suns', 'teamShortName': 'Phoeni
x', 'confName': 'West', 'divName': 'Pacific'}, {'isNBAFranchise': True, 'isAll
Star': False, 'city': 'Portland', 'altCityName': 'Portland', 'fullName': 'Port
land Trail Blazers', 'tricode': 'POR', 'teamId': '1610612757', 'nickname': 'Tr
```

```
ail Blazers', 'urlName': 'blazers', 'teamShortName': 'Portland', 'confName':
'West', 'divName': 'Northwest'}, {'isNBAFranchise': False, 'isAllStar': False,
'city': 'Rio de Janeiro', 'altCityName': 'Rio de Janeiro', 'fullName': 'Rio de
Janeiro Flamengo', 'tricode': 'FLA', 'teamId': '12325', 'nickname': 'Flameng
o', 'urlName': 'flamengo', 'teamShortName': 'Flamengo', 'confName': 'Intl', 'd
ivName': ''}, {'isNBAFranchise': True, 'isAllStar': False, 'city': 'Sacrament
o', 'altCityName': 'Sacramento', 'fullName': 'Sacramento Kings', 'tricode': 'S
AC', 'teamId': '1610612758', 'nickname': 'Kings', 'urlName': 'kings', 'teamSho
rtName': 'Sacramento', 'confName': 'West', 'divName': 'Pacific'}, {'isNBAFranc
hise': True, 'isAllStar': False, 'city': 'San Antonio', 'altCityName': 'San An
tonio', 'fullName': 'San Antonio Spurs', 'tricode': 'SAS', 'teamId': '16106127
59', 'nickname': 'Spurs', 'urlName': 'spurs', 'teamShortName': 'San Antonio',
'confName': 'West', 'divName': 'Southwest'}, {'isNBAFranchise': False, 'isAllS
tar': False, 'city': 'Shanghai', 'altCityName': 'Shanghai', 'fullName': 'Shang
hai Sharks', 'tricode': 'SDS', 'teamId': '12329', 'nickname': 'Sharks', 'urlNa
me': 'shanghai_sharks', 'teamShortName': 'Shanghai', 'confName': 'Intl', 'divN
ame': ''}, {'isNBAFranchise': False, 'isAllStar': False, 'city': 'Sydney', 'al
tCityName': 'Sydney', 'fullName': 'Sydney Kings', 'tricode': 'SYD', 'teamId':
'15015', 'nickname': 'Kings', 'urlName': 'sydkings', 'teamShortName': 'Sydne
y', 'confName': 'Intl', 'divName': ''}, {'isNBAFranchise': False, 'isAllStar':
True, 'city': 'Team', 'altCityName': 'Team', 'fullName': 'All-Stars', 'tricod
e': 'EST', 'teamId': '1699999999', 'nickname': 'All-Stars', 'urlName': 'assn_a
way', 'confName': 'East', 'divName': 'East'}, {'isNBAFranchise': False, 'isAll
Star': True, 'city': 'Team', 'altCityName': 'Team', 'fullName': 'All-Stars',
'tricode': 'WST', 'teamId': '1699999998', 'nickname': 'All-Stars', 'urlName':
'assn_home', 'confName': 'West', 'divName': 'West'}, {'isNBAFranchise': False,
'isAllStar': True, 'city': 'Team Giannis', 'altCityName': 'Team Giannis', 'ful
lName': 'Team Giannis', 'tricode': 'GNS', 'teamId': '1610616833', 'nickname':
'Team Giannis', 'urlName': 'team_giannis', 'teamShortName': 'Team Giannis', 'c
onfName': 'East', 'divName': 'East'}, {'isNBAFranchise': False, 'isAllStar': T
rue, 'city': 'Team LeBron', 'altCityName': 'Team LeBron', 'fullName': 'Team Le
Bron', 'tricode': 'LBN', 'teamId': '1610616834', 'nickname': 'Team LeBron', 'u
rlName': 'team_lebron', 'teamShortName': 'Team LeBron', 'confName': 'West', 'd
ivName': 'West'}, {'isNBAFranchise': True, 'isAllStar': False, 'city': 'Toront
o', 'altCityName': 'Toronto', 'fullName': 'Toronto Raptors', 'tricode': 'TOR',
'teamId': '1610612761', 'nickname': 'Raptors', 'urlName': 'raptors', 'teamShor
tName': 'Toronto', 'confName': 'East', 'divName': 'Atlantic'}, {'isNBAFranchis
e': False, 'isAllStar': True, 'city': 'USA', 'altCityName': 'USA', 'fullName':
```

```
'USA', 'tricode': 'USA', 'teamId': '1610616843', 'nickname': 'USA', 'urlName':
'usa', 'teamShortName': 'USA', 'confName': 'East', 'divName': 'East'}, {'isNBA
Franchise': True, 'isAllStar': False, 'city': 'Utah', 'altCityName': 'Utah',
'fullName': 'Utah Jazz', 'tricode': 'UTA', 'teamId': '1610612762', 'nickname':
'Jazz', 'urlName': 'jazz', 'teamShortName': 'Utah', 'confName': 'West', 'divNa
me': 'Northwest'}, {'isNBAFranchise': True, 'isAllStar': False, 'city': 'Washi
ngton', 'altCityName': 'Washington', 'fullName': 'Washington Wizards', 'tricod
e': 'WAS', 'teamId': '1610612764', 'nickname': 'Wizards', 'urlName': 'wizard
s', 'teamShortName': 'Washington', 'confName': 'East', 'divName': 'Southeas
t'}, {'isNBAFranchise': False, 'isAllStar': True, 'city': 'World', 'altCityNam
e': 'World', 'fullName': 'World', 'tricode': 'WLD', 'teamId': '1610616844', 'n
ickname': 'World', 'urlName': 'world', 'teamShortName': 'World', 'confName':
'East', 'divName': 'East'}], 'africa': [{'isNBAFranchise': False, 'isAllStar':
False, 'city': 'Team', 'altCityName': 'Team', 'fullName': 'Team USA', 'tricod
e': 'USA', 'teamId': '22', 'nickname': 'USA', 'urlName': 'nhs_usa', 'teamShort
Name': 'USA', 'confName': '', 'divName': ''}, {'isNBAFranchise': False, 'isAll
Star': False, 'city': 'Team', 'altCityName': 'Team', 'fullName': 'Team World',
'tricode': 'WLD', 'teamId': '21', 'nickname': 'World', 'urlName': 'nhs_world',
'teamShortName': 'World', 'confName': '', 'divName': ''}], 'sacramento': [{'is
NBAFranchise': True, 'isAllStar': False, 'city': 'Golden State', 'altCityNam
e': 'Golden State', 'fullName': 'Golden State Warriors', 'tricode': 'GSW', 'te
amId': '1610612744', 'nickname': 'Warriors', 'urlName': 'warriors', 'teamShort
Name': 'Golden State', 'confName': 'Sacramento', 'divName': ''}, {'isNBAFranch
ise': True, 'isAllStar': False, 'city': 'Los Angeles', 'altCityName': 'Los Ang
eles Lakers', 'fullName': 'Los Angeles Lakers', 'tricode': 'LAL', 'teamId': '1
610612747', 'nickname': 'Lakers', 'urlName': 'lakers', 'teamShortName': 'L.A.
Lakers', 'confName': 'Sacramento', 'divName': ''}, {'isNBAFranchise': True, 'i
sAllStar': False, 'city': 'Miami', 'altCityName': 'Miami', 'fullName': 'Miami
Heat', 'tricode': 'MIA', 'teamId': '1610612748', 'nickname': 'Heat', 'urlNam
e': 'heat', 'teamShortName': 'Miami', 'confName': 'Sacramento', 'divName':
''}, {'isNBAFranchise': True, 'isAllStar': False, 'city': 'Sacramento', 'altCi
tyName': 'Sacramento', 'fullName': 'Sacramento Kings', 'tricode': 'SAC', 'team
Id': '1610612758', 'nickname': 'Kings', 'urlName': 'kings', 'teamShortName':
'Sacramento', 'confName': 'Sacramento', 'divName': ''}], 'vegas': [{'isNBAFran
chise': True, 'isAllStar': False, 'city': 'Atlanta', 'altCityName': 'Atlanta',
'fullName': 'Atlanta Hawks', 'tricode': 'ATL', 'teamId': '1610612737', 'nickna
me': 'Hawks', 'urlName': 'hawks', 'teamShortName': 'Atlanta', 'confName': 'sum
mer', 'divName': ''}, {'isNBAFranchise': True, 'isAllStar': False, 'city': 'Bo
```

```
ston', 'altCityName': 'Boston', 'fullName': 'Boston Celtics', 'tricode': 'BO
S', 'teamId': '1610612738', 'nickname': 'Celtics', 'urlName': 'celtics', 'team
ShortName': 'Boston', 'confName': 'summer', 'divName': ''}, {'isNBAFranchise':
True, 'isAllStar': False, 'city': 'Brooklyn', 'altCityName': 'Brooklyn', 'full
Name': 'Brooklyn Nets', 'tricode': 'BKN', 'teamId': '1610612751', 'nickname':
'Nets', 'urlName': 'nets', 'teamShortName': 'Brooklyn', 'confName': 'summer',
'divName': ''}, {'isNBAFranchise': True, 'isAllStar': False, 'city': 'Charlott
e', 'altCityName': 'Charlotte', 'fullName': 'Charlotte Hornets', 'tricode': 'C
HA', 'teamId': '1610612766', 'nickname': 'Hornets', 'urlName': 'hornets', 'tea
mShortName': 'Charlotte', 'confName': 'summer', 'divName': ''}, {'isNBAFranchi
se': True, 'isAllStar': False, 'city': 'Chicago', 'altCityName': 'Chicago', 'f
ullName': 'Chicago Bulls', 'tricode': 'CHI', 'teamId': '1610612741', 'nicknam
e': 'Bulls', 'urlName': 'bulls', 'teamShortName': 'Chicago', 'confName': 'summ
er', 'divName': ''}, {'isNBAFranchise': False, 'isAllStar': False, 'city': 'Ch
ina', 'altCityName': 'China', 'fullName': 'Team China', 'tricode': 'CHN', 'tea
mId': '45', 'nickname': 'China', 'urlName': 'china', 'teamShortName': 'China',
'confName': 'summer', 'divName': ''}, {'isNBAFranchise': True, 'isAllStar': Fa
lse, 'city': 'Cleveland', 'altCityName': 'Cleveland', 'fullName': 'Cleveland C
avaliers', 'tricode': 'CLE', 'teamId': '1610612739', 'nickname': 'Cavaliers',
'urlName': 'cavaliers', 'teamShortName': 'Cleveland', 'confName': 'summer', 'd
ivName': ''}, {'isNBAFranchise': False, 'isAllStar': False, 'city': 'Croatia',
'altCityName': 'Croatia', 'fullName': 'Team Croatia', 'tricode': 'CRO', 'teamI
d': '70', 'nickname': 'Croatia', 'urlName': 'croatia', 'teamShortName': 'Croat
ia', 'confName': 'summer', 'divName': ''}, {'isNBAFranchise': True, 'isAllSta
r': False, 'city': 'Dallas', 'altCityName': 'Dallas', 'fullName': 'Dallas Mave
ricks', 'tricode': 'DAL', 'teamId': '1610612742', 'nickname': 'Mavericks', 'ur
lName': 'mavericks', 'teamShortName': 'Dallas', 'confName': 'summer', 'divNam
e': ''}, {'isNBAFranchise': True, 'isAllStar': False, 'city': 'Denver', 'altCi
tyName': 'Denver', 'fullName': 'Denver Nuggets', 'tricode': 'DEN', 'teamId':
'1610612743', 'nickname': 'Nuggets', 'urlName': 'nuggets', 'teamShortName': 'D
enver', 'confName': 'summer', 'divName': ''}, {'isNBAFranchise': True, 'isAllS
tar': False, 'city': 'Detroit', 'altCityName': 'Detroit', 'fullName': 'Detroit
Pistons', 'tricode': 'DET', 'teamId': '1610612765', 'nickname': 'Pistons', 'ur
lName': 'pistons', 'teamShortName': 'Detroit', 'confName': 'summer', 'divNam
e': ''}, {'isNBAFranchise': True, 'isAllStar': False, 'city': 'Golden State',
'altCityName': 'Golden State', 'fullName': 'Golden State Warriors', 'tricode':
'GSW', 'teamId': '1610612744', 'nickname': 'Warriors', 'urlName': 'warriors',
'teamShortName': 'Golden State', 'confName': 'summer', 'divName': ''}, {'isNBA
```

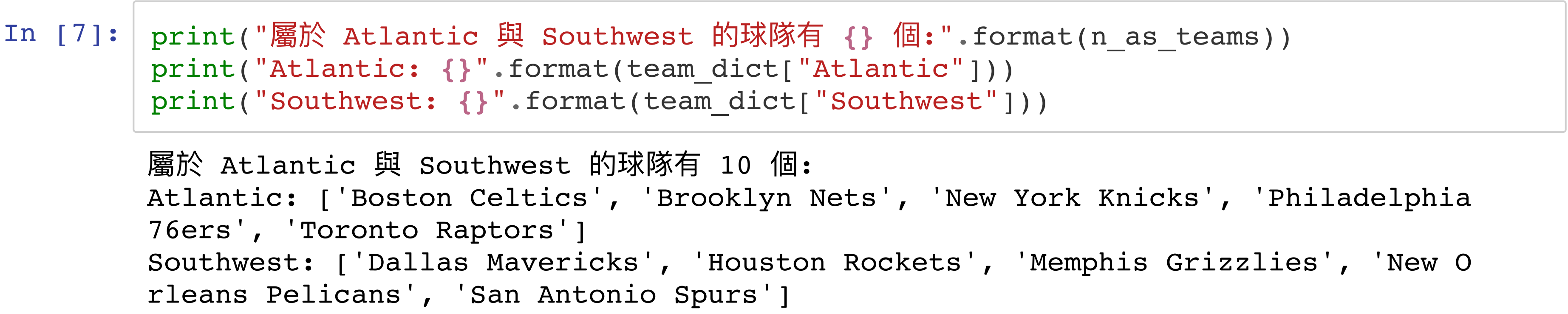
```
Franchise': True, 'isAllStar': False, 'city': 'Houston', 'altCityName': 'Houst
on', 'fullName': 'Houston Rockets', 'tricode': 'HOU', 'teamId': '1610612745',
'nickname': 'Rockets', 'urlName': 'rockets', 'teamShortName': 'Houston', 'conf
Name': 'summer', 'divName': ''}, {'isNBAFranchise': True, 'isAllStar': False,
'city': 'Indiana', 'altCityName': 'Indiana', 'fullName': 'Indiana Pacers', 'tr
icode': 'IND', 'teamId': '1610612754', 'nickname': 'Pacers', 'urlName': 'pacer
s', 'teamShortName': 'Indiana', 'confName': 'summer', 'divName': ''}, {'isNBAF
ranchise': True, 'isAllStar': False, 'city': 'LA', 'altCityName': 'LA Clipper
s', 'fullName': 'LA Clippers', 'tricode': 'LAC', 'teamId': '1610612746', 'nick
name': 'Clippers', 'urlName': 'clippers', 'teamShortName': 'LA Clippers', 'con
fName': 'summer', 'divName': ''}, {'isNBAFranchise': True, 'isAllStar': False,
'city': 'Los Angeles', 'altCityName': 'Los Angeles Lakers', 'fullName': 'Los A
ngeles Lakers', 'tricode': 'LAL', 'teamId': '1610612747', 'nickname': 'Laker
s', 'urlName': 'lakers', 'teamShortName': 'L.A. Lakers', 'confName': 'summer',
'divName': ''}, {'isNBAFranchise': True, 'isAllStar': False, 'city': 'Memphi
s', 'altCityName': 'Memphis', 'fullName': 'Memphis Grizzlies', 'tricode': 'ME
M', 'teamId': '1610612763', 'nickname': 'Grizzlies', 'urlName': 'grizzlies',
'teamShortName': 'Memphis', 'confName': 'summer', 'divName': ''}, {'isNBAFranc
hise': True, 'isAllStar': False, 'city': 'Miami', 'altCityName': 'Miami', 'ful
lName': 'Miami Heat', 'tricode': 'MIA', 'teamId': '1610612748', 'nickname': 'H
eat', 'urlName': 'heat', 'teamShortName': 'Miami', 'confName': 'summer', 'divN
ame': ''}, {'isNBAFranchise': True, 'isAllStar': False, 'city': 'Milwaukee',
'altCityName': 'Milwaukee', 'fullName': 'Milwaukee Bucks', 'tricode': 'MIL',
'teamId': '1610612749', 'nickname': 'Bucks', 'urlName': 'bucks', 'teamShortNam
e': 'Milwaukee', 'confName': 'summer', 'divName': ''}, {'isNBAFranchise': Tru
e, 'isAllStar': False, 'city': 'Minnesota', 'altCityName': 'Minnesota', 'fullN
ame': 'Minnesota Timberwolves', 'tricode': 'MIN', 'teamId': '1610612750', 'nic
kname': 'Timberwolves', 'urlName': 'timberwolves', 'teamShortName': 'Minnesot
a', 'confName': 'summer', 'divName': ''}, {'isNBAFranchise': True, 'isAllSta
r': False, 'city': 'New Orleans', 'altCityName': 'New Orleans', 'fullName': 'N
ew Orleans Pelicans', 'tricode': 'NOP', 'teamId': '1610612740', 'nickname': 'P
elicans', 'urlName': 'pelicans', 'teamShortName': 'New Orleans', 'confName':
'summer', 'divName': ''}, {'isNBAFranchise': True, 'isAllStar': False, 'city':
'New York', 'altCityName': 'New York', 'fullName': 'New York Knicks', 'tricod
e': 'NYK', 'teamId': '1610612752', 'nickname': 'Knicks', 'urlName': 'knicks',
'teamShortName': 'New York', 'confName': 'summer', 'divName': ''}, {'isNBAFran
chise': True, 'isAllStar': False, 'city': 'Oklahoma City', 'altCityName': 'Okl
ahoma City', 'fullName': 'Oklahoma City Thunder', 'tricode': 'OKC', 'teamId':
```

## 隨堂練習：2019-2020 球季 NBA 有幾支球隊?

