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 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

請求資料

- 使用 Quick JavaScript Switcher (https://chrome.google.com/webstore/detail/quick-javascript-switcher/geddoclleiomckbhadiaipdggiiccfje) 與 Chrome 開發者工具判斷網頁資料類型
- 以 `requests` 或 `selenium` 發送 HTTP 請求獲得網頁資料

常用的 `requests` 方法、屬性

- `requests.get()`：進行 GET 請求
- `r.status_code`：查看 HTTP 狀態碼
- `r.json()`：將回應直接轉換為 Python 的資料結構（`list` 或 `dict`）
- `r.content`：將回應轉換為 `bytes`
- `r.text`：將回應轉換為 `str`

解析資料

- 如果資料是 JSON 格式：以 `requests` 獲取後可直接以 Python 資料結構解析
- 如果資料是 HTML 格式：以 `bs4`、`pyquery` 或 `selenium` 搭配 CSS Selector/XPath 解析

擷取 JSON 格式網頁資料

JSON 格式網頁資料範例

- 空氣品質指標(AQI) (https://opendata.epa.gov.tw/ws/Data/AQI/?$format=json)
- 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

- `requests.get()`
- `r.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 [1]:

```
import requests

teams_url = "http://data.nba.net/prod/v2/2019/teams.json"
r = requests.get(teams_url)
teams = r.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': 'Shanghai Shark
s', 'urlName': 'shanghai_sharks', 'teamShortName': 'Shanghai', 'confName': 'In
tl', 'divName': ''}, {'isNBAFranchise': False, 'isAllStar': False, 'city': 'Sy
dney', 'altCityName': 'Sydney', 'fullName': 'Sydney Kings', 'tricode': 'SYD',
'teamId': '15015', 'nickname': 'Kings', 'urlName': 'sydkings', 'teamShortNam
e': 'Sydney', 'confName': 'Intl', 'divName': ''}, {'isNBAFranchise': False, 'i
sAllStar': True, 'city': 'Team', 'altCityName': 'Team', 'fullName': 'All-Star
s', 'tricode': 'EST', 'teamId': '1699999999', 'nickname': 'All-Stars', 'urlNam
e': 'assn_away', 'confName': 'East', 'divName': 'East'}, {'isNBAFranchise': Fa
lse, 'isAllStar': True, 'city': 'Team', 'altCityName': 'Team', 'fullName': 'Al
l-Stars', 'tricode': 'WST', 'teamId': '1699999998', 'nickname': 'All-Stars',
'urlName': 'assn_home', 'confName': 'West', 'divName': 'West'}, {'isNBAFranchi
se': False, 'isAllStar': True, 'city': 'Team Giannis', 'altCityName': 'Team Gi
annis', 'fullName': 'Team Giannis', 'tricode': 'GNS', 'teamId': '1610616833',
'nickname': 'Team Giannis', 'urlName': 'team_giannis', 'teamShortName': 'Team
Giannis', 'confName': 'East', 'divName': 'East'}, {'isNBAFranchise': False, 'i
sAllStar': True, 'city': 'Team LeBron', 'altCityName': 'Team LeBron', 'fullNam
e': 'Team LeBron', 'tricode': 'LBN', 'teamId': '1610616834', 'nickname': 'Team
LeBron', 'urlName': 'team_lebron', 'teamShortName': 'Team LeBron', 'confName':
'West', 'divName': 'West'}, {'isNBAFranchise': True, 'isAllStar': False, 'cit
y': 'Toronto', 'altCityName': 'Toronto', 'fullName': 'Toronto Raptors', 'trico
de': 'TOR', 'teamId': '1610612761', 'nickname': 'Raptors', 'urlName': 'raptor
s', 'teamShortName': 'Toronto', 'confName': 'East', 'divName': 'Atlantic'},
{'isNBAFranchise': False, 'isAllStar': True, 'city': 'USA', 'altCityName': 'US

A', 'fullName': 'USA', 'tricode': 'USA', 'teamId': '1610616843', 'nickname':
'USA', 'urlName': 'usa', 'teamShortName': 'USA', 'confName': 'East', 'divNam
e': 'East'}, {'isNBAFranchise': True, 'isAllStar': False, 'city': 'Utah', 'alt
CityName': 'Utah', 'fullName': 'Utah Jazz', 'tricode': 'UTA', 'teamId': '16106
12762', 'nickname': 'Jazz', 'urlName': 'jazz', 'teamShortName': 'Utah', 'confN
ame': 'West', 'divName': 'Northwest'}, {'isNBAFranchise': True, 'isAllStar': F
alse, 'city': 'Washington', 'altCityName': 'Washington', 'fullName': 'Washingt
on Wizards', 'tricode': 'WAS', 'teamId': '1610612764', 'nickname': 'Wizards',
'urlName': 'wizards', 'teamShortName': 'Washington', 'confName': 'East', 'divN
ame': 'Southeast'}, {'isNBAFranchise': False, 'isAllStar': True, 'city': 'Worl
d', 'altCityName': 'World', 'fullName': 'World', 'tricode': 'WLD', 'teamId':
'1610616844', 'nickname': 'World', 'urlName': 'world', 'teamShortName': 'Worl
d', 'confName': 'East', 'divName': 'East'}], 'africa': [{'isNBAFranchise': Fal
se, 'isAllStar': False, 'city': 'Team', 'altCityName': 'Team', 'fullName': 'Te
am USA', 'tricode': 'USA', 'teamId': '22', 'nickname': 'USA', 'urlName': 'nhs_
usa', 'teamShortName': 'USA', 'confName': '', 'divName': ''}, {'isNBAFranchis
