

靜態網頁擷取

郭耀仁

網頁資料擷取的核心任務

- 如何獲得資料
- 如何解析資料
- 如何清理整併資料

網頁資料來源

- JSON (<https://www.json.org/>).
- HTML (<https://zh.wikipedia.org/zh-tw/HTML>).

JSON

以 json 套件解析 JSON 文字為 dict

```
In [14]: import json

json_str = """
{
    "teamName": "Chicago Bulls",
    "season": "1995-96",
    "records": {
        "wins": 72,
        "losses": 10
    },
    "startingLineup": ["Ron Harper", "Michael Jordan", "Scottie Pippen", "Dennis Rod
man", "Luc Longley"],
    "isChampion": true
}
"""
json_dict = json.loads(json_str)
```

```
In [13]: print(type(json_dict))  
         json_dict
```

```
<class 'dict'>
```

```
Out[13]: {'isChampion': True,  
          'records': {'losses': 10, 'wins': 72},  
          'season': '1995-96',  
          'startingLineup': ['Ron Harper',  
                              'Michael Jordan',  
                              'Scottie Pippen',  
                              'Dennis Rodman',  
                              'Luc Longley'],  
          'teamName': 'Chicago Bulls'}
```

如果是本機的 .json 檔案

```
In [16]: import json

with open('chicago_bulls_19951996.json') as json_data:
    json_dict = json.load(json_data)
```

```
In [17]: print(type(json_dict))  
         json_dict
```

```
<class 'dict'>
```

```
Out[17]: {'isChampion': True,  
          'records': {'losses': 10, 'wins': 72},  
          'season': '1995-96',  
          'startingLineup': ['Ron Harper',  
                              'Michael Jordan',  
                              'Scottie Pippen',  
                              'Dennis Rodman',  
                              'Luc Longley'],  
          'teamName': 'Chicago Bulls'}
```


如果是網路上的 .json 檔案

```
In [19]: from requests import get

json_response = get("https://storage.googleapis.com/ds_data_import/chicago_bulls_1995_1996.json")
json_dict = json_response.json()
```

```
In [20]: print(type(json_dict))
json_dict
```

```
<class 'dict'>
```

```
Out[20]: {'assistant_coach': ['Jim Cleamons',
                              'John Paxson',
                              'Jimmy Rodgers',
                              'Tex Winter'],
          'coach': 'Phil Jackson',
          'records': {'losses': 10, 'wins': 72},
          'starting_lineups': {'C': 'Luc Longley',
                                'PF': 'Dennis Rodman',
                                'PG': 'Ron Harper',
                                'SF': 'Scottie Pippen',
                                'SG': 'Michael Jordan'},
          'team_name': 'Chicago Bulls'}
```