```
In [5]: print("2019-2020 球季 NBA 有 {} 支球隊".format(n_nba_teams))
```

```
2019-2020 球季 NBA 有 30 支球隊
```

# 隨堂練習：屬於 Atlantic 與 Southwest 的球隊有幾個？各隊名為？



```
In [7]: print("屬於 Atlantic 與 Southwest 的球隊有 {} 個:".format(n_as_teams))
        print("Atlantic: {}".format(team_dict["Atlantic"]))
        print("Southwest: {}".format(team_dict["Southwest"]))
```

```
屬於 Atlantic 與 Southwest 的球隊有 10 個:
Atlantic: ['Boston Celtics', 'Brooklyn Nets', 'New York Knicks', 'Philadelphia
76ers', 'Toronto Raptors']
Southwest: ['Dallas Mavericks', 'Houston Rockets', 'Memphis Grizzlies', 'New O
rleans Pelicans', 'San Antonio Spurs']
```

# 擷取 XML 格式網頁資料

## 擷取 XML 格式網頁資料步驟

- 使用 `requests` 請求資料
- 使用回應的 `.content` 屬性，例如 `response.content`
- 以 `lxml` 搭配 XPath 解析

## 以 https://emap.pcsc.com.tw (https://emap.pcsc.com.tw) 示範

In [8]:

```
import requests

#進行 POST 請求時要攜帶資料
form_data = {
    "commandid": "GetTown",
    "cityid": "01"
}
request_url = "https://emap.pcsc.com.tw/EMapSDK.aspx"
response = requests.post(request_url, data=form_data)
print(response.status_code)
```

```
200
```

# 使用 .content 屬性

```
In [9]: response_content = response.content
        print(response_content)
```

```
b'<?xml version="1.0" encoding="utf-8"?><iMapSDKOutput><MessageID>00000</Messa
geID><CommandID>GetTown</CommandID><Status>\xe9\x80\xa3\xe7\xb7\x9a\xe6\x88\x9
0\xe5\x8a\x9f</Status><TimeStamp>2019/11/29 \xe4\xb8\x8b\xe5\x8d\x88 04:04:18
</TimeStamp><GeoPosition><TownID>01</TownID><TownName>\xe6\x9d\xbe\xe5\xb1\xb1
\xe5\x8d\x80</TownName><X>121577218</X><Y>25049837</Y></GeoPosition><GeoPositi
on><TownID>02</TownID><TownName>\xe4\xbf\xa1\xe7\xbe\xa9\xe5\x8d\x80</TownName
><X>121567161</X><Y>25033147</Y></GeoPosition><GeoPosition><TownID>03</TownID>
<TownName>\xe5\xa4\xa7\xe5\xae\x89\xe5\x8d\x80</TownName><X>121534593</X><Y>25
026482</Y></GeoPosition><GeoPosition><TownID>04</TownID><TownName>\xe4\xb8\xad
\xe5\xb1\xb1\xe5\x8d\x80</TownName><X>121533655</X><Y>25064427</Y></GeoPositio
n><GeoPosition><TownID>05</TownID><TownName>\xe4\xb8\xad\xe6\xad\xa3\xe5\x8d\x
80</TownName><X>121518245</X><Y>25032251</Y></GeoPosition><GeoPosition><TownID
>06</TownID><TownName>\xe5\xa4\xa7\xe5\x90\x8c\xe5\x8d\x80</TownName><X>121515
830</X><Y>25066142</Y></GeoPosition><GeoPosition><TownID>07</TownID><TownName>
\xe8\x90\xac\xe8\x8f\xaf\xe5\x8d\x80</TownName><X>121499745</X><Y>25034807</Y>
</GeoPosition><GeoPosition><TownID>08</TownID><TownName>\xe6\x96\x87\xe5\xb1\x
b1\xe5\x8d\x80</TownName><X>121570280</X><Y>24989800</Y></GeoPosition><GeoPosi
tion><TownID>09</TownID><TownName>\xe5\x8d\x97\xe6\xb8\xaf\xe5\x8d\x80</TownNa
me><X>121607043</X><Y>25054684</Y></GeoPosition><GeoPosition><TownID>10</TownI
D><TownName>\xe5\x85\xa7\xe6\xb9\x96\xe5\x8d\x80</TownName><X>121589471</X><Y>
25069353</Y></GeoPosition><GeoPosition><TownID>11</TownID><TownName>\xe5\xa3\x
ab\xe6\x9e\x97\xe5\x8d\x80</TownName><X>121525380</X><Y>25090430</Y></GeoPosit
ion><GeoPosition><TownID>12</TownID><TownName>\xe5\x8c\x97\xe6\x8a\x95\xe5\x8d
\x80</TownName><X>121503066</X><Y>25132054</Y></GeoPosition></iMapSDKOutput>'
```

# 以 lxml 解析

In [10]:

```
from lxml import etree
from io import BytesIO

file = BytesIO(response_content)
tree = etree.parse(file)
town_names = [t.text for t in tree.xpath("//TownName")]
print(town_names)
```

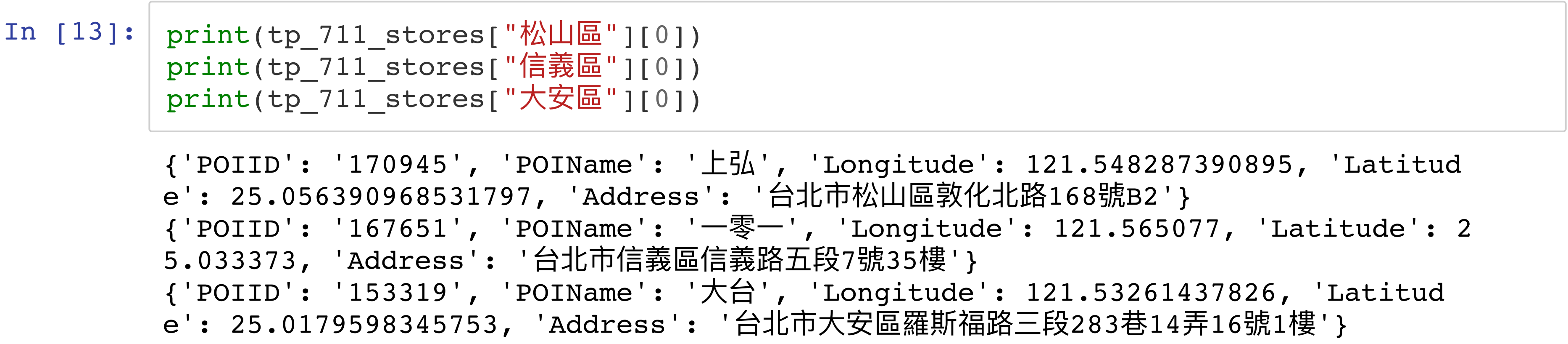
```
['松山區', '信義區', '大安區', '中山區', '中正區', '大同區', '萬華區', '文山區', '南港
區', '內湖區', '士林區', '北投區']
```

In [11]:

```
from lxml import etree
from io import BytesIO

file = BytesIO(response_content)
tree = etree.parse(file)
town_names = [t.text for t in tree.xpath("//TownName")]
print(town_names)
```

```
['松山區', '信義區', '大安區', '中山區', '中正區', '大同區', '萬華區', '文山區', '南港
區', '內湖區', '士林區', '北投區']
```

# 隨堂練習：擷取台北市所有 7-11 便利商店資訊



```
In [13]: print(tp_711_stores["松山區"][0])
         print(tp_711_stores["信義區"][0])
         print(tp_711_stores["大安區"][0])
```