e': False, 'isAllStar': False, 'city': 'Team', 'altCityName': 'Team', 'fullNam
e': 'Team World', 'tricode': 'WLD', 'teamId': '21', 'nickname': 'World', 'urlN
ame': 'nhs_world', 'teamShortName': 'World', 'confName': '', 'divName': ''}],
'sacramento': [{'isNBAFranchise': True, 'isAllStar': False, 'city': 'Golden St
ate', 'altCityName': 'Golden State', 'fullName': 'Golden State Warriors', 'tri
code': 'GSW', 'teamId': '1610612744', 'nickname': 'Warriors', 'urlName': 'warr
iors', 'teamShortName': 'Golden State', 'confName': 'Sacramento', 'divName':
''}, {'isNBAFranchise': True, 'isAllStar': False, 'city': 'Los Angeles', 'altC
ityName': 'Los Angeles Lakers', 'fullName': 'Los Angeles Lakers', 'tricode':
'LAL', 'teamId': '1610612747', 'nickname': 'Lakers', 'urlName': 'lakers', 'tea
mShortName': 'L.A. Lakers', 'confName': 'Sacramento', 'divName': ''}, {'isNBAF
ranchise': True, 'isAllStar': False, 'city': 'Miami', 'altCityName': 'Miami',
'fullName': 'Miami Heat', 'tricode': 'MIA', 'teamId': '1610612748', 'nicknam
e': 'Heat', 'urlName': 'heat', 'teamShortName': 'Miami', 'confName': 'Sacramen
to', 'divName': ''}, {'isNBAFranchise': True, 'isAllStar': False, 'city': 'Sac
ramento', 'altCityName': 'Sacramento', 'fullName': 'Sacramento Kings', 'tricod
e': 'SAC', 'teamId': '1610612758', 'nickname': 'Kings', 'urlName': 'kings', 't
eamShortName': 'Sacramento', 'confName': 'Sacramento', 'divName': ''}], 'vega
s': [{'isNBAFranchise': True, 'isAllStar': False, 'city': 'Atlanta', 'altCityN
ame': 'Atlanta', 'fullName': 'Atlanta Hawks', 'tricode': 'ATL', 'teamId': '161
0612737', 'nickname': 'Hawks', 'urlName': 'hawks', 'teamShortName': 'Atlanta',
'confName': 'summer', 'divName': ''}, {'isNBAFranchise': True, 'isAllStar': Fa

lse, 'city': 'Boston', 'altCityName': 'Boston', 'fullName': 'Boston Celtics',
'tricode': 'BOS', 'teamId': '1610612738', 'nickname': 'Celtics', 'urlName': 'c
eltics', 'teamShortName': 'Boston', 'confName': 'summer', 'divName': ''}, {'is
NBAFranchise': True, 'isAllStar': False, 'city': 'Brooklyn', 'altCityName': 'B
rooklyn', 'fullName': 'Brooklyn Nets', 'tricode': 'BKN', 'teamId': '161061275
1', 'nickname': 'Nets', 'urlName': 'nets', 'teamShortName': 'Brooklyn', 'confN
ame': 'summer', 'divName': ''}, {'isNBAFranchise': True, 'isAllStar': False,
'city': 'Charlotte', 'altCityName': 'Charlotte', 'fullName': 'Charlotte Hornet
s', 'tricode': 'CHA', 'teamId': '1610612766', 'nickname': 'Hornets', 'urlNam
e': 'hornets', 'teamShortName': 'Charlotte', 'confName': 'summer', 'divName':
''}, {'isNBAFranchise': True, 'isAllStar': False, 'city': 'Chicago', 'altCityN
ame': 'Chicago', 'fullName': 'Chicago Bulls', 'tricode': 'CHI', 'teamId': '161
0612741', 'nickname': 'Bulls', 'urlName': 'bulls', 'teamShortName': 'Chicago',
'confName': 'summer', 'divName': ''}, {'isNBAFranchise': False, 'isAllStar': F
alse, 'city': 'China', 'altCityName': 'China', 'fullName': 'Team China', 'tric
ode': 'CHN', 'teamId': '45', 'nickname': 'China', 'urlName': 'china', 'teamSho
rtName': 'China', 'confName': 'summer', 'divName': ''}, {'isNBAFranchise': Tru
e, 'isAllStar': False, 'city': 'Cleveland', 'altCityName': 'Cleveland', 'fullN
ame': 'Cleveland Cavaliers', 'tricode': 'CLE', 'teamId': '1610612739', 'nickna
me': 'Cavaliers', 'urlName': 'cavaliers', 'teamShortName': 'Cleveland', 'confN
ame': 'summer', 'divName': ''}, {'isNBAFranchise': False, 'isAllStar': False,
'city': 'Croatia', 'altCityName': 'Croatia', 'fullName': 'Team Croatia', 'tric
ode': 'CRO', 'teamId': '70', 'nickname': 'Croatia', 'urlName': 'croatia', 'tea
mShortName': 'Croatia', 'confName': 'summer', 'divName': ''}, {'isNBAFranchis
e': True, 'isAllStar': False, 'city': 'Dallas', 'altCityName': 'Dallas', 'full
Name': 'Dallas Mavericks', 'tricode': 'DAL', 'teamId': '1610612742', 'nicknam
e': 'Mavericks', 'urlName': 'mavericks', 'teamShortName': 'Dallas', 'confNam
e': 'summer', 'divName': ''}, {'isNBAFranchise': True, 'isAllStar': False, 'ci
ty': 'Denver', 'altCityName': 'Denver', 'fullName': 'Denver Nuggets', 'tricod
e': 'DEN', 'teamId': '1610612743', 'nickname': 'Nuggets', 'urlName': 'nugget