以 pchome 商品頁面為例

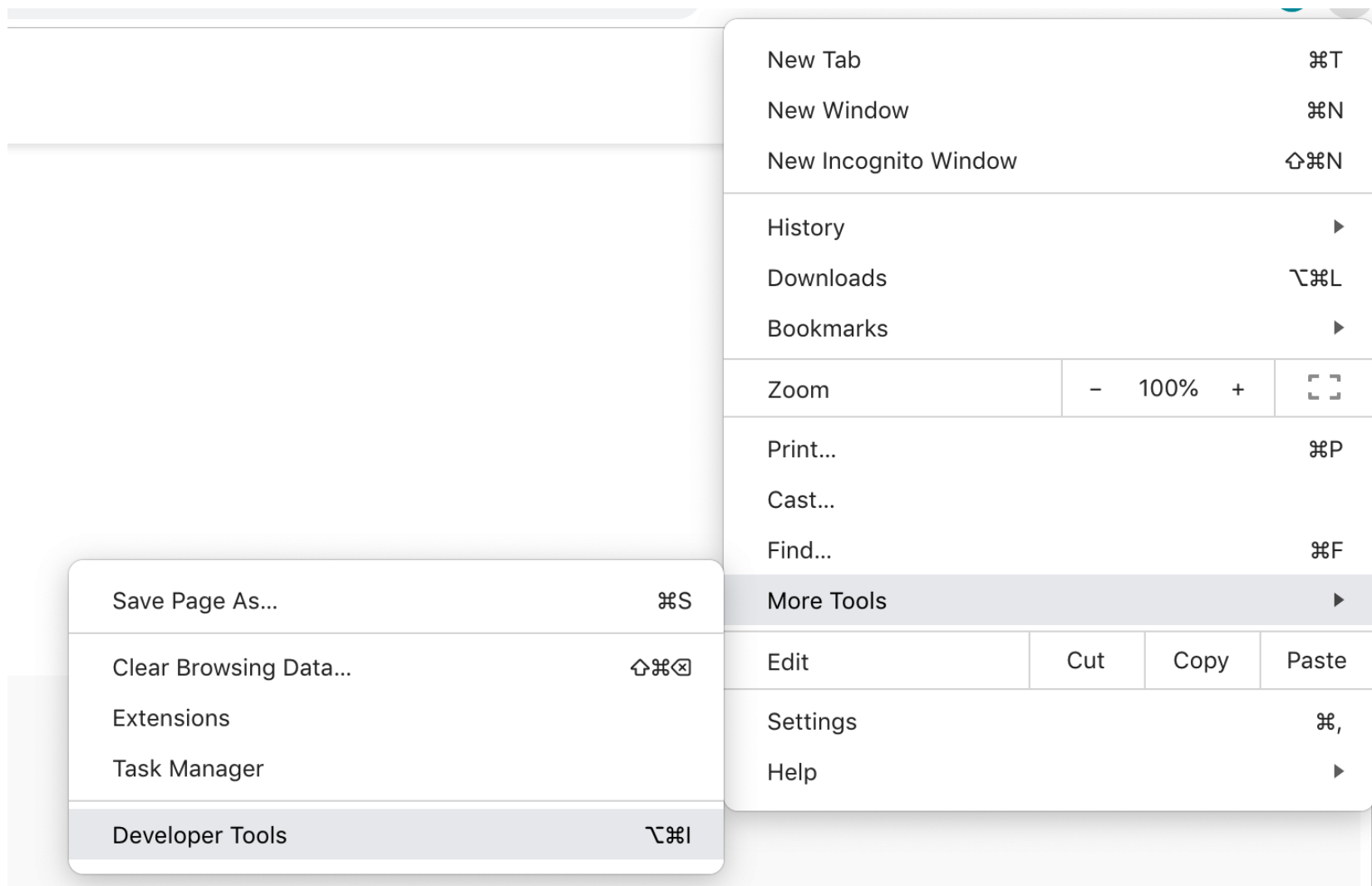
<https://ecshweb.pchome.com.tw/search/v3.3/?q=mac>
(<https://ecshweb.pchome.com.tw/search/v3.3/?q=mac>).

先下載 Chrome 外掛

- Quick Javascript Switcher (<https://chrome.google.com/webstore/detail/quick-javascript-switcher/geddoclleiomckbhadiapdggiiccfje>): 關閉 JavaScript 的作用

關閉 JavaScript 之後商品資訊都不見了

打開 Chrome 開發者工具



打開 Network 標籤，點選 XHR

```
In [21]: import json

from requests import get

json_response = get("https://ecshweb.pchome.com.tw/search/v3.3/all/results?q=mac&p
age=1&sort=rnk/dc")
json_dict = json_response.json()
```

```
In [24]: print(json_dict['prods'][0])
```

```
{'Id': 'DYAJBD-A90097XCW', 'cateId': 'DYAJBD', 'picS': '/items/DYAJBDA90097XCW/000002_1531729125.png', 'picB': '/items/DYAJBDA90097XCW/000001_1542854872.jpg', 'name': 'MacBook Pro 13-inch : 2.3GHz dual-core i5, 128GB - Space Grey (MPXQ2TA/A)', 'describe': '▼限時88折優惠▼MacBook Pro 13-(太空灰) : 2.3GHz dual-core i5, 128GB - Space Grey太空灰 (MPXQ2TA/A)\r\n\r\n★此機型無觸控列和touch id\r\n\r\n\r\n★限時88折優惠 11 01 (四) 10:00 至 11 30 (五) 10:00止\r\n\r\n數量有限, 售完為止\r\n\r\n網路價$41900. 限時價\r\n$3 6 8 7 2\r\n\r\n\r\n商品特色\r\n\r\n● 2.3 ghz 處理器\r\n\r\n● 128 gb 儲存空間\r\n\r\n● 2.3 ghz 雙核心第七代 intel core i5 處理器\r\n\r\n● turbo boost 可達 3.6 ghz\r\n\r\n● 8 gb 2133 mhz lpddr 3 記憶體\r\n\r\n● 128 gb ssd 儲存裝置\r\n\r\n● intel iris plus graphics 640\r\n\r\n● 兩個 thunderbolt 3 埠\r\n\r\n\r\n\r\n館長小叮嚀: 儲值購買最划算~\r\n\r\n【11 1~30限定】刷卡買3 000儲值金\r\n\r\n送150刷卡金→最高回饋5%↑11月儲值優惠活動, 請點我\r\n\r\n\r\n11月刷卡限時優惠 限量需登錄\r\n\r\n\r\n↑銀行刷卡活動詳情請點圖片了解\r\n\r\n\r\n\r\n\r\n注意事項', 'price': 36872, 'originPrice': 36872, 'author': '', 'brand': '', 'publishDate': '', 'sellerId': '', 'isPChome': 1, 'isNC17': 0, 'couponActid': [], 'BU': 'ec'}
```