```
{'POIID': '170945', 'POIName': '上弘', 'Longitude': 121.548287390895, 'Latitud
e': 25.056390968531797, 'Address': '台北市松山區敦化北路168號B2'}
{'POIID': '167651', 'POIName': '一零一', 'Longitude': 121.565077, 'Latitude': 2
5.033373, 'Address': '台北市信義區信義路五段7號35樓'}
{'POIID': '153319', 'POIName': '大台', 'Longitude': 121.53261437826, 'Latitud
e': 25.0179598345753, 'Address': '台北市大安區羅斯福路三段283巷14弄16號1樓'}
```

# 擷取 HTML 格式網頁資料

## 擷取 HTML 格式網頁資料步驟

- 使用 `requests` 請求資料
- 使用回應的 `.text` 屬性，例如 `response.text`
- 以 `bs4` 搭配 Tag Name/CSS Selector 解析

# 常見用來標示 HTML 資料的方法

- **HTML 的標籤名稱**
- HTML 標籤中給予的 id
- HTML 標籤中給予的 class
- **資料所在的 CSS 選擇器（CSS Selector）**
- 資料所在的 XPath

# 幫助定位 CSS 選擇器的 Chrome 外掛

SelectorGadget
(https://chrome.google.com/webstore/detail/selectorgadget/mhjhnkcfbdhnjickkkdbjoemdmb

## [SelectorGadget (https://chrome.google.com/webstore/detail/selectorgadget/mhjhnkcfbdhnjickkkdbjoemdmbfginb)](https://chrome.google.com/webstore/detail/selectorgadget/mhjhnkcfbdhnjickkkdbjoemdmbfginb) 的使用方法

1. 點選 SelectorGadget 的外掛圖示
2. 留意 SelectorGadget 的 CSS 選擇器
3. 移動滑鼠到想要定位的元素
4. 在想要定位的資料上面點選左鍵，留意 Clear 後面數字表示有多少個元素被選擇到
5. 移動滑鼠點選不要選擇的元素（改以紅底標記），並同時注意 CSS 選擇器位址與 Clear 後面數字

## 以 Avengers: Endgame (2019) (https://www.imdb.com/title/tt4154796) 示範 SelectorGadget (https://chrome.google.com/webstore/detail/selectorgadget/mhjhnkcfbdhnjickkkdbjoemdmbfginb) 的使用方法

- 電影名稱
- 電影海報
- 評分
- 劇情類型
- 演員陣容

以 Avengers: Endgame (2019) (https://www.imdb.com/title/tt4154796) 示範 bs4

# 常用的 bs4 函數

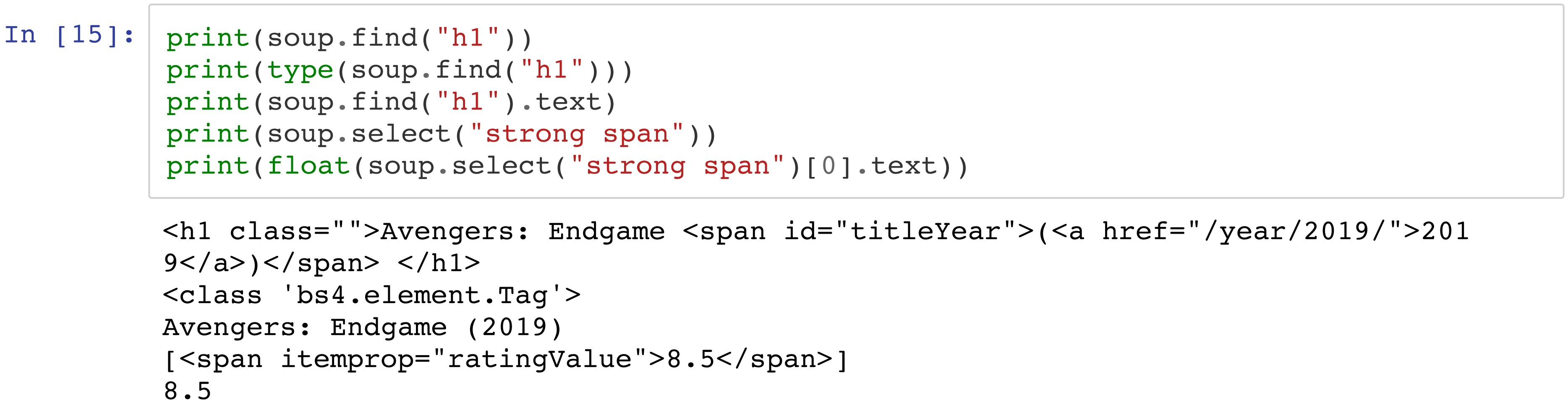
`BeautifulSoup()`：創建 `soup` 類別

```
In [14]: import requests
         from bs4 import BeautifulSoup

         request_url = "https://www.imdb.com/title/tt4154796"
         response = requests.get(request_url)
         response_text = response.text
         soup = BeautifulSoup(response_text)
         print(type(soup))
```

```
<class 'bs4.BeautifulSoup'>
```

## 常用的 soup 方法

- `soup.find()`：尋找第一個符合標記名稱的資料
- `soup.find_all()`：尋找所有符合標記名稱的資料
- `soup.select()`：尋找所有符合 CSS 選擇的資料



In [15]:

```
print(soup.find("h1"))
print(type(soup.find("h1")))
print(soup.find("h1").text)
print(soup.select("strong span"))
print(float(soup.select("strong span")[0].text))
```

```
<h1 class="">Avengers: Endgame <span id="titleYear">(<a href="/year/2019/">201
9</a>)</span> </h1>
<class 'bs4.element.Tag'>
Avengers: Endgame (2019)
[<span itemprop="ratingValue">8.5</span>]
8.5
```

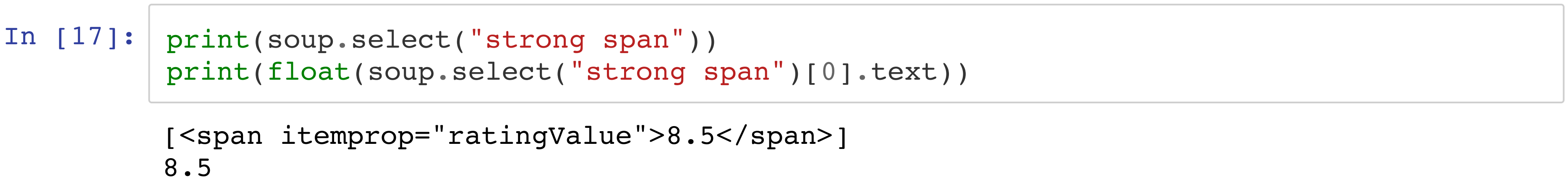
## 常用的 element.Tag 屬性、方法

- `element.Tag.text`：取出標記中的文字值
- `element.Tag.get(attr)`：取出標記中的指定屬性

In [16]:

```
print(len(soup.find_all("img")))
print(soup.find_all("img")[2])
print(soup.find_all("img")[2].get("alt"))
print(soup.find_all("img")[2].get("src"))
```

```
90
<img class="pro_logo" src="https://m.media-amazon.com/images/G/01/imdb/IMDbCon
sumerSiteProTitleViews/images/logo/pro_logo_dark-3176609149._CB455053166_.pn
g"/>
None
https://m.media-amazon.com/images/G/01/imdb/IMDbConsumerSiteProTitleViews/imag
es/logo/pro_logo_dark-3176609149._CB455053166_.png
```



In [17]:

```
print(soup.select("strong span"))
print(float(soup.select("strong span")[0].text))
```

```
[<span itemprop="ratingValue">8.5</span>]
8.5
```

## 隨堂練習：以 `requests` 搭配 `bs4` 擷取 Avengers: Endgame (2019) (https://www.imdb.com/title/tt4154796) 的劇情類型

In [19]:

```python
for g in genre:
    print(g.text)
```

```
Action
Adventure
Drama
```

## 隨堂練習：以 requests 搭配 bs4 擷取 Avengers: Endgame (2019) (https://www.imdb.com/title/tt4154796) 的演員陣容

In [21]:

```python
print(cast)
```

```
['Robert Downey Jr.', 'Chris Evans', 'Mark Ruffalo', 'Chris Hemsworth', 'Scarl
ett Johansson', 'Jeremy Renner', 'Don Cheadle', 'Paul Rudd', 'Benedict Cumberb
atch', 'Chadwick Boseman', 'Brie Larson', 'Tom Holland', 'Karen Gillan', 'Zoe
Saldana', 'Evangeline Lilly']
```

# 隨堂練習：自訂一個函數 get_movie_data(movie_url)

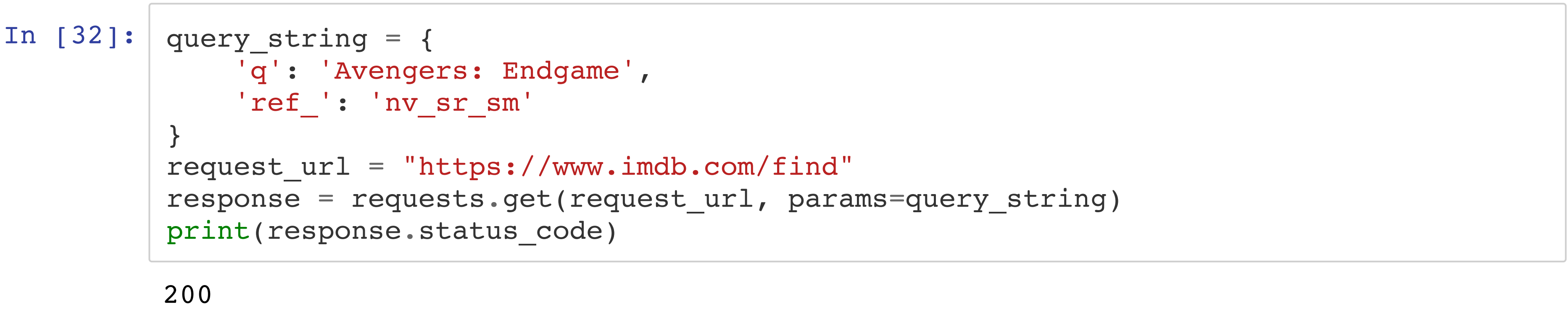
```
In [31]: get_movie_data("https://www.imdb.com/title/tt4154796")
```

```
Out[31]: {'movieTitle': 'Avengers: Endgame (2019)',
          'moviePoster': 'https://m.media-amazon.com/images/M/MV5BMTc5MDE2ODcwNV5BMl5Ba
         nBnXkFtZTgwMzI2NzQ2NzM@._V1_UX182_CR0,0,182,268_AL_.jpg',
          'movieRating': 8.5,
          'movieGenre': ['Action', 'Adventure', 'Drama'],
          'movieCast': ['Robert Downey Jr.',
           'Chris Evans',
           'Mark Ruffalo',
           'Chris Hemsworth',
           'Scarlett Johansson',
           'Jeremy Renner',
           'Don Cheadle',
           'Paul Rudd',
           'Benedict Cumberbatch',
           'Chadwick Boseman',
           'Brie Larson',
           'Tom Holland',
           'Karen Gillan',
           'Zoe Saldana',
           'Evangeline Lilly']}
```

# 讓 `get_movie_data()` 更方便使用

- 可以輸入電影名稱，而非 URL！
- 觀察 [https://www.imdb.com/find?q=Avengers%3A+Endgame&ref_=nv_sr_sm](https://www.imdb.com/find?q=Avengers%3A+Endgame&ref_=nv_sr_sm) (https://www.imdb.com/find?q=Avengers%3A+Endgame&ref_=nv_sr_sm)

在 `get()` 中加入 `params`



In [32]:

```
query_string = {
    'q': 'Avengers: Endgame',
    'ref_': 'nv_sr_sm'
}
request_url = "https://www.imdb.com/find"
response = requests.get(request_url, params=query_string)
print(response.status_code)
```

```
200
```

# 隨堂練習：自訂一個函數 get_movie_data(movie_title)

```
In [34]: get_movie_data("Avengers: Endgame (2019)")
```

```
Out[34]: {'movieTitle': 'Avengers: Endgame (2019)',
          'moviePoster': 'https://m.media-amazon.com/images/M/MV5BMTc5MDE2ODcwNV5BMl5Ba
         nBnXkFtZTgwMzI2NzQ2NzM@._V1_UX182_CR0,0,182,268_AL_.jpg',
          'movieRating': 8.5,
          'movieGenre': ['Action', 'Adventure', 'Drama'],
          'movieCast': ['Robert Downey Jr.',
           'Chris Evans',
           'Mark Ruffalo',
           'Chris Hemsworth',
           'Scarlett Johansson',
           'Jeremy Renner',
           'Don Cheadle',
           'Paul Rudd',
           'Benedict Cumberbatch',
           'Chadwick Boseman',
           'Brie Larson',
           'Tom Holland',
           'Karen Gillan',
           'Zoe Saldana',
           'Evangeline Lilly']}
```

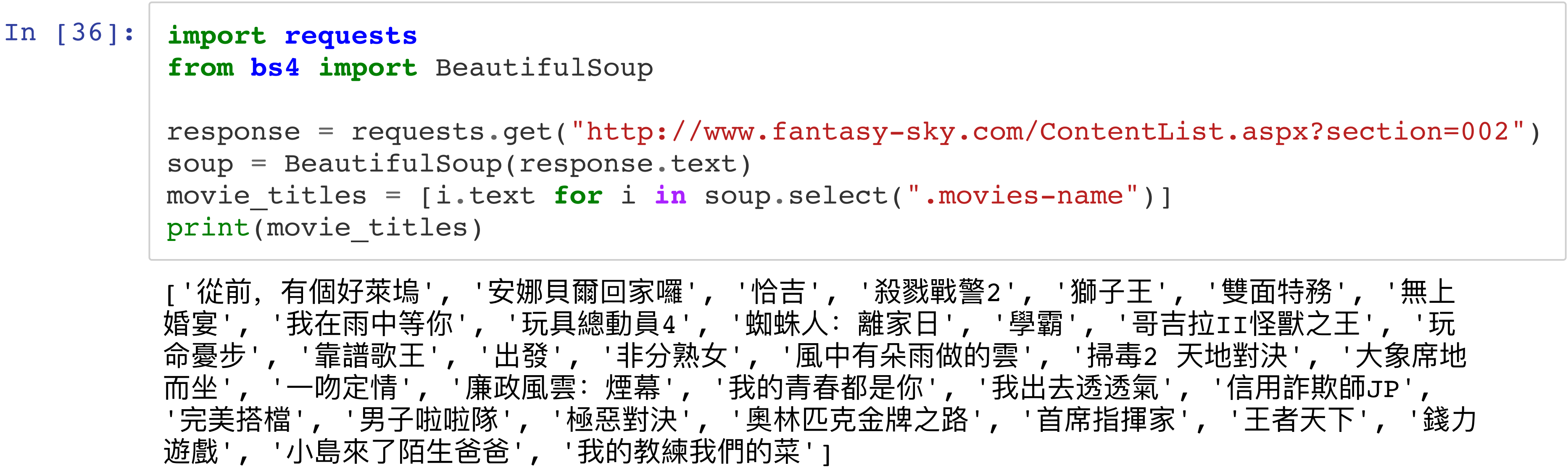
## 有時候 `requests` 送出的請求需要攜帶餅乾（cookies），否則回傳的資料會不符合預期

- PTT 八卦版 (https://www.ptt.cc/bbs/Gossiping/index.html)
- 華航機上電影清單 (http://www.fantasy-sky.com/ContentList.aspx?section=002)