s', 'teamShortName': 'Denver', 'confName': 'summer', 'divName': ''}, {'isNBAFr
anchise': True, 'isAllStar': False, 'city': 'Detroit', 'altCityName': 'Detroi
t', 'fullName': 'Detroit Pistons', 'tricode': 'DET', 'teamId': '1610612765',
'nickname': 'Pistons', 'urlName': 'pistons', 'teamShortName': 'Detroit', 'conf
Name': 'summer', 'divName': ''}, {'isNBAFranchise': True, 'isAllStar': False,
'city': 'Golden State', 'altCityName': 'Golden State', 'fullName': 'Golden Sta
te Warriors', 'tricode': 'GSW', 'teamId': '1610612744', 'nickname': 'Warrior
s', 'urlName': 'warriors', 'teamShortName': 'Golden State', 'confName': 'summe

r', 'divName': ''}, {'isNBAFranchise': True, 'isAllStar': False, 'city': 'Hous
ton', 'altCityName': 'Houston', 'fullName': 'Houston Rockets', 'tricode': 'HO
U', 'teamId': '1610612745', 'nickname': 'Rockets', 'urlName': 'rockets', 'team
ShortName': 'Houston', 'confName': 'summer', 'divName': ''}, {'isNBAFranchis
e': True, 'isAllStar': False, 'city': 'Indiana', 'altCityName': 'Indiana', 'fu
llName': 'Indiana Pacers', 'tricode': 'IND', 'teamId': '1610612754', 'nicknam
e': 'Pacers', 'urlName': 'pacers', 'teamShortName': 'Indiana', 'confName': 'su
mmer', 'divName': ''}, {'isNBAFranchise': True, 'isAllStar': False, 'city': 'L
A', 'altCityName': 'LA Clippers', 'fullName': 'LA Clippers', 'tricode': 'LAC',
'teamId': '1610612746', 'nickname': 'Clippers', 'urlName': 'clippers', 'teamSh
ortName': 'LA Clippers', 'confName': 'summer', 'divName': ''}, {'isNBAFranchis
e': True, 'isAllStar': False, 'city': 'Los Angeles', 'altCityName': 'Los Angel
es Lakers', 'fullName': 'Los Angeles Lakers', 'tricode': 'LAL', 'teamId': '161
0612747', 'nickname': 'Lakers', 'urlName': 'lakers', 'teamShortName': 'L.A. La
kers', 'confName': 'summer', 'divName': ''}, {'isNBAFranchise': True, 'isAllSt
ar': False, 'city': 'Memphis', 'altCityName': 'Memphis', 'fullName': 'Memphis
Grizzlies', 'tricode': 'MEM', 'teamId': '1610612763', 'nickname': 'Grizzlies',
'urlName': 'grizzlies', 'teamShortName': 'Memphis', 'confName': 'summer', 'div
Name': ''}, {'isNBAFranchise': True, 'isAllStar': False, 'city': 'Miami', 'alt
CityName': 'Miami', 'fullName': 'Miami Heat', 'tricode': 'MIA', 'teamId': '161
0612748', 'nickname': 'Heat', 'urlName': 'heat', 'teamShortName': 'Miami', 'co
nfName': 'summer', 'divName': ''}, {'isNBAFranchise': True, 'isAllStar': Fals
e, 'city': 'Milwaukee', 'altCityName': 'Milwaukee', 'fullName': 'Milwaukee Buc
ks', 'tricode': 'MIL', 'teamId': '1610612749', 'nickname': 'Bucks', 'urlName':
'bucks', 'teamShortName': 'Milwaukee', 'confName': 'summer', 'divName': ''},
{'isNBAFranchise': True, 'isAllStar': False, 'city': 'Minnesota', 'altCityNam
e': 'Minnesota', 'fullName': 'Minnesota Timberwolves', 'tricode': 'MIN', 'team
Id': '1610612750', 'nickname': 'Timberwolves', 'urlName': 'timberwolves', 'tea
mShortName': 'Minnesota', 'confName': 'summer', 'divName': ''}, {'isNBAFranchi
se': True, 'isAllStar': False, 'city': 'New Orleans', 'altCityName': 'New Orle
ans', 'fullName': 'New Orleans Pelicans', 'tricode': 'NOP', 'teamId': '1610612
740', 'nickname': 'Pelicans', 'urlName': 'pelicans', 'teamShortName': 'New Orl
eans', 'confName': 'summer', 'divName': ''}, {'isNBAFranchise': True, 'isAllSt
ar': False, 'city': 'New York', 'altCityName': 'New York', 'fullName': 'New Yo
rk Knicks', 'tricode': 'NYK', 'teamId': '1610612752', 'nickname': 'Knicks', 'u
rlName': 'knicks', 'teamShortName': 'New York', 'confName': 'summer', 'divNam
e': ''}, {'isNBAFranchise': True, 'isAllStar': False, 'city': 'Oklahoma City',
'altCityName': 'Oklahoma City', 'fullName': 'Oklahoma City Thunder', 'tricod

隨堂練習：2019-2020 球季 NBA 有幾支球隊?

In [3]:
```
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 [5]:
```
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 HTML 格式網頁資料

擷取 HTML 格式網頁資料步驟

- `requests.get()`
- `r.text`
- 以 `bs4` 或 `pyquery` 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 CSS 選擇的 Chrome 外掛

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

SelectorGadget (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

使用 bs4 或 pyquery 解析網頁資料