HTML

先安裝 Chrome 外掛

- SelectorGadget
(<https://chrome.google.com/webstore/detail/selectorgadget/mhjhnkcfbdhnjickkkdbjoenhl=zh-TW>): 幫助 CSS 選擇

CSS 選擇

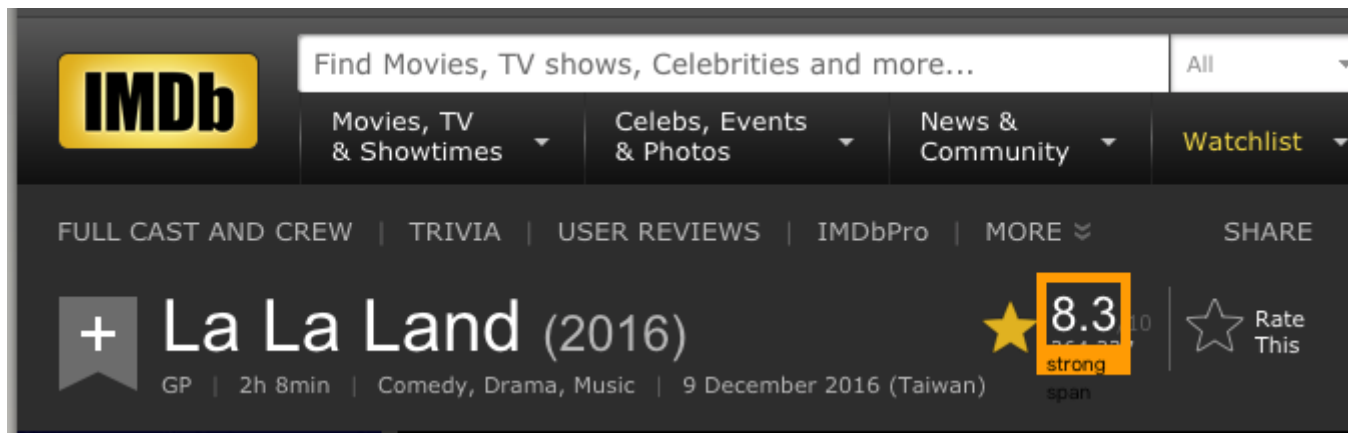
- 利用文件物件模型（Document Object Model, DOM）選擇網頁元素的技巧

CSS 選擇 (2)

- 以 IMDB: La La Land (<http://www.imdb.com/title/tt3783958/>) 為例
- 選出評等 (Rating)