```
In [35]: import requests

         response = requests.get("https://www.ptt.cc/bbs/Gossiping/index.html")
         print(response.text)
```

```
<!DOCTYPE html>
<html>
        <head>
                <meta charset="utf-8">


<meta name="viewport" content="width=device-width, initial-scale=1">

<title>批踢踢實業坊</title>

<link rel="stylesheet" type="text/css" href="//images.ptt.cc/bbs/v2.27/bbs-com
mon.css">
<link rel="stylesheet" type="text/css" href="//images.ptt.cc/bbs/v2.27/bbs-bas
e.css" media="screen">
<link rel="stylesheet" type="text/css" href="//images.ptt.cc/bbs/v2.27/bbs-cus
tom.css">
<link rel="stylesheet" type="text/css" href="//images.ptt.cc/bbs/v2.27/pushstr
eam.css" media="screen">
<link rel="stylesheet" type="text/css" href="//images.ptt.cc/bbs/v2.27/bbs-pri
nt.css" media="print">




        </head>
    <body>

<div class="bbs-screen bbs-content">
    <div class="over18-notice">
        <p>本網站已依網站內容分級規定處理</p>

        <p>警告：您即將進入之看板內容需滿十八歲方可瀏覽。</p>
        <p>警告：您即將進入之看板內容需滿十八歲方可瀏覽。</p>

        <p>若您尚未年滿十八歲，請點選離開。若您已滿十八歲，亦不可將本區之內容派發、傳閱、出
售、出租、交給或借予年齡未滿18歲的人士瀏覽，或將本網站內容向該人士出示、播放或放映。</p>
    </div>
</div>

<div class="bbs-screen bbs-content center clear">
    <form action="/ask/over18" method="post">
        <input type="hidden" name="from" value="/bbs/Gossiping/index.html">
        <div class="over18-button-container">
            <button class="btn-big" type="submit" name="yes" value="yes">我同
意，我已年滿十八歲<br><small>進入</small></button>
        </div>
        <div class="over18-button-container">
            <button class="btn-big" type="submit" name="no" value="no">未滿十八
歲或不同意本條款<br><small>離開</small></button>
        </div>
    </form>
</div>




<script>
  (function(i,s,o,g,r,a,m){i['GoogleAnalyticsObject']=r;i[r]=i[r]||function(){
  (i[r].q=i[r].q||[]).push(arguments)},i[r].l=1*new Date();a=s.createElement
(o),
  m=s.getElementsByTagName(o)[0];a.async=1;a.src=g;m.parentNode.insertBefore
(a,m)
  })(window,document,'script','https://www.google-analytics.com/analytics.j
s','ga');

  ga('create', 'UA-32365737-1', {
    cookieDomain: 'ptt.cc',
    legacyCookieDomain: 'ptt.cc'
  });
  ga('send', 'pageview');
</script>
```

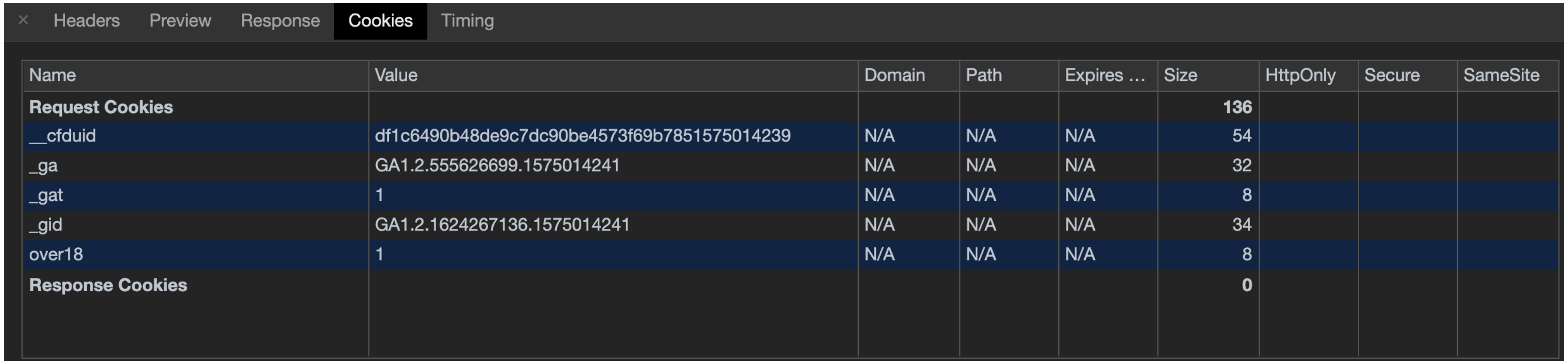
```
<script src="//ajax.googleapis.com/ajax/libs/jquery/2.1.1/jquery.min.js"></scr
ipt>
<script src="//images.ptt.cc/bbs/v2.27/bbs.js"></script>

    </body>
</html>
```



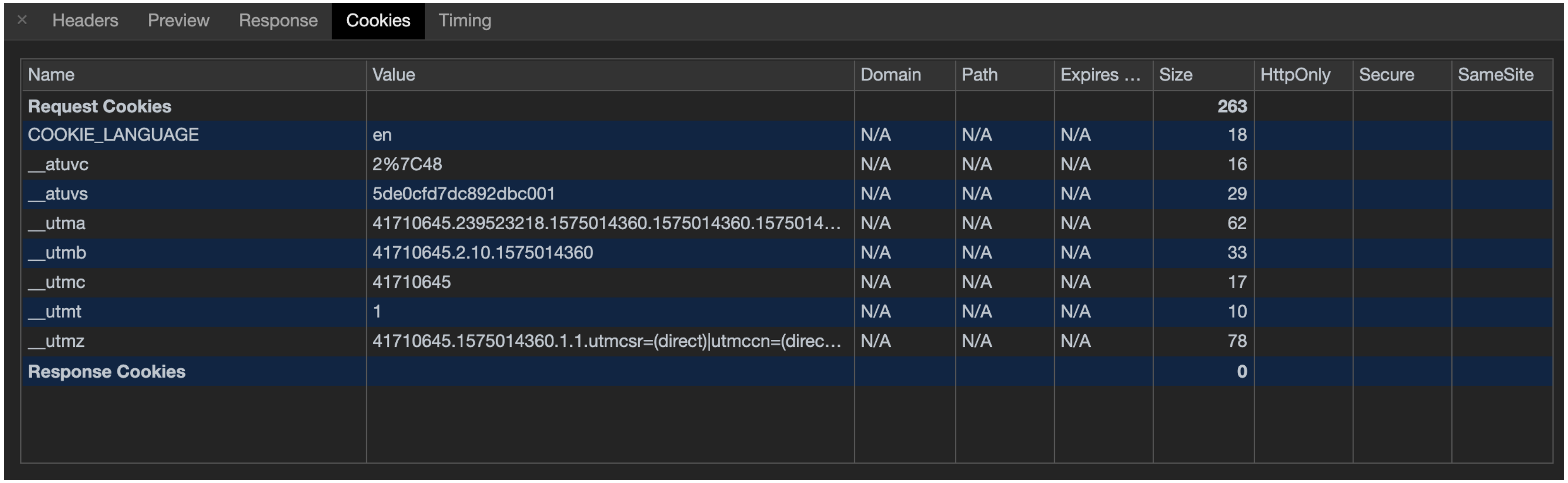
```
In [36]: import requests
         from bs4 import BeautifulSoup

         response = requests.get("http://www.fantasy-sky.com/ContentList.aspx?section=002")
         soup = BeautifulSoup(response.text)
         movie_titles = [i.text for i in soup.select(".movies-name")]
         print(movie_titles)
```

```
['從前，有個好萊塢', '安娜貝爾回家囉', '恰吉', '殺戮戰警2', '獅子王', '雙面特務', '無上
婚宴', '我在雨中等你', '玩具總動員4', '蜘蛛人：離家日', '學霸', '哥吉拉II怪獸之王', '玩
命憂步', '靠譜歌王', '出發', '非分熟女', '風中有朵雨做的雲', '掃毒2 天地對決', '大象席地
而坐', '一吻定情', '廉政風雲：煙幕', '我的青春都是你', '我出去透透氣', '信用詐欺師JP',
'完美搭檔', '男子啦啦隊', '極惡對決', '奧林匹克金牌之路', '首席指揮家', '王者天下', '錢力
遊戲', '小島來了陌生爸爸', '我的教練我們的菜']
```

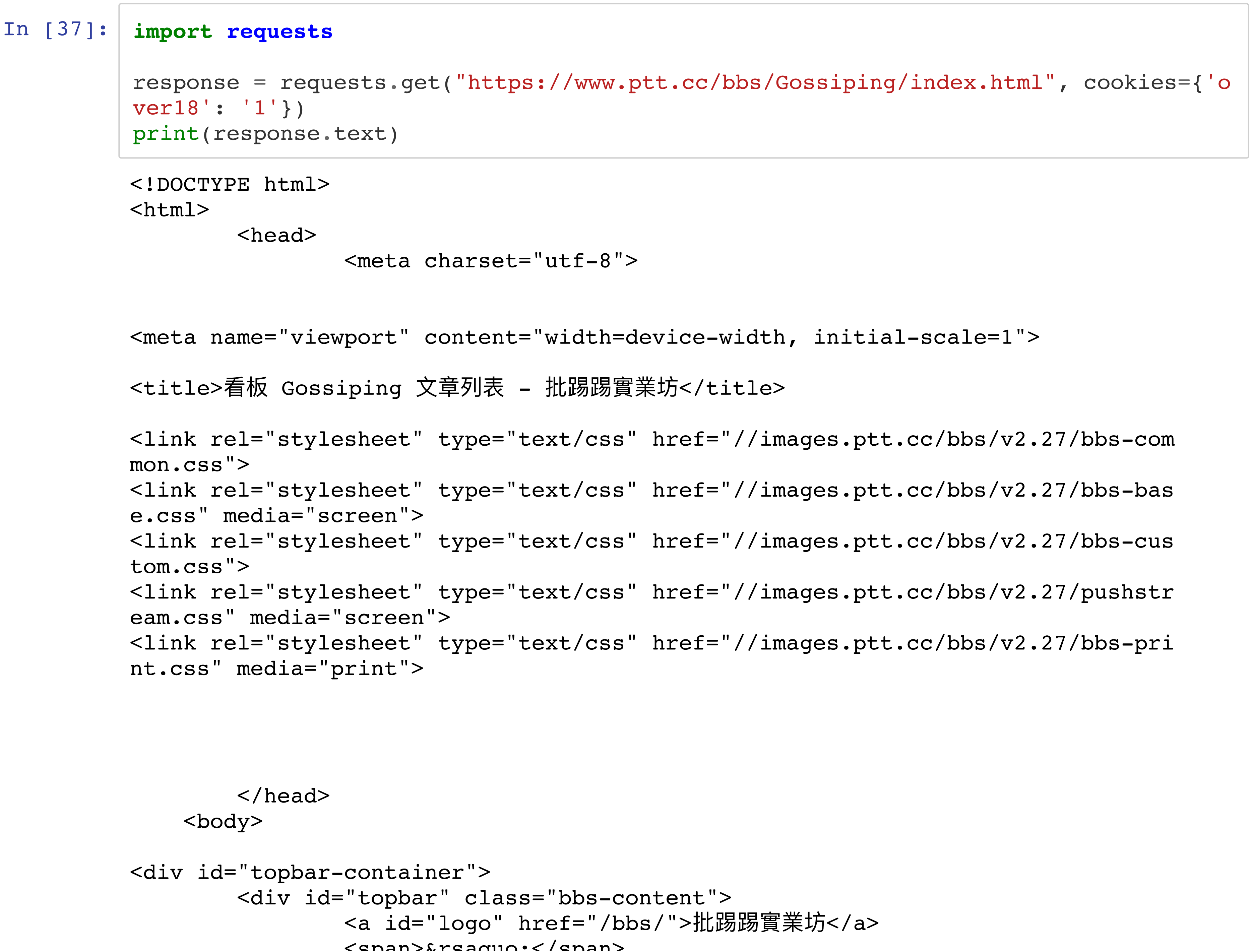
# 從開發人員工具檢視 Cookies



| Name | Value | Domain | Path | Expires ... | Size | HttpOnly | Secure | SameSite |
|---|---|---|---|---|---|---|---|---|
| **Request Cookies** | | | | | **136** | | | |
| __cfduid | df1c6490b48de9c7dc90be4573f69b7851575014239 | N/A | N/A | N/A | 54 | | | |
| _ga | GA1.2.555626699.1575014241 | N/A | N/A | N/A | 32 | | | |
| _gat | 1 | N/A | N/A | N/A | 8 | | | |
| _gid | GA1.2.1624267136.1575014241 | N/A | N/A | N/A | 34 | | | |
| over18 | 1 | N/A | N/A | N/A | 8 | | | |
| **Response Cookies** | | | | | **0** | | | |



| Name | Value | Domain | Path | Expires … | Size | HttpOnly | Secure | SameSite |
|---|---|---|---|---|---|---|---|---|
| **Request Cookies** | | | | | **263** | | | |
| COOKIE_LANGUAGE | en | N/A | N/A | N/A | 18 | | | |
| __atuvc | 2%7C48 | N/A | N/A | N/A | 16 | | | |
| __atuvs | 5de0cfd7dc892dbc001 | N/A | N/A | N/A | 29 | | | |
| __utma | 41710645.239523218.1575014360.1575014360.1575014… | N/A | N/A | N/A | 62 | | | |
| __utmb | 41710645.2.10.1575014360 | N/A | N/A | N/A | 33 | | | |
| __utmc | 41710645 | N/A | N/A | N/A | 17 | | | |
| __utmt | 1 | N/A | N/A | N/A | 10 | | | |
| __utmz | 41710645.1575014360.1.1.utmcsr=(direct)\|utmccn=(direc… | N/A | N/A | N/A | 78 | | | |
| **Response Cookies** | | | | | **0** | | | |



In [37]:

```python
import requests

response = requests.get("https://www.ptt.cc/bbs/Gossiping/index.html", cookies={'over18': '1'})
print(response.text)
```

```
<!DOCTYPE html>
<html>
        <head>
                <meta charset="utf-8">


<meta name="viewport" content="width=device-width, initial-scale=1">

<title>看板 Gossiping 文章列表 - 批踢踢實業坊</title>

<link rel="stylesheet" type="text/css" href="//images.ptt.cc/bbs/v2.27/bbs-common.css">
<link rel="stylesheet" type="text/css" href="//images.ptt.cc/bbs/v2.27/bbs-base.css" media="screen">
<link rel="stylesheet" type="text/css" href="//images.ptt.cc/bbs/v2.27/bbs-custom.css">
<link rel="stylesheet" type="text/css" href="//images.ptt.cc/bbs/v2.27/pushstream.css" media="screen">
<link rel="stylesheet" type="text/css" href="//images.ptt.cc/bbs/v2.27/bbs-print.css" media="print">




        </head>
    <body>

<div id="topbar-container">
        <div id="topbar" class="bbs-content">
                <a id="logo" href="/bbs/">批踢踢實業坊</a>
                <span>&rsaquo;</span>
                <span>&rsaquo;</span>
                <a class="board" href="/bbs/Gossiping/index.html"><span class
="board-label">看板 </span>Gossiping</a>
                <a class="right small" href="/about.html">關於我們</a>
                <a class="right small" href="/contact.html">聯絡資訊</a>
        </div>
</div>

<div id="main-container">
        <div id="action-bar-container">
                <div class="action-bar">
                        <div class="btn-group btn-group-dir">
                                <a class="btn selected" href="/bbs/Gossiping/i
ndex.html">看板</a>
                                <a class="btn" href="/man/Gossiping/index.htm
l">精華區</a>
                        </div>
                        <div class="btn-group btn-group-paging">
                                <a class="btn wide" href="/bbs/Gossiping/index
1.html">最舊</a>
                                <a class="btn wide" href="/bbs/Gossiping/index
39286.html">&lsaquo; 上頁</a>
                                <a class="btn wide disabled">下頁 &rsaquo;</a>
                                <a class="btn wide" href="/bbs/Gossiping/inde
x.html">最新</a>
                        </div>
                </div>
        </div>

        <div class="r-list-container action-bar-margin bbs-screen">
                <div class="search-bar">
                        <form type="get" action="search" id="search-bar">
                                <input class="query" type="text" name="q" valu
e="" placeholder="搜尋文章&#x22ef;">
                        </form>
                </div>
                <div class="r-ent">
                        <div class="nrec"></div>
                        <div class="title">

                                <a href="/bbs/Gossiping/M.1575014242.A.DED.htm
l">[新聞] 睡夢中一陣痛又舒服「暖流」襲來　醒來驚見他趴在床尾</a>

                        </div>
                        <div class="meta">
                                <div class="author">chinaeatshit</div>
                                <div class="article-menu">

                                        <div class="trigger">&#x22ef;</div>
                                        <div class="dropdown">
                                                <div class="item"><a href="/bb
s/Gossiping/search?q=thread%3A%5B%E6%96%B0%E8%81%9E%5D&#43;%E7%9D%A1%E5%A4%A2%
E4%B8%AD%E4%B8%80%E9%99%A3%E7%97%9B%E5%8F%88%E8%88%92%E6%9C%8D%E3%80%8C%E6%9A%
96%E6%B5%81%E3%80%8D%E8%A5%B2%E4%BE%86%E3%80%80%E9%86%92%E4%BE%86%E9%A9%9A%E8%
A6%8B%E4%BB%96%E8%B6%B4%E5%9C%A8%E5%BA%8A%E5%B0%BE">搜尋同標題文章</a></div>

                                                <div class="item"><a href="/bb
s/Gossiping/search?q=author%3Achinaeatshit">搜尋看板內 chinaeatshit 的文章</a></d
iv>

                                        </div>

                                </div>
                                <div class="date">11/29</div>
                                <div class="mark"></div>
                        </div>
                </div>
                <div class="r-ent">
                        <div class="nrec"></div>
                        <div class="title">

                                <a href="/bbs/Gossiping/M.1575014250.A.21C.htm
l">Re: [新聞] 韓國瑜又改口! 軍公教免費出國進修沒</a>

                        </div>
                        <div class="meta">
                                <div class="author">ihaveseven</div>
                                <div class="article-menu">

                                        <div class="trigger">&#x22ef;</div>
                                        <div class="dropdown">
                                                <div class="item"><a href="/bb
s/Gossiping/search?q=thread%3A%5B%E6%96%B0%E8%81%9E%5D&#43;%E9%9F%93%E5%9C%8B%
E7%91%9C%E5%8F%88%E6%94%B9%E5%8F%A3%EF%BC%81%E8%BB%8D%E5%85%AC%E6%95%99%E5%85%
8D%E8%B2%BB%E5%87%BA%E5%9C%8B%E9%80%B2%E4%BF%AE%E6%B2%92">搜尋同標題文章</a></di
v>

                                                <div class="item"><a href="/bb
s/Gossiping/search?q=author%3Aihaveseven">搜尋看板內 ihaveseven 的文章</a></div>

                                        </div>

                                </div>
                                <div class="date">11/29</div>
                                <div class="mark"></div>
                        </div>
                </div>




                <div class="r-ent">
                        <div class="nrec"><span class="hl f2">4</span></div>
                        <div class="title">

                                <a href="/bbs/Gossiping/M.1575014261.A.A8F.htm
l">[問卦] 為什麼巨乳對男生如此有吸引力? </a>

                        </div>
                        <div class="meta">
                                <div class="author">xzcb2008</div>
                                <div class="article-menu">

                                        <div class="trigger">&#x22ef;</div>
                                        <div class="dropdown">
                                                <div class="item"><a href="/bb
s/Gossiping/search?q=thread%3A%5B%E5%95%8F%E5%8D%A6%5D&#43;%E7%82%BA%E4%BB%80%
E9%BA%BC%E5%B7%A8%E4%B9%B3%E5%B0%8D%E7%94%B7%E7%94%9F%E5%A6%82%E6%AD%A4%E6%9C%
89%E5%90%B8%E5%BC%95%E5%8A%9B%EF%BC%9F">搜尋同標題文章</a></div>

                                                <div class="item"><a href="/bb
s/Gossiping/search?q=author%3Axzcb2008">搜尋看板內 xzcb2008 的文章</a></div>

                                        </div>

                                </div>
                                <div class="date">11/29</div>
                                <div class="mark"></div>
                        </div>
                </div>




                <div class="r-ent">
                        <div class="nrec"></div>
                        <div class="title">

                                <a href="/bbs/Gossiping/M.1575014276.A.92B.htm
l">[問卦] 肥宅從事砂石業有加分嗎? </a>
```

```
                        </div>
                        <div class="meta">
                                <div class="author">remmurds</div>
                                <div class="article-menu">

                                        <div class="trigger">&#x22ef;</div>
                                        <div class="dropdown">
                                                <div class="item"><a href="/bb
s/Gossiping/search?q=thread%3A%5B%E5%95%8F%E5%8D%A6%5D&#43;%E8%82%A5%E5%AE%85%
E5%BE%9E%E4%BA%8B%E7%A0%82%E7%9F%B3%E6%A5%AD%E6%9C%89%E5%8A%A0%E5%88%86%E5%97%
8E%EF%BC%9F">搜尋同標題文章</a></div>

                                                <div class="item"><a href="/bb
s/Gossiping/search?q=author%3Aremmurds">搜尋看板內 remmurds 的文章</a></div>

                                        </div>

                                </div>
                                <div class="date">11/29</div>
                                <div class="mark"></div>
                        </div>
                </div>





                <div class="r-ent">
                        <div class="nrec"><span class="hl f2">3</span></div>
                        <div class="title">

                                <a href="/bbs/Gossiping/M.1575014323.A.F5B.htm
l">[問卦] 要如何加入砂石業</a>

                        </div>
                        <div class="meta">
                                <div class="author">BrandonRoy7</div>
                                <div class="article-menu">

                                        <div class="trigger">&#x22ef;</div>
                                        <div class="dropdown">
                                                <div class="item"><a href="/bb
s/Gossiping/search?q=thread%3A%5B%E5%95%8F%E5%8D%A6%5D&#43;%E8%A6%81%E5%A6%82%
E4%BD%95%E5%8A%A0%E5%85%A5%E7%A0%82%E7%9F%B3%E6%A5%AD">搜尋同標題文章</a></div>

                                                <div class="item"><a href="/bb
s/Gossiping/search?q=author%3ABrandonRoy7">搜尋看板內 BrandonRoy7 的文章</a></div
>

                                        </div>

                                </div>
                                <div class="date">11/29</div>
                                <div class="mark"></div>
                        </div>
                </div>






                <div class="r-ent">
                        <div class="nrec"><span class="hl f2">3</span></div>
                        <div class="title">

                                <a href="/bbs/Gossiping/M.1575014421.A.DB8.htm
l">[問卦] 苗栗國跟花蓮國哪一國比較團結? </a>

                        </div>
                        <div class="meta">
                                <div class="author">ronnyvvang</div>
                                <div class="article-menu">

                                        <div class="trigger">&#x22ef;</div>
                                        <div class="dropdown">
```

```
                                                <div class="item"><a href="/bb
s/Gossiping/search?q=thread%3A%5B%E5%95%8F%E5%8D%A6%5D&#43;%E8%8B%97%E6%A0%97%
E5%9C%8B%E8%B7%9F%E8%8A%B1%E8%93%AE%E5%9C%8B%E5%93%AA%E4%B8%80%E5%9C%8B%E6%AF%
94%E8%BC%83%E5%9C%98%E7%B5%90%EF%BC%9F">搜尋同標題文章</a></div>

                                                <div class="item"><a href="/bb
s/Gossiping/search?q=author%3Aronnyvvang">搜尋看板內 ronnyvvang 的文章</a></div>

                                        </div>

                                </div>
                                <div class="date">11/29</div>
                                <div class="mark"></div>
                        </div>
                </div>





                <div class="r-ent">
                        <div class="nrec"><span class="hl f2">3</span></div>
                        <div class="title">

                                <a href="/bbs/Gossiping/M.1575014468.A.8AF.htm
l">[新聞] 要求核廢遷出蘭嶼 長老批蔡政府25億收買</a>

                        </div>
                        <div class="meta">
                                <div class="author">ff760725</div>
                                <div class="article-menu">

                                        <div class="trigger">&#x22ef;</div>
                                        <div class="dropdown">
                                                <div class="item"><a href="/bb
s/Gossiping/search?q=thread%3A%5B%E6%96%B0%E8%81%9E%5D&#43;%E8%A6%81%E6%B1%82%
E6%A0%B8%E5%BB%A2%E9%81%B7%E5%87%BA%E8%98%AD%E5%B6%BC&#43;%E9%95%B7%E8%80%81%E
6%89%B9%E8%94%A1%E6%94%BF%E5%BA%9C25%E5%84%84%E6%94%B6%E8%B2%B7">搜尋同標題文章
```

```
</a></div>

                                                <div class="item"><a href="/bb
s/Gossiping/search?q=author%3Aff760725">搜尋看板內 ff760725 的文章</a></div>

                                        </div>

                                </div>
                                <div class="date">11/29</div>
                                <div class="mark"></div>
                        </div>
                </div>





                <div class="r-ent">
                        <div class="nrec"><span class="hl f2">9</span></div>
                        <div class="title">

                                <a href="/bbs/Gossiping/M.1575014481.A.A70.htm
l">[新聞] 批吳怡農送肥皂賄選?  蔡宜芳哽咽道歉</a>

                        </div>
                        <div class="meta">
                                <div class="author">pshuang</div>
                                <div class="article-menu">

                                        <div class="trigger">&#x22ef;</div>
                                        <div class="dropdown">
                                                <div class="item"><a href="/bb
s/Gossiping/search?q=thread%3A%5B%E6%96%B0%E8%81%9E%5D&#43;%E6%89%B9%E5%90%B3%
E6%80%A1%E8%BE%B2%E9%80%81%E8%82%A5%E7%9A%82%E8%B3%84%E9%81%B8%EF%BC%9F%E3%80%
80%E8%94%A1%E5%AE%9C%E8%8A%B3%E5%93%BD%E5%92%BD%E9%81%93%E6%AD%89">搜尋同標題文
章</a></div>