```
from bs4 import BeautifulSoup
from pyquery import PyQuery as pq
```

常用的 bs4 方法、屬性

- `BeautifulSoup()`：創建 soup 類別
- `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 Selectors 的資料
- `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

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

In [8]:
```
import requests
from bs4 import BeautifulSoup

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

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

In [9]:
```
print(soup.find("h1"))
print(type(soup.find("h1")))
print(soup.find("h1").text)
```

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

In [10]:

```
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"))
```

```
98
<img alt="Avengers: Endgame Poster" src="https://m.media-amazon.com/images/M/M
V5BMTc5MDE2ODcwNV5BMl5BanBnXkFtZTgwMzI2NzQ2NzM@._V1_UX182_CR0,0,182,268_AL_.jp
g" title="Avengers: Endgame Poster"/>
Avengers: Endgame Poster
https://m.media-amazon.com/images/M/MV5BMTc5MDE2ODcwNV5BMl5BanBnXkFtZTgwMzI2Nz
Q2NzM@._V1_UX182_CR0,0,182,268_AL_.jpg
```

In [11]:

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

```
[<span itemprop="ratingValue">8.6</span>]
8.6
```

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 [13]:
```
for g in genre:
    print(g.text)
```

```
Action
Adventure
Sci-Fi
```

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 [15]:
```
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']
```

常用的 pyquery 方法、屬性

- `PyQuery()`：創建 `PyQuery` 類別
- `d("YOUR-CSS-SELECTOR")`：尋找所有符合 CSS Selectors 的資料
- `d("YOUR-CSS-SELECTOR").items()`：回傳所有符合 CSS Selectors 的文字與屬性
- `elem.text()`：取出標記中的文字值
- `elem.attr(ATTR)`：取出標記中的指定屬性

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

In [16]:

```
import requests
from pyquery import PyQuery as pq

r = requests.get("https://www.imdb.com/title/tt4154796")
html_doc = r.text
d = pq(html_doc)
print(type(d))
```

```
<class 'pyquery.pyquery.PyQuery'>
```

In [17]:

```
print(d("h1"))
print(type(d("h1")))
for i in d("h1").items():
    print(i.text())
```

```
<h1 class="">Avengers: Endgame <span id="titleYear">(<a href="/year/2019/">201
9</a>)</span>            </h1>

<class 'pyquery.pyquery.PyQuery'>
Avengers: Endgame (2019)
```

In [18]:
```
for i in d(".poster img").items():
    print(i.attr("src"))
```

```
https://m.media-amazon.com/images/M/MV5BMTc5MDE2ODcwNV5BMl5BanBnXkFtZTgwMzI2Nz
Q2NzM@._V1_UX182_CR0,0,182,268_AL_.jpg
```

In [19]:
```
for i in d("strong span").items():
    print(float(i.text()))
```

```
8.6
```

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

In [21]:
```
print(genre)
```

```
['Action', 'Adventure', 'Sci-Fi']
```

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

In [23]:
```
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 [25]:
```
get_movie_data("https://www.imdb.com/title/tt4154796")
```

Out[25]:
```
{'movieTitle': 'Avengers: Endgame (2019)',
 'moviePoster': 'https://m.media-amazon.com/images/M/MV5BMTc5MDE2ODcwNV5BMl5Ba
nBnXkFtZTgwMzI2NzQ2NzM@._V1_UX182_CR0,0,182,268_AL_.jpg',
 'movieRating': 8.6,
 'movieGenre': ['Action', 'Adventure', 'Sci-Fi'],
 '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！
- 以 `urllib.parse.quote_plus()` 製作 query string

In [26]:
```
from urllib.parse import quote_plus

query_str = quote_plus("Avengers: Endgame (2019)")
print(query_str)
```

```
Avengers%3A+Endgame+%282019%29
```