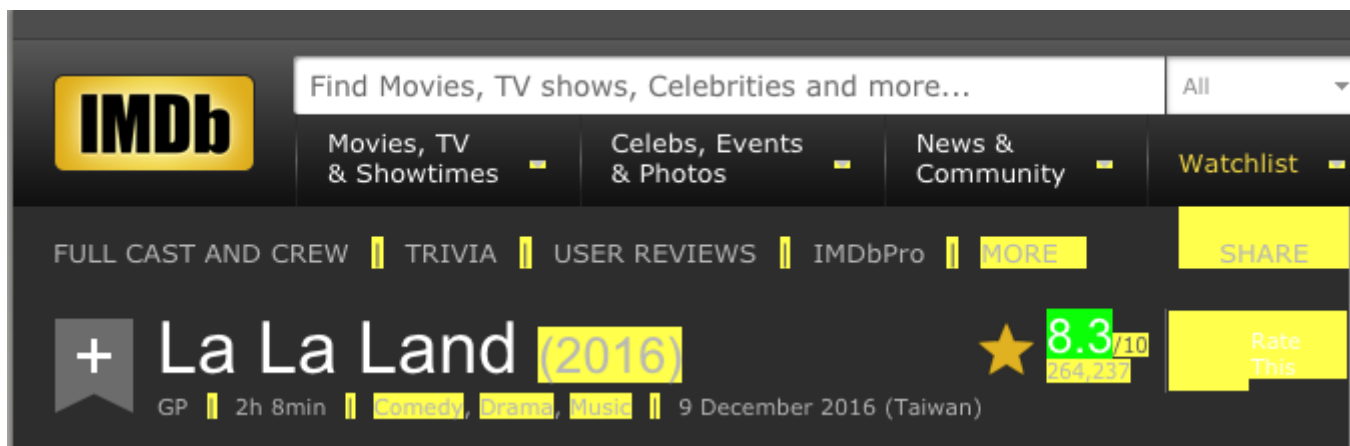
CSS 選擇 (3)

- 點選 SelectorGadget 的外掛圖示
- 先點選評等 (Rating)



CSS 選擇 (4)

- 會有多個元素被選出（標註黃色），因為這頁有很多 ``



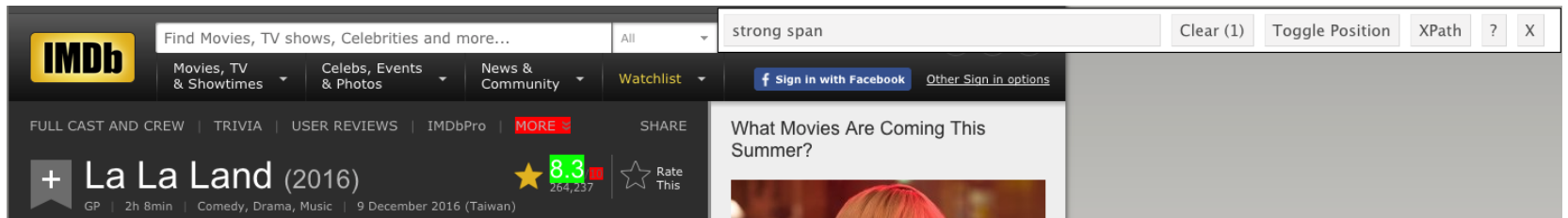
CSS 選擇 (5)

- 接著點選不要的元素（標註紅色）













CSS 選擇 (6)

- 這時我們就得到正確的 CSS 選擇: `strong span`



CSS 選擇 (7)

- 再練習一次，這次我們要將這頁的演員名單（Cast）擷取出來：

Cast			Edit
Cast overview, first billed only:			
	Ryan Gosling	...	Sebastian
	Emma Stone	...	Mia
	Amiée Conn	...	Famous Actress
	Terry Walters	...	Linda (Coffee Shop Manager)
	Thom Shelton	...	Coffee Spiller
	Cinda Adams	...	Casting Director (First Audition)
	Callie Hernandez	...	Tracy
	Jessica Rothe	...	Alexis
	Sonoya Mizuno	...	Caitlin
	Rosemarie DeWitt	...	Laura

CSS 選擇 (8)

- 先點選男主角 Ryan Gosling
- 乍看之下好像 `.itemprop` 這個類別就對了

Cast

[Edit](#)

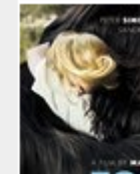
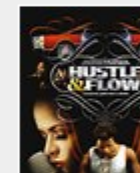
Cast overview, first billed only:

	Ryan Gosling	...	Sebastian
	Emma Stone	...	Mia
	Amiée Conn	...	Famous Actress
	Terry Walters	...	Linda (Coffee Shop Manager)
	Thom Shelton	...	Coffee Spiller
	Cinda Adams	...	Casting Director (First Audition)
	Callie Hernandez	...	Tracy
	Jessica Rothe	...	Alexis
	Sonoya Mizuno	...	Caitlin
	Rosemarie DeWitt	...	Laura
	J.K. Simmons	...	Bill
	Claudine Claudio	...	Karen (Waitress)
	Jason Fuchs	...	Carlo
	D.A. Wallach	...	'80s Singer



User List

Related lis



[See all rel](#)

Connect



CSS 選擇 (9)

- 但是其實其他的元素也有使用 `.itemprop` 這個類別

A jazz pianist falls for an aspiring actress in Los Angeles.