                                                <div class="item"><a href="/bb
```

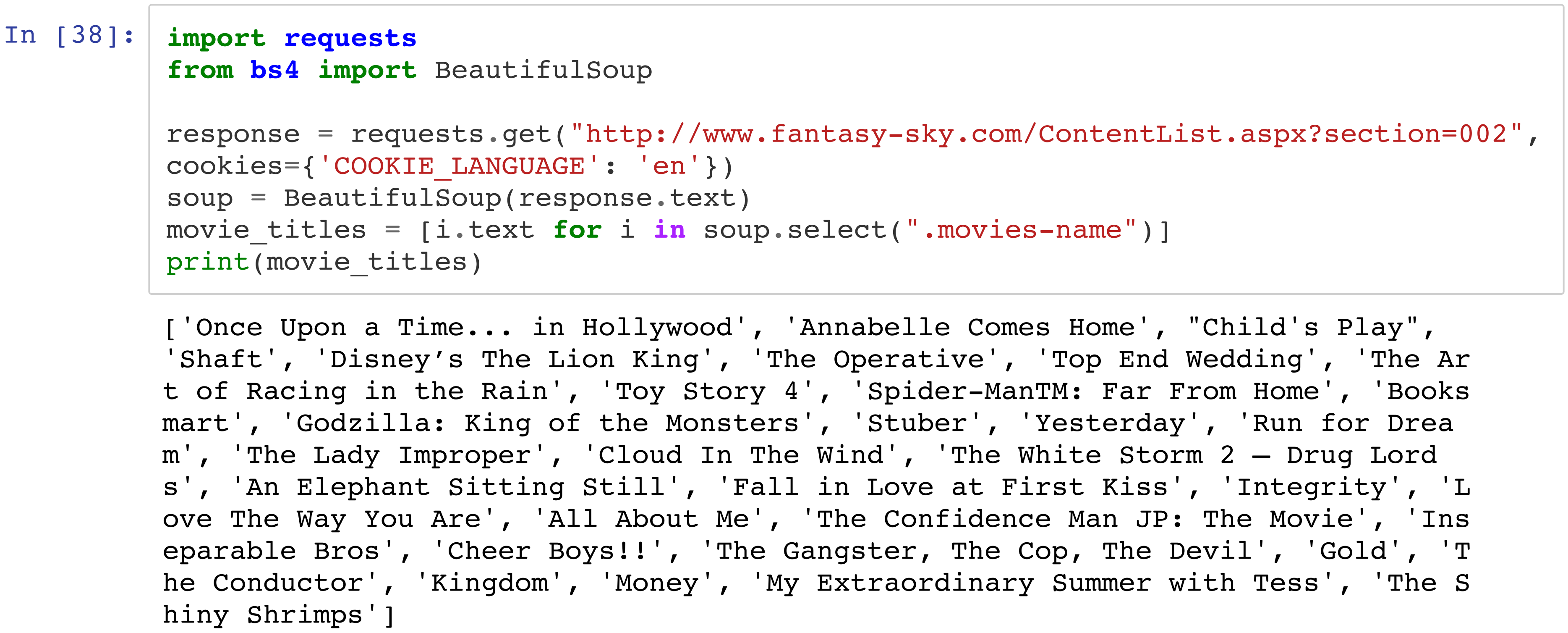
```
s/Gossiping/search?q=author%3Apshuang">搜尋看板內 pshuang 的文章</a></div>

                                        </div>

                                </div>
                                <div class="date">11/29</div>
                                <div class="mark"></div>
                        </div>
                </div>




                <div class="r-ent">
                        <div class="nrec"></div>
                        <div class="title">

                                <a href="/bbs/Gossiping/M.1575014487.A.8BF.htm
l">[問卦] 請問這是什麼鳥?</a>

                        </div>
                        <div class="meta">
                                <div class="author">graywater</div>
                                <div class="article-menu">

                                        <div class="trigger">&#x22ef;</div>
                                        <div class="dropdown">
                                                <div class="item"><a href="/bb
s/Gossiping/search?q=thread%3A%5B%E5%95%8F%E5%8D%A6%5D&#43;%E8%AB%8B%E5%95%8F%
E9%80%99%E6%98%AF%E4%BB%80%E9%BA%BC%E9%B3%A5%3F">搜尋同標題文章</a></div>

                                                <div class="item"><a href="/bb
s/Gossiping/search?q=author%3Agraywater">搜尋看板內 graywater 的文章</a></div>

                                        </div>

                                </div>
                                <div class="date">11/29</div>
                                <div class="mark"></div>
                        </div>
                </div>





                <div class="r-ent">
                        <div class="nrec"><span class="hl f2">4</span></div>
                        <div class="title">

                                <a href="/bbs/Gossiping/M.1575014558.A.D82.htm
l">[問卦] 如果上節目酸的是同性戀 風向怎麼吹? </a>

                        </div>
                        <div class="meta">
                                <div class="author">Friend5566</div>
                                <div class="article-menu">

                                        <div class="trigger">&#x22ef;</div>
                                        <div class="dropdown">
                                                <div class="item"><a href="/bb
s/Gossiping/search?q=thread%3A%5B%E5%95%8F%E5%8D%A6%5D&#43;%E5%A6%82%E6%9E%9C%
E4%B8%8A%E7%AF%80%E7%9B%AE%E9%85%B8%E7%9A%84%E6%98%AF%E5%90%8C%E6%80%A7%E6%88%
80&#43;%E9%A2%A8%E5%90%91%E6%80%8E%E9%BA%BC%E5%90%B9%EF%BC%9F">搜尋同標題文章</a
></div>

                                                <div class="item"><a href="/bb
s/Gossiping/search?q=author%3AFriend5566">搜尋看板內 Friend5566 的文章</a></div>

                                        </div>

                                </div>
                                <div class="date">11/29</div>
                                <div class="mark"></div>
                        </div>
                </div>





                <div class="r-ent">
                        <div class="nrec"><span class="hl f2">1</span></div>
                        <div class="title">

                                <a href="/bbs/Gossiping/M.1575014583.A.DF1.htm
l">Re: [問卦] 京華城關門 下一個是不是美麗華???</a>

                        </div>
                        <div class="meta">
                                <div class="author">DoraeCookie</div>
                                <div class="article-menu">

                                        <div class="trigger">&#x22ef;</div>
                                        <div class="dropdown">
                                                <div class="item"><a href="/bb
s/Gossiping/search?q=thread%3A%5B%E5%95%8F%E5%8D%A6%5D&#43;%E4%BA%AC%E8%8F%AF%
E5%9F%8E%E9%97%9C%E9%96%80&#43;%E4%B8%8B%E4%B8%80%E5%80%8B%E6%98%AF%E4%B8%8D%E
6%98%AF%E7%BE%8E%E9%BA%97%E8%8F%AF%3F%3F%3F">搜尋同標題文章</a></div>

                                                <div class="item"><a href="/bb
s/Gossiping/search?q=author%3ADoraeCookie">搜尋看板內 DoraeCookie 的文章</a></div
>

                                        </div>

                                </div>
                                <div class="date">11/29</div>
                                <div class="mark"></div>
                        </div>
                </div>
                <div class="r-ent">
                        <div class="nrec"></div>
                        <div class="title">

                                <a href="/bbs/Gossiping/M.1575014622.A.365.htm
l">[問卦] 一隻穿雲箭怎麼還沒澄清砂石案？</a>

                        </div>
                        <div class="meta">
                                <div class="author">ununnihao</div>
                                <div class="article-menu">

                                        <div class="trigger">&#x22ef;</div>
                                        <div class="dropdown">
                                                <div class="item"><a href="/bb
s/Gossiping/search?q=thread%3A%5B%E5%95%8F%E5%8D%A6%5D&#43;%E4%B8%80%E9%9A%BB%
E7%A9%BF%E9%9B%B2%E7%AE%AD%E6%80%8E%E9%BA%BC%E9%82%84%E6%B2%92%E6%BE%84%E6%B8%
85%E7%A0%82%E7%9F%B3%E6%A1%88%EF%BC%9F">搜尋同標題文章</a></div>

                                                <div class="item"><a href="/bb
s/Gossiping/search?q=author%3Aununnihao">搜尋看板內 ununnihao 的文章</a></div>

                                        </div>

                                </div>
                                <div class="date">11/29</div>
                                <div class="mark"></div>
                        </div>
                </div>




                <div class="r-ent">
```

```
                        <div class="nrec"></div>
                        <div class="title">

                                <a href="/bbs/Gossiping/M.1575014663.A.1F0.htm
l">[問卦] 大廈外面人行道擋住當停車場合法嗎</a>

                        </div>
                        <div class="meta">
                                <div class="author">ufoshooter</div>
                                <div class="article-menu">

                                        <div class="trigger">&#x22ef;</div>
                                        <div class="dropdown">
                                                <div class="item"><a href="/bb
s/Gossiping/search?q=thread%3A%5B%E5%95%8F%E5%8D%A6%5D&#43;%E5%A4%A7%E5%BB%88%
E5%A4%96%E9%9D%A2%E4%BA%BA%E8%A1%8C%E9%81%93%E6%93%8B%E4%BD%8F%E7%95%B6%E5%81%
9C%E8%BB%8A%E5%A0%B4%E5%90%88%E6%B3%95%E5%97%8E">搜尋同標題文章</a></div>

                                                <div class="item"><a href="/bb
s/Gossiping/search?q=author%3Aufoshooter">搜尋看板內 ufoshooter 的文章</a></div>

                                        </div>

                                </div>
                                <div class="date">11/29</div>
                                <div class="mark"></div>
                        </div>
                </div>




                <div class="r-ent">
                        <div class="nrec"><span class="hl f2">1</span></div>
                        <div class="title">

                                <a href="/bbs/Gossiping/M.1575014674.A.362.htm
```

```
l">[新聞] 妙齡女爛醉夢中喊暗戀男名字 醒來看內褲</a>

                        </div>
                        <div class="meta">
                                <div class="author">jodojeda</div>
                                <div class="article-menu">

                                        <div class="trigger">&#x22ef;</div>
                                        <div class="dropdown">
                                                <div class="item"><a href="/bb
s/Gossiping/search?q=thread%3A%5B%E6%96%B0%E8%81%9E%5D&#43;%E5%A6%99%E9%BD%A1%
E5%A5%B3%E7%88%9B%E9%86%89%E5%A4%A2%E4%B8%AD%E5%96%8A%E6%9A%97%E6%88%80%E7%94%
B7%E5%90%8D%E5%AD%97&#43;%E9%86%92%E4%BE%86%E7%9C%8B%E5%85%A7%E8%A4%B2">搜尋同
標題文章</a></div>

                                                <div class="item"><a href="/bb
s/Gossiping/search?q=author%3Ajodojeda">搜尋看板內 jodojeda 的文章</a></div>

                                        </div>

                                </div>
                                <div class="date">11/29</div>
                                <div class="mark"></div>
                        </div>
                </div>





                <div class="r-ent">
                        <div class="nrec"></div>
                        <div class="title">

                                <a href="/bbs/Gossiping/M.1575014676.A.2A8.htm
l">[問卦] 看的最膩的月經文</a>

                        </div>
                        <div class="meta">
                                <div class="author">AsllaPiscu</div>
                                <div class="article-menu">

                                        <div class="trigger">&#x22ef;</div>
                                        <div class="dropdown">
                                                <div class="item"><a href="/bb
s/Gossiping/search?q=thread%3A%5B%E5%95%8F%E5%8D%A6%5D&#43;%E7%9C%8B%E7%9A%84%
E6%9C%80%E8%86%A9%E7%9A%84%E6%9C%88%E7%B6%93%E6%96%87">搜尋同標題文章</a></div>

                                                <div class="item"><a href="/bb
s/Gossiping/search?q=author%3AAsllaPiscu">搜尋看板內 AsllaPiscu 的文章</a></div>

                                        </div>

                                </div>
                                <div class="date">11/29</div>
                                <div class="mark"></div>
                        </div>
                </div>





                <div class="r-ent">
                        <div class="nrec"></div>
                        <div class="title">

                                <a href="/bbs/Gossiping/M.1575014698.A.3F2.htm
l">Re: [問卦] 知道多年來喊媽媽的不是親生媽媽怎麼調適</a>

                        </div>
                        <div class="meta">
                                <div class="author">pauljet</div>
                                <div class="article-menu">

                                        <div class="trigger">&#x22ef;</div>
                                        <div class="dropdown">
                                                <div class="item"><a href="/bb
s/Gossiping/search?q=thread%3A%5B%E5%95%8F%E5%8D%A6%5D&#43;%E7%9F%A5%E9%81%93%
E5%A4%9A%E5%B9%B4%E4%BE%86%E5%96%8A%E5%AA%BD%E5%AA%BD%E7%9A%84%E4%B8%8D%E6%98%
AF%E8%A6%AA%E7%94%9F%E5%AA%BD%E5%AA%BD%E6%80%8E%E9%BA%BC%E8%AA%BF%E9%81%A9">搜
尋同標題文章</a></div>

                                                <div class="item"><a href="/bb
s/Gossiping/search?q=author%3Apauljet">搜尋看板內 pauljet 的文章</a></div>

                                        </div>

                                </div>
                                <div class="date">11/29</div>
                                <div class="mark"></div>
                        </div>
                </div>



        <div class="r-list-sep"></div>




                <div class="r-ent">
                        <div class="nrec"></div>
                        <div class="title">

                                <a href="/bbs/Gossiping/M.1566347622.A.9C7.htm
l">[公告] 八卦板板規(2019.08.21)</a>

                        </div>
                        <div class="meta">
                                <div class="author">arsonlolita</div>
                                <div class="article-menu">

                                        <div class="trigger">&#x22ef;</div>
                                        <div class="dropdown">
                                                <div class="item"><a href="/bb
s/Gossiping/search?q=thread%3A%5B%E5%85%AC%E5%91%8A%5D&#43;%E5%85%AB%E5%8D%A6%
E6%9D%BF%E6%9D%BF%E8%A6%8F%282019.08.21%29">搜尋同標題文章</a></div>