In [27]:
```
query_str = quote_plus("Avengers: Endgame (2019)")
query_url = "https://www.imdb.com/find?q={}&s=tt&ttype=ft&ref_=fn_ft".format(query
_str)
print(query_url)
```

```
https://www.imdb.com/find?q=Avengers%3A+Endgame+%282019%29&s=tt&ttype=ft&ref_=
fn_ft
```

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

In [28]:

```
import requests
from pyquery import PyQuery as pq
from urllib.parse import quote_plus

def get_movie_data(movie_title):
    query_str = quote_plus(movie_title)
    query_url = "https://www.imdb.com/find?q={}&s=tt&ttype=ft&ref_=fn_ft".format(q
uery_str)
    r = requests.get(query_url)
    d = pq(r.text)
    search_results = [i.attr("href") for i in d(".result_text a").items()]
    movie_url = "https://www.imdb.com" + search_results[0]
    r = requests.get(movie_url)
    d = pq(r.text)
    movie_title = [i.text().replace("\xa0", " ") for i in d("h1").items()][0]
    movie_poster = [i.attr("src") for i in d(".poster img").items()][0]
    movie_rating = [float(i.text()) for i in d("strong span").items()][0]
    movie_genre = [i.text() for i in d(".subtext a").items()]
    movie_genre.pop()
    movie_cast = [i.text() for i in d(".primary_photo+ td a").items()]
    movie_data = {
        "movieTitle": movie_title,
        "moviePoster": movie_poster,
        "movieRating": movie_rating,
        "movieGenre": movie_genre,
        "movieCast": movie_cast
    }
    return movie_data
```

In [29]:
```
get_movie_data("Avengers: Endgame (2019)")
```

Out[29]:
```
{'movieTitle': 'Avengers: Endgame (2019)',
 'moviePoster': 'https://m.media-amazon.com/images/M/MV5BMTc5MDE2ODcwNV5BMl5Ba
nBnXkFtZTgwMzI2NzQ2NzM@._V1_UX182_CR0,0,182,268_AL_.jpg',
 'movieRating': 8.6,
 'movieGenre': ['Action', 'Adventure', 'Sci-Fi'],
 '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 [30]:
```
import requests

r = requests.get("https://www.ptt.cc/bbs/Gossiping/index.html")
print(r.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.js','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 [31]:

```
import requests
from bs4 import BeautifulSoup

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

```
['阿拉丁', '狗狗的旅程', 'MIB星際戰警：跨國行動', '名偵探皮卡丘', 'X戰警：黑鳳凰', '火箭人', '寵物當家2', '捍衛任務3：全面開戰', '復仇者聯盟4：終局之戰', '沙贊！', '庫爾斯克號：深海救援', '阿波羅11號', '紅色密令', '斯德哥爾摩症候', '魔力女聲', '大象席地而坐', '廉政風雲：煙幕', '拾芳', '老大人', '最是橙黃橘綠時', '後勁：王建民', '神探蒲松齡', '半世界', '飛翔吧！埼玉', '地球最後的夜晚', '看不見的旋律', '漫長的告別', '警察小姐', '90分鐘末日倒數', '逆轉劫局', '上流世界', '七個會議', '飆風特警隊', '96', '別問我是誰']
```

幫助檢視 cookies 的 Chrome 外掛

EditThisCookie
(https://chrome.google.com/webstore/detail/editthiscookie/fngmhnnpilhplaeedifhccceomclg

In [32]:
```
import requests

r = requests.get("https://www.ptt.cc/bbs/Gossiping/index.html", cookies={'over18':
'1'})
print(r.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-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 id="topbar-container">
        <div id="topbar" class="bbs-content">
                <a id="logo" href="/bbs/">批踢踢實業坊</a>
                <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
39024.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"><span class="hl f2">4</span></div>
                        <div class="title">

                                <a href="/bbs/Gossiping/M.1569265925.A.D44.htm
l">[問卦] 80歲阿嬤看返校會不會太刺激</a>

                        </div>
                        <div class="meta">
                                <div class="author">shrape</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;80%E6%AD%B2%E9%98%B
F%E5%AC%A4%E7%9C%8B%E8%BF%94%E6%A0%A1%E6%9C%83%E4%B8%8D%E6%9C%83%E5%A4%AA%E5%8
8%BA%E6%BF%80">搜尋同標題文章</a></div>

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

                                        </div>

                                </div>
                                <div class="date"> 9/24</div>
                                <div class="mark"></div>
                        </div>
                </div>




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

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

                                <a href="/bbs/Gossiping/M.1569266016.A.BE3.htm
l">[問卦] 基努李維好像我，我的錯覺? </a>

                        </div>
                        <div class="meta">
                                <div class="author">Dinenger</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%9F%BA%E5%8A%AA%
E6%9D%8E%E7%B6%AD%E5%A5%BD%E5%83%8F%E6%88%91%EF%BC%8C%E6%88%91%E7%9A%84%E9%8C%
AF%E8%A6%BA%EF%BC%9F">搜尋同標題文章</a></div>

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

                                        </div>

                                </div>
                                <div class="date"> 9/24</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.1569266069.A.E55.htm
```