Director: [Damien Chazelle](#)

Writer: [Damien Chazelle](#)

Stars: [Ryan Gosling](#), [Emma Stone](#), [Rosemarie DeWitt](#) | [See full cast & crew »](#)

Storyline

[Edit](#)

Mia, an aspiring actress, serves lattes to movie stars in between auditions and Sebastian, a jazz musician, scrapes by playing cocktail party gigs in dingy bars, but as success mounts they are faced with decisions that begin to fray the fragile fabric of their love affair, and the dreams they worked so hard to maintain in each other threaten to rip them apart. *Written by [Eirini](#)*

[Plot Summary](#) | [Plot Synopsis](#)

Plot Keywords: [pianist](#) | [aspiring actress](#) | [jazz musician](#) | [movie set](#) | [jazz music](#)
| [See All \(247\) »](#)

CSS 選擇 (10)

- 於是點選不要的元素（標註紅色）

A jazz pianist falls for an aspiring actress in Los Angeles.

Director: [Damien Chazelle](#)

Writer: [Damien Chazelle](#)

Stars: [Ryan Gosling](#), [Emma Stone](#), [Rosemarie DeWitt](#) | [See full cast & crew »](#)

93 [Metascore](#)
From [metacritic.com](#)

Reviews
[1,075 user](#) | [592 critic](#)

 **Popularity**
45 (↓ 5)

CSS 選擇 (11)

- 這時我們就得到正確的 CSS 選擇: `.itemprop .itemprop`

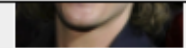
Cast

Edit

Cast overview, first billed only:

	Ryan Gosling	...	Sebastian
	Emma Stone	...	Mia
	Amiée Conn	...	Famous Actress
	Terry Walters	...	Linda (Coffee Shop Manager)
	Thom Shelton	...	Coffee Spiller
	Cinda Adams	...	Casting Director (First Audition)
	Callie Hernandez	...	Tracy
	Jessica Rothe	...	Alexis
	Sonoya Mizuno	...	Caitlin
	Rosemarie DeWitt	...	Laura
	J.K. Simmons	...	Bill
	Claudine Claudio	...	Karen (Waitress)

.itemprop .itemprop



User Lists

Related lists from IMDb



Nog kijk

a list of 48
created 27



To watch

a list of 38
created 06



Best and

a list of 42
created 08



Sight & of 2016

a list of 30
created 6



Watchmen

a list of 21
created 2

See all related lists »

Pyquery

- 用來解析 HTML 與 XML 檔案的 Python 模組
- 先在終端機安裝模組

```
# Local  
pip install pyquery # terminal  
# Google Colab  
!pip install pyquery # cell
```

In [1]: **from pyquery import** PyQuery **as** pq

```
lalaland_url = "http://www.imdb.com/title/tt3783958/"  
html_doc = pq(lalaland_url)  
html_doc.contents()
```

Out[1]: ['\n', <Element head at 0x109ed3778>, '\n', <Element body at 0x109ed3818>, '\n']


```
In [2]: rating_css = "strong span"
print(html_doc(rating_css))
print(html_doc(rating_css).text())
```

```
<span itemprop="ratingValue">8.1</span>
8.1
```

```
In [3]: lalaland_rating = html_doc(rating_css).text()  
        type(lalaland_rating)
```

```
Out[3]: str
```

隨堂練習：從網頁中擷取出電影類型

<http://www.imdb.com/title/tt3783958/> (<http://www.imdb.com/title/tt3783958/>).

```
In [5]: get_genre()
```

```
Out[5]: ['Comedy', 'Drama', 'Music']
```

隨堂練習：從網頁中擷取出演員名單

<http://www.imdb.com/title/tt3783958/> (<http://www.imdb.com/title/tt3783958/>)

```
In [7]: get_cast()
```

```
Out[7]: ['Ryan Gosling',  
        'Emma Stone',  
        'Amiée Conn',  
        'Terry Walters',  
        'Thom Shelton',  
        'Cinda Adams',  
        'Callie Hernandez',  
        'Jessica Rothe',  
        'Sonoya Mizuno',  
        'Rosemarie DeWitt',  
        'J.K. Simmons',  
        'Claudine Claudio',  
        'Jason Fuchs',  
        'D.A. Wallach']
```