                                                <div class="item"><a href="/bb
s/Gossiping/search?q=author%3Aarsonlolita">搜尋看板內 arsonlolita 的文章</a></div
>

                                        </div>

                                </div>
                                <div class="date"> 8/21</div>
                                <div class="mark">!</div>
                        </div>
                </div>




                <div class="r-ent">
                        <div class="nrec"><span class="hl f3">22</span></div>
                        <div class="title">

                                <a href="/bbs/Gossiping/M.1573743900.A.A37.htm
l">尋求9/23敦化南路(國泰綜合證卷前)行車紀錄影像</a>

                        </div>
                        <div class="meta">
                                <div class="author">zyx0809449</div>
                                <div class="article-menu">

                                        <div class="trigger">&#x22ef;</div>
                                        <div class="dropdown">
                                                <div class="item"><a href="/bb
s/Gossiping/search?q=thread%3A%E5%B0%8B%E6%B1%829%2F23%E6%95%A6%E5%8C%96%E5%8
D%97%E8%B7%AF%28%E5%9C%8B%E6%B3%B0%E7%B6%9C%E5%90%88%E8%AD%89%E5%8D%B7%E5%89%8
```

```
D%29%E8%A1%8C%E8%BB%8A%E7%B4%80%E9%8C%84%E5%BD%B1%E5%83%8F">搜尋同標題文章</a></
div>

                                                <div class="item"><a href="/bb
s/Gossiping/search?q=author%3Azyx0809449">搜尋看板內 zyx0809449 的文章</a></div>

                                        </div>

                                </div>
                                <div class="date">11/14</div>
                                <div class="mark">M</div>
                        </div>
                </div>






                <div class="r-ent">
                        <div class="nrec"><span class="hl f3">12</span></div>
                        <div class="title">

                                <a href="/bbs/Gossiping/M.1574212486.A.222.htm
l">[公告] 宣導禁止回文政問</a>

                        </div>
                        <div class="meta">
                                <div class="author">arsonlolita</div>
                                <div class="article-menu">

                                        <div class="trigger">&#x22ef;</div>
                                        <div class="dropdown">
                                                <div class="item"><a href="/bb
s/Gossiping/search?q=thread%3A%5B%E5%85%AC%E5%91%8A%5D&#43;%E5%AE%A3%E5%B0%8E%
E7%A6%81%E6%AD%A2%E5%9B%9E%E6%96%87%E6%94%BF%E5%95%8F">搜尋同標題文章</a></div>

                                                <div class="item"><a href="/bb
s/Gossiping/search?q=author%3Aarsonlolita">搜尋看板內 arsonlolita 的文章</a></div
```

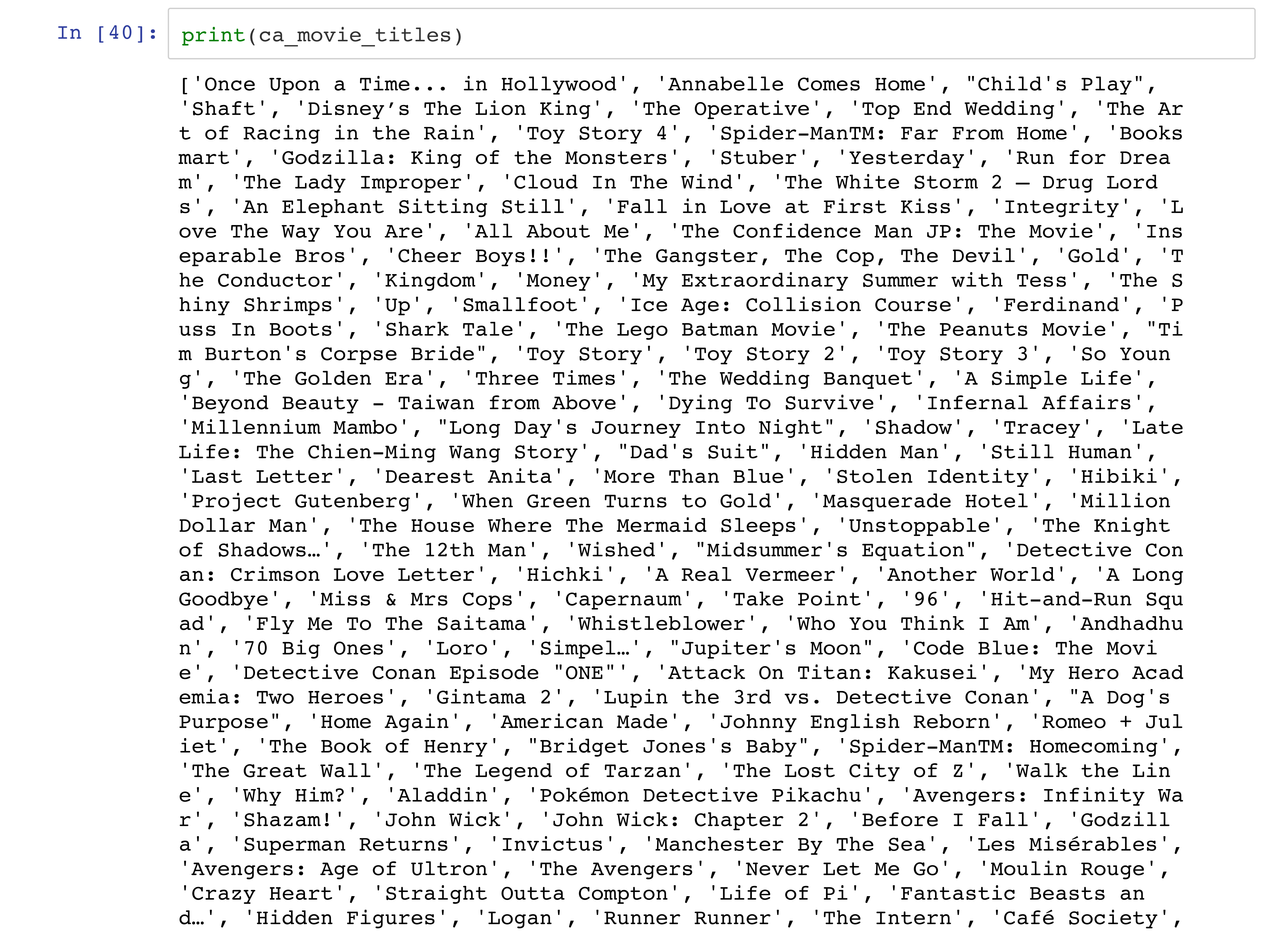
```
>

                                        </div>

                                </div>
                                <div class="date">11/20</div>
                                <div class="mark">M</div>
                        </div>
                </div>





                <div class="r-ent">
                        <div class="nrec"><span class="hl f2">7</span></div>
                        <div class="title">

                                <a href="/bbs/Gossiping/M.1574602085.A.66D.htm
l">[協尋] 11/20高市大順一路548號順明街行車記錄</a>

                        </div>
                        <div class="meta">
                                <div class="author">witty</div>
                                <div class="article-menu">

                                        <div class="trigger">&#x22ef;</div>
                                        <div class="dropdown">
                                                <div class="item"><a href="/bb
s/Gossiping/search?q=thread%3A%5B%E5%8D%94%E5%B0%8B%5D&#43;11%2F20%E9%AB%98%E
5%B8%82%E5%A4%A7%E9%A0%86%E4%B8%80%E8%B7%AF548%E8%99%9F%E9%A0%86%E6%98%8E%E8%A
1%97%E8%A1%8C%E8%BB%8A%E8%A8%98%E9%8C%84">搜尋同標題文章</a></div>

                                                <div class="item"><a href="/bb
s/Gossiping/search?q=author%3Awitty">搜尋看板內 witty 的文章</a></div>

                                        </div>
                                </div>
                                <div class="date">11/24</div>
                                <div class="mark"></div>
                        </div>
                </div>





                <div class="r-ent">
                        <div class="nrec"><span class="hl f2">8</span></div>
                        <div class="title">

                                <a href="/bbs/Gossiping/M.1572625135.A.19E.htm
l">[公告] 文殊起劍，十一月份置底閒聊區</a>

                        </div>
                        <div class="meta">
                                <div class="author">Bignana</div>
                                <div class="article-menu">

                                        <div class="trigger">&#x22ef;</div>
                                        <div class="dropdown">
                                                <div class="item"><a href="/bb
s/Gossiping/search?q=thread%3A%5B%E5%85%AC%E5%91%8A%5D&#43;%E6%96%87%E6%AE%8A%
E8%B5%B7%E5%8A%8D%EF%BC%8C%E5%8D%81%E4%B8%80%E6%9C%88%E4%BB%BD%E7%BD%AE%E5%BA%
95%E9%96%92%E8%81%8A%E5%8D%80">搜尋同標題文章</a></div>

                                                <div class="item"><a href="/bb
s/Gossiping/search?q=author%3ABignana">搜尋看板內 Bignana 的文章</a></div>

                                        </div>

                                </div>
                                <div class="date">11/02</div>
                                <div class="mark">M</div>
                        </div>
                </div>



        </div>


<div class="bbs-screen bbs-footer-message">本網站已依台灣網站內容分級規定處理。此區域
為限制級，未滿十八歲者不得瀏覽。</div>

</div>



<script>
  (function(i,s,o,g,r,a,m){i['GoogleAnalyticsObject']=r;i[r]=i[r]||function(){
  (i[r].q=i[r].q||[]).push(arguments)},i[r].l=1*new Date();a=s.createElement
(o),
  m=s.getElementsByTagName(o)[0];a.async=1;a.src=g;m.parentNode.insertBefore
(a,m)
  })(window,document,'script','https://www.google-analytics.com/analytics.j
s','ga');

  ga('create', 'UA-32365737-1', {
    cookieDomain: 'ptt.cc',
    legacyCookieDomain: 'ptt.cc'
  });
  ga('send', 'pageview');
</script>



<script src="//ajax.googleapis.com/ajax/libs/jquery/2.1.1/jquery.min.js"></scr
ipt>
<script src="//images.ptt.cc/bbs/v2.27/bbs.js"></script>

    </body>
</html>
```



```
In [38]: import requests
         from bs4 import BeautifulSoup

         response = requests.get("http://www.fantasy-sky.com/ContentList.aspx?section=002",
         cookies={'COOKIE_LANGUAGE': 'en'})
         soup = BeautifulSoup(response.text)
         movie_titles = [i.text for i in soup.select(".movies-name")]
         print(movie_titles)
```

```
['Once Upon a Time... in Hollywood', 'Annabelle Comes Home', "Child's Play",
'Shaft', 'Disney’s The Lion King', 'The Operative', 'Top End Wedding', 'The Ar
t of Racing in the Rain', 'Toy Story 4', 'Spider-ManTM: Far From Home', 'Books
mart', 'Godzilla: King of the Monsters', 'Stuber', 'Yesterday', 'Run for Drea
m', 'The Lady Improper', 'Cloud In The Wind', 'The White Storm 2 — Drug Lord
s', 'An Elephant Sitting Still', 'Fall in Love at First Kiss', 'Integrity', 'L
ove The Way You Are', 'All About Me', 'The Confidence Man JP: The Movie', 'Ins
eparable Bros', 'Cheer Boys!!', 'The Gangster, The Cop, The Devil', 'Gold', 'T
he Conductor', 'Kingdom', 'Money', 'My Extraordinary Summer with Tess', 'The S
hiny Shrimps']
```

# 隨堂練習：擷取所有華航機上電影清單



```
In [40]: print(ca_movie_titles)
```

```
['Once Upon a Time... in Hollywood', 'Annabelle Comes Home', "Child's Play",
'Shaft', 'Disney’s The Lion King', 'The Operative', 'Top End Wedding', 'The Ar
t of Racing in the Rain', 'Toy Story 4', 'Spider-ManTM: Far From Home', 'Books
mart', 'Godzilla: King of the Monsters', 'Stuber', 'Yesterday', 'Run for Drea
m', 'The Lady Improper', 'Cloud In The Wind', 'The White Storm 2 — Drug Lord
s', 'An Elephant Sitting Still', 'Fall in Love at First Kiss', 'Integrity', 'L
ove The Way You Are', 'All About Me', 'The Confidence Man JP: The Movie', 'Ins
eparable Bros', 'Cheer Boys!!', 'The Gangster, The Cop, The Devil', 'Gold', 'T
he Conductor', 'Kingdom', 'Money', 'My Extraordinary Summer with Tess', 'The S
hiny Shrimps', 'Up', 'Smallfoot', 'Ice Age: Collision Course', 'Ferdinand', 'P
uss In Boots', 'Shark Tale', 'The Lego Batman Movie', 'The Peanuts Movie', "Ti
m Burton's Corpse Bride", 'Toy Story', 'Toy Story 2', 'Toy Story 3', 'So Youn
g', 'The Golden Era', 'Three Times', 'The Wedding Banquet', 'A Simple Life',
'Beyond Beauty - Taiwan from Above', 'Dying To Survive', 'Infernal Affairs',
'Millennium Mambo', "Long Day's Journey Into Night", 'Shadow', 'Tracey', 'Late
Life: The Chien-Ming Wang Story', "Dad's Suit", 'Hidden Man', 'Still Human',
'Last Letter', 'Dearest Anita', 'More Than Blue', 'Stolen Identity', 'Hibiki',
'Project Gutenberg', 'When Green Turns to Gold', 'Masquerade Hotel', 'Million
Dollar Man', 'The House Where The Mermaid Sleeps', 'Unstoppable', 'The Knight
of Shadows…', 'The 12th Man', 'Wished', "Midsummer's Equation", 'Detective Con
an: Crimson Love Letter', 'Hichki', 'A Real Vermeer', 'Another World', 'A Long
Goodbye', 'Miss & Mrs Cops', 'Capernaum', 'Take Point', '96', 'Hit-and-Run Squ
ad', 'Fly Me To The Saitama', 'Whistleblower', 'Who You Think I Am', 'Andhadhu
n', '70 Big Ones', 'Loro', 'Simpel…', "Jupiter's Moon", 'Code Blue: The Movi
e', 'Detective Conan Episode "ONE"', 'Attack On Titan: Kakusei', 'My Hero Acad
emia: Two Heroes', 'Gintama 2', 'Lupin the 3rd vs. Detective Conan', "A Dog's
Purpose", 'Home Again', 'American Made', 'Johnny English Reborn', 'Romeo + Jul
iet', 'The Book of Henry', "Bridget Jones's Baby", 'Spider-ManTM: Homecoming',
'The Great Wall', 'The Legend of Tarzan', 'The Lost City of Z', 'Walk the Lin
e', 'Why Him?', 'Aladdin', 'Pokémon Detective Pikachu', 'Avengers: Infinity Wa
r', 'Shazam!', 'John Wick', 'John Wick: Chapter 2', 'Before I Fall', 'Godzill
a', 'Superman Returns', 'Invictus', 'Manchester By The Sea', 'Les Misérables',
'Avengers: Age of Ultron', 'The Avengers', 'Never Let Me Go', 'Moulin Rouge',
'Crazy Heart', 'Straight Outta Compton', 'Life of Pi', 'Fantastic Beasts an
d…', 'Hidden Figures', 'Logan', 'Runner Runner', 'The Intern', 'Café Society',
```

# 隨堂練習：找出華航機上最高評等的電影