```
l">[問卦] 女森灑花 囊森要灑什麼? </a>

                        </div>
                        <div class="meta">
                                <div class="author">jason83813</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%A5%B3%E6%A3%AE%
E7%81%91%E8%8A%B1&#43;%E5%9B%8A%E6%A3%AE%E8%A6%81%E7%81%91%E4%BB%80%E9%BA%BC%E
F%BC%9F">搜尋同標題文章</a></div>

                                                <div class="item"><a href="/bb
s/Gossiping/search?q=author%3Ajason83813">搜尋看板內 jason83813 的文章</a></div>

                                        </div>

                                </div>
                                <div class="date"> 9/24</div>
                                <div class="mark"></div>
                        </div>
                </div>




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

                                <a href="/bbs/Gossiping/M.1569266507.A.0AF.htm
l">Re: [問卦] 活著是為了什麼? </a>

                        </div>
                        <div class="meta">
```

```
                                <div class="author">killerchi</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;%E6%B4%BB%E8%91%97%
E6%98%AF%E7%82%BA%E4%BA%86%E4%BB%80%E9%BA%BC%EF%BC%9F">搜尋同標題文章</a></div>

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

                                        </div>

                                </div>
                                <div class="date"> 9/24</div>
                                <div class="mark"></div>
                        </div>
                </div>




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

                                <a href="/bbs/Gossiping/M.1569266625.A.E60.htm
l">Re: [問卦] 台灣男森在東南亞有多帥? </a>

                        </div>
                        <div class="meta">
                                <div class="author">deepdish</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%8F%B0%E7%81%A3%
E7%94%B7%E6%A3%AE%E5%9C%A8%E6%9D%B1%E5%8D%97%E4%BA%9E%E6%9C%89%E5%A4%9A%E5%B8%
A5%EF%BC%9F">搜尋同標題文章</a></div>

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

                                        </div>

                                </div>
                                <div class="date"> 9/24</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.1569266815.A.627.htm
l">Re: [問卦] 我媽的房門打不開了（認真求解</a>

                        </div>
                        <div class="meta">
                                <div class="author">Dinenger</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;%E6%88%91%E5%AA%BD%
E7%9A%84%E6%88%BF%E9%96%80%E6%89%93%E4%B8%8D%E9%96%8B%E4%BA%86%EF%BC%88%E8%AA%
8D%E7%9C%9F%E6%B1%82%E8%A7%A3">搜尋同標題文章</a></div>
```

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

                                        </div>

                                </div>
                                <div class="date"> 9/24</div>
                                <div class="mark"></div>
                        </div>
                </div>




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

                                <a href="/bbs/Gossiping/M.1569266863.A.BBE.htm
l">[問卦] 一般人入黨可以幹嘛</a>

                        </div>
                        <div class="meta">
                                <div class="author">clever0705</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%E8%88%AC%
E4%BA%BA%E5%85%A5%E9%BB%A8%E5%8F%AF%E4%BB%A5%E5%B9%B9%E5%98%9B">搜尋同標題文章</
a></div>

                                                <div class="item"><a href="/bb
s/Gossiping/search?q=author%3Aclever0705">搜尋看板內 clever0705 的文章</a></div>
```

```
                                        </div>

                                </div>
                                <div class="date"> 9/24</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.1569266888.A.533.htm
l">[問卦] 大家會拒讀義大嗎</a>

                        </div>
                        <div class="meta">
                                <div class="author">bupayu</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%AE%B6%
E6%9C%83%E6%8B%92%E8%AE%80%E7%BE%A9%E5%A4%A7%E5%97%8E">搜尋同標題文章</a></div>

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

                                        </div>

                                </div>
                                <div class="date"> 9/24</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.html">[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="/bbs/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="/bbs/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">11</span></div>
                        <div class="title">

                                <a href="/bbs/Gossiping/M.1568569396.A.D48.htm
l">[公告] 呼籲!!請勿於看板發布販賣帳號之訊息</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;%E5%91%BC%E7%B1%B2%
21%21%E8%AB%8B%E5%8B%BF%E6%96%BC%E7%9C%8B%E6%9D%BF%E7%99%BC%E5%B8%83%E8%B2%A9%
E8%B3%A3%E5%B8%B3%E8%99%9F%E4%B9%8B%E8%A8%8A%E6%81%AF">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"> 9/16</div>
                                <div class="mark">M</div>
                        </div>
                </div>
```

```
                <div class="r-ent">
                        <div class="nrec"><span class="hl f3">48</span></div>
                        <div class="title">

                                <a href="/bbs/Gossiping/M.1569227318.A.583.htm
l">[公告] 關於客家人水桶事件，可申請改判。</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;%E9%97%9C%E6%96%BC%
E5%AE%A2%E5%AE%B6%E4%BA%BA%E6%B0%B4%E6%A1%B6%E4%BA%8B%E4%BB%B6%EF%BC%8C%E5%8F%
AF%E7%94%B3%E8%AB%8B%E6%94%B9%E5%88%A4%E3%80%82">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"> 9/23</div>
                                <div class="mark">M</div>
                        </div>
                </div>




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

[公告] 九月份中秋線上烤肉兼置底閒聊區

Bignana

⋯

搜尋同標題文章

搜尋看板內 Bignana 的文章

 9/13
M

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。

```
<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 [33]:

```
import requests
from bs4 import BeautifulSoup

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

```
['Aladdin', "A Dog's Journey", 'Men in Black: International', 'Pokémon Detecti
ve Pikachu', 'X-Men: Dark Phoenix', 'Rocketman', 'The Secret Life of Pets 2',
'John Wick: Chapter 3 - Parabellum', 'Avengers: Endgame', 'Shazam!', 'Kursk',
'Apollo 11', 'Red Joan', 'Stockholm', 'Teen Spirit', 'An Elephant Sitting Stil
l', 'Integrity', 'Dearest Anita', "Dad's Suit", 'When Green Turns to Gold', 'L
ate Life: The Chien-Ming Wang Story', 'The Knight of Shadows…', 'Another Worl
d', 'Fly Me To The Saitama', "Long Day's Journey Into Night", 'Andhadhun', 'A
Long Goodbye', 'Miss & Mrs Cops', 'Take Point', '70 Big Ones', 'Loro', 'Whistl
eblower', 'Hit-and-Run Squad', '96', 'Who You Think I Am']
```

隨堂練習：擷取所有華航機上電影清單

In [35]:
```
print(ca_movie_titles)
```

```
['Aladdin', "A Dog's Journey", 'Men in Black: International', 'Pokémon Detecti
ve Pikachu', 'X-Men: Dark Phoenix', 'Rocketman', 'The Secret Life of Pets 2',
'John Wick: Chapter 3 - Parabellum', 'Avengers: Endgame', 'Shazam!', 'Kursk',
'Apollo 11', 'Red Joan', 'Stockholm', 'Teen Spirit', 'An Elephant Sitting Stil
l', 'Integrity', 'Dearest Anita', "Dad's Suit", 'When Green Turns to Gold', 'L
ate Life: The Chien-Ming Wang Story', 'The Knight of Shadows…', 'Another Worl
d', 'Fly Me To The Saitama', "Long Day's Journey Into Night", 'Andhadhun', 'A
Long Goodbye', 'Miss & Mrs Cops', 'Take Point', '70 Big Ones', 'Loro', 'Whistl
eblower', 'Hit-and-Run Squad', '96', 'Who You Think I Am', "Tim Burton's Corps
e Bride", 'The Peanuts Movie', 'Sing', 'Puss In Boots', 'The Lego Batman Movi
e', 'Ferdinand', 'The Polar Express', 'Hop', 'Pan', 'Where the Wild Things Ar
e', 'Three Times', 'Beyond Beauty - Taiwan from Above', 'The Wedding Banquet',
'Millennium Mambo', 'On Happiness Road', 'Father to Son', 'Long Time No Sea',
'More Than Blue', 'Handan', 'Master Z: Ip Man Legacy', 'Shadow', 'Hidden Man',
'Still Human', 'Last Letter', 'Tracey', 'Dragon Ball Super: Broly', 'Stolen Id
entity', 'Hibiki', 'Project Gutenberg', 'Default', 'Masquerade Hotel', 'Millio
n Dollar Man', 'The House Where The Mermaid Sleeps', 'Unstoppable', 'Extreme J
ob', 'Innocent Witness', 'Badhaai Ho', 'Border', 'Family is Family', 'Capernau
m', 'U - July 22', 'A Better Tomorrow', 'A Chinese Ghost Story', 'Rouge', 'Gin
tama 2', 'A Real Vermeer', 'Cafe Funiculi Funicula', 'Code Blue: The Movie',
"Jupiter's Moon", 'Simpel…', 'Shoplifters', "Midsummer's Equation", 'Attack On
Titan: Kakusei', 'Detective Conan: Crimson Love Letter', 'Detective Conan Epis
ode "ONE"', 'Hichki', 'My Hero Academia: Two Heroes', 'Mirai', 'Lupin the 3rd
vs. Detective Conan', 'The Negotiation', 'Euforia', 'The Benefit of the Doub
t', 'Hello Mr. Billionaire', 'Wished', 'The Square', 'Le Brio', 'Pad Man', 'Jo
hn Wick', 'X-Men', 'X-Men: First Class', 'X-Men: Days of Future Past', 'X-Men:
Apocalypse', 'The Mule', 'Aquaman', 'Iron Man', 'Iron Man 2', 'Iron Man 3', 'T
he Avengers', 'Avengers: Age of Ultron', 'Avengers: Infinity War', 'Alita: Bat
tle Angel', "Isn't it Romantic", 'Eddie The Eagle', 'Les Misérables', 'Knight
and Day', 'Invictus', 'Café Society', 'Going In Style', 'Fantastic Beasts an
d…', 'Crazy, Stupid, Love.', 'J. Edgar', 'John Wick: Chapter 2', 'A Star is Bo
rn', 'Manchester By The Sea', 'Hitman: Agent 47', 'I, Daniel Blake', 'My Cousi
n Rachel', 'Prometheus', 'Crazy Heart', 'Ruby Sparks', 'Life of Pi', 'Moulin R
ouge', 'Hidden Figures', 'Logan', 'Runner Runner', 'Creed', 'Sherlock Holmes:
A Game of Shadows', 'Deepwater Horizon', 'The Intern', 'Captain America: Civil
```

隨堂練習：找出華航機上最高評等的電影

In [39]:
```
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/) 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 [ ]:

```
#!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

In [ ]:

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

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

隨堂練習：以 selenium 實作 get_movie_data(movie_title)

In [41]:
```
get_movie_data("Avengers: Endgame (2019)")
```

Out[41]:
```
{'movieTitle': 'Avengers: Endgame (2019)',
 'moviePosterLink': 'https://m.media-amazon.com/images/M/MV5BMTc5MDE2ODcwNV5BM
l5BanBnXkFtZTgwMzI2NzQ2NzM@._V1_UX182_CR0,0,182,268_AL_.jpg',
 'movieRating': 8.6,
 'movieGenre': ['Action', 'Adventure', 'Sci-Fi'],
 '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']}
```

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)", "Aven
gers: Infinity War (2018)", "Avengers: Endgame (2019)"]
```