```
In [44]: print(ca_movie_titles[best_movie_index])
```

```
The Shawshank Redemption
```

# 瀏覽器自動化

## 在研究如何使 `get_movie_data()` 更方便的過程中我們做了幾個動作

1. 前往 [https://www.imdb.com/ (https://www.imdb.com/)](https://www.imdb.com/) 首頁
2. 輸入電影名稱
3. 點選搜尋
4. 點選 Movie 分類標籤
5. 點選相似度最高的搜尋結果

這些操作可以利用 `selenium` 來自動化！

# 什麼是 Selenium

- Selenium 是瀏覽器自動化測試的解決方案
- Python 透過 Selenium WebDriver 呼叫瀏覽器驅動程式，再由瀏覽器驅動程式去呼叫瀏覽器
- 對 Google Chrome 與 Mozilla Firefox 兩個主流瀏覽器的支援最好

# Selenium 環境設定： Chrome

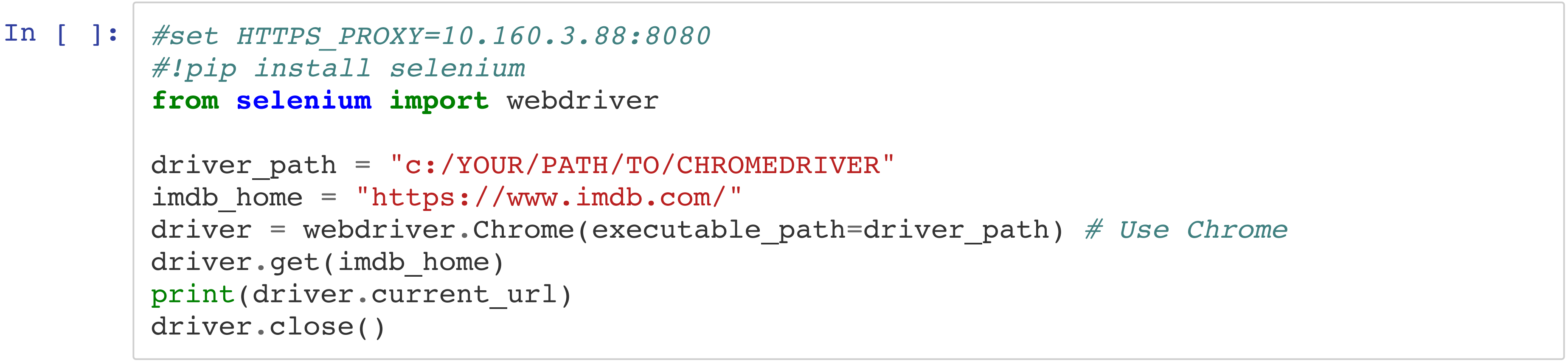
- 前往 Chrome 官方網站 (https://www.google.com/chrome/)下載最新版的瀏覽器
- 下載最新版的瀏覽器驅動程式 ChromeDriver (http://chromedriver.chromium.org/)
- 下載完成以後解壓縮在熟悉路徑讓後續指派較為方便

# Selenium 環境設定： Firefox

- 前往 Firefox 官方網站 (https://www.mozilla.org/zh-TW/firefox/new/) 下載最新版的瀏覽器
- 下載最新版的瀏覽器驅動程式 geckodriver (https://github.com/mozilla/geckodriver/releases)
- 下載完成以後解壓縮在熟悉路徑讓後續指派較為方便

# 測試 Chrome 是否設定完成

用程式碼透過 ChromeDriver 操控 Chrome 瀏覽器前往 IMDB 首頁並將首頁的網址印出再關閉瀏覽器

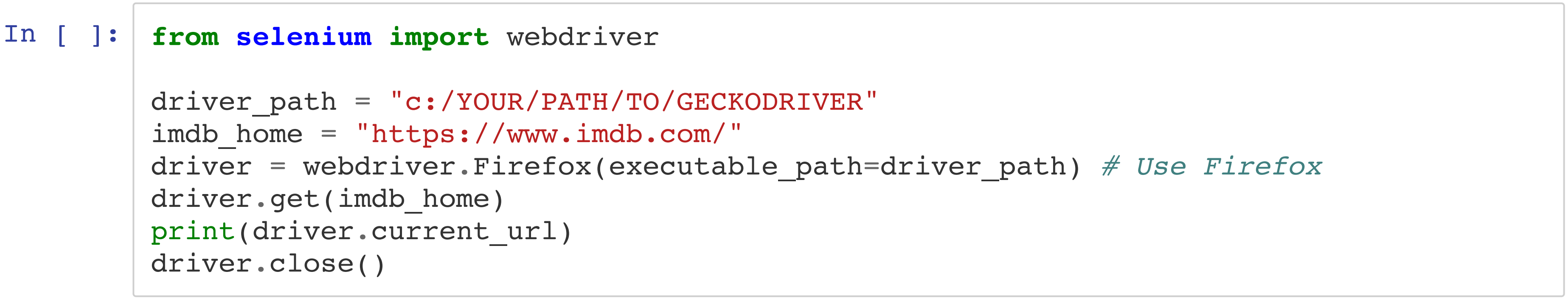


In [ ]:

```
#set HTTPS_PROXY=10.160.3.88:8080
#!pip install selenium
from selenium import webdriver

driver_path = "c:/YOUR/PATH/TO/CHROMEDRIVER"
imdb_home = "https://www.imdb.com/"
driver = webdriver.Chrome(executable_path=driver_path) # Use Chrome
driver.get(imdb_home)
print(driver.current_url)
driver.close()
```

## 測試 Firefox 是否設定完成

用程式碼透過 geckodriver 操控 Firefox 瀏覽器前往 IMDB 首頁並將首頁的網址印出再關閉瀏覽器



```
from selenium import webdriver

driver_path = "c:/YOUR/PATH/TO/GECKODRIVER"
imdb_home = "https://www.imdb.com/"
driver = webdriver.Firefox(executable_path=driver_path) # Use Firefox
driver.get(imdb_home)
print(driver.current_url)
driver.close()
```

# 常使用的 `driver` 方法、屬性

- `driver.get()`：前往指定網址
- `driver.find_element_by_css_selector()`：定位搜尋欄位、搜尋按鈕與搜尋結果連結（單數）
- `driver.find_elements_by_css_selector()`：定位搜尋欄位、搜尋按鈕與搜尋結果連結（複數）
- `driver.find_element_by_xpath()`：定位搜尋欄位、搜尋按鈕與搜尋結果連結（單數）
- `driver.find_elements_by_xpath()`：定位搜尋欄位、搜尋按鈕與搜尋結果連結（複數）
- `driver.current_url`：取得當下瀏覽器的網址

# 幫助檢視 XPath 的 Chrome 外掛

XPath Helper (https://chrome.google.com/webstore/detail/xpath-helper/hgimnogjllphhhkhlmebbmlgjoejdpjl)

## XPath Helper (https://chrome.google.com/webstore/detail/xpath-helper/hgimnogjllphhhkhlmebbmlgjoejdpjl) 的使用方法

- 點選 XPath Helper 的外掛圖示
- 留意 XPath Helper 介面左邊的 XPath 與右邊被定位到的資料
- 按住 shift 鍵移動滑鼠到想要定位的元素
- 試著縮減 XPath，從最前面開始刪減並置換為 //

## 以 Avengers: Endgame (2019) (https://www.imdb.com/title/tt4154796) 示範 XPath Helper (https://chrome.google.com/webstore/detail/xpath-helper/hgimnogjllphhhkhlmebbmlgjoejdpjl) 的使用方法

- 電影名稱
- 電影海報
- 評分
- 劇情類型
- 演員陣容

## 常使用的 element 方法、屬性

- `element.send_keys()`：輸入文字
- `element.click()`：按下搜尋按鈕與連結
- `element.text`：取出標記中的文字值
- `element.get_attribute(ATTR)`：取出標記中的指定屬性

## 隨堂練習：以 selenium 實作 get_movie_data(movie_title)

```
In [ ]: get_movie_data("Avengers: Endgame (2019)")
```

# 隨堂練習：以 selenium 擷取四部復仇者聯盟的電影資訊

```
avengers_movies = ["The Avengers (2012)", "Avengers: Age of Ultron (2015)", "Avengers: Infinity War (2018)", "Avengers: Endgame (2019)"]
```

In [ ]:

```
print(avengers_movie_data)
```

# 將擷取的電影資訊匯出

In [ ]:

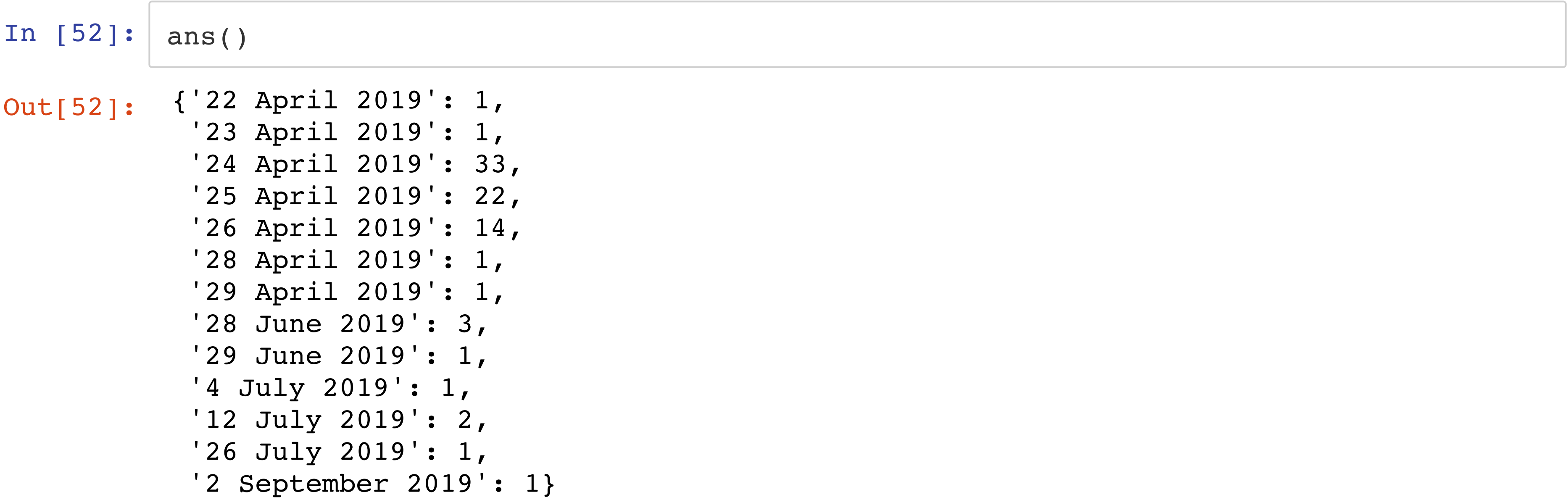
```python
import json

with open("avengers.json", "w") as f:
    json.dump(avengers_movie_data, f)
```

## 延伸閱讀

- Requests: HTTP for Humans (http://docs.python-requests.org/en/master/)
- Beautiful Soup Documentation (https://www.crummy.com/software/BeautifulSoup/bs4/doc/#)
- Selenium with Python (https://selenium-python.readthedocs.io/)
- Python 與網頁資料擷取 - DataInPoint (https://medium.com/datainpoint/web-scraping-with-python/home)

# 作業

擷取 Avengers: Endgame (2019) (https://www.imdb.com/title/tt4154796/releaseinfo) 的上映日期列表，最多的上映日期為哪一天？有幾個國家在那天上映？



```
In [52]: ans()
```

```
Out[52]: {'22 April 2019': 1,
          '23 April 2019': 1,
          '24 April 2019': 33,
          '25 April 2019': 22,
          '26 April 2019': 14,
          '28 April 2019': 1,
          '29 April 2019': 1,
          '28 June 2019': 3,
          '29 June 2019': 1,
          '4 July 2019': 1,
          '12 July 2019': 2,
          '26 July 2019': 1,
          '2 September 2019': 1}
```