In [43]:
```
print(avengers_movie_data)
```

```
[{'movieTitle': 'The Avengers (2012)', 'moviePosterLink': 'https://m.media-ama
zon.com/images/M/MV5BNDYxNjQyMjAtNTdiOS00NGYwLWFmNTAtNThmYjU5ZGI2YTI1XkEyXkFqc
GdeQXVyMTMxODk2OTU@._V1_UX182_CR0,0,182,268_AL_.jpg', 'movieRating': 8.0, 'mov
ieGenre': ['Action', 'Adventure', 'Sci-Fi'], 'movieCast': ['Robert Downey J
r.', 'Chris Evans', 'Mark Ruffalo', 'Chris Hemsworth', 'Scarlett Johansson',
'Jeremy Renner', 'Tom Hiddleston', 'Clark Gregg', 'Cobie Smulders', 'Stellan S
karsgård', 'Samuel L. Jackson', 'Gwyneth Paltrow', 'Paul Bettany', 'Alexis Den
isof', 'Tina Benko']}, {'movieTitle': 'Avengers: Age of Ultron (2015)', 'movie
PosterLink': 'https://m.media-amazon.com/images/M/MV5BMTM4OGJmNWMtOTM4Ni00NTE3
LTg3MDItZmQxYjc4N2JhNmUxXkEyXkFqcGdeQXVyNTgzMDMzMTg@._V1_UX182_CR0,0,182,268_A
L_.jpg', 'movieRating': 7.3, 'movieGenre': ['Action', 'Adventure', 'Sci-Fi'],
'movieCast': ['Robert Downey Jr.', 'Chris Hemsworth', 'Mark Ruffalo', 'Chris E
vans', 'Scarlett Johansson', 'Jeremy Renner', 'James Spader', 'Samuel L. Jacks
on', 'Don Cheadle', 'Aaron Taylor-Johnson', 'Elizabeth Olsen', 'Paul Bettany',
'Cobie Smulders', 'Anthony Mackie', 'Hayley Atwell']}, {'movieTitle': 'Avenger
s: Infinity War (2018)', 'moviePosterLink': 'https://m.media-amazon.com/image
s/M/MV5BMjMxNjY2MDU1OV5BMl5BanBnXkFtZTgwNzY1MTUwNTM@._V1_UX182_CR0,0,182,268_A
L_.jpg', 'movieRating': 8.5, 'movieGenre': ['Action', 'Adventure', 'Sci-Fi'],
'movieCast': ['Robert Downey Jr.', 'Chris Hemsworth', 'Mark Ruffalo', 'Chris E
vans', 'Scarlett Johansson', 'Don Cheadle', 'Benedict Cumberbatch', 'Tom Holla
nd', 'Chadwick Boseman', 'Zoe Saldana', 'Karen Gillan', 'Tom Hiddleston', 'Pau
l Bettany', 'Elizabeth Olsen', 'Anthony Mackie']}, {'movieTitle': 'Avengers: E
ndgame (2019)', 'moviePosterLink': 'https://m.media-amazon.com/images/M/MV5BMT
c5MDE2ODcwNV5BMl5BanBnXkFtZTgwMzI2NzQ2NzM@._V1_UX182_CR0,0,182,268_AL_.jpg',
'movieRating': 8.6, 'movieGenre': ['Action', 'Adventure', 'Sci-Fi'], 'movieCas
t': ['Robert Downey Jr.', 'Chris Evans', 'Mark Ruffalo', 'Chris Hemsworth', 'S
carlett Johansson', 'Jeremy Renner', 'Don Cheadle', 'Paul Rudd', 'Benedict Cum
berbatch', 'Chadwick Boseman', 'Brie Larson', 'Tom Holland', 'Karen Gillan',
'Zoe Saldana', 'Evangeline Lilly']}]
```

將擷取的電影資訊匯出

In [ ]:

```
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/#)
- pyquery: a jquery-like library for python (https://pythonhosted.org/pyquery/)
- 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 [45]:
```
ans()
```

Out[45]:
```
{'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}
```