動態網頁擷取

郭耀仁

建立一個新的虛擬環境

```
conda create -n pyscraper python=3
(source) activate pyscraper
pip install pandas matplotlib selenium ipykernel
```

對應 Kernel

```
python -m ipykernel install --user --name pyscraper --display-name "pyscraper"
```

更新 Chrome 瀏覽器

https://www.google.com/chrome/browser/desktop/index.html
(https://www.google.com/chrome/browser/desktop/index.html)

下載驅動 Chrome 瀏覽器的 Driver

https://sites.google.com/a/chromium.org/chromedriver/downloads
(https://sites.google.com/a/chromium.org/chromedriver/downloads)

測試

```
from selenium import webdriver

driver_path = "C:/Users/user/Downloads/chromedriver.exe" # 改為你的 driver 路徑
driver = webdriver.Chrome(executable_path = driver_path)
driver.get("http://www.python.org")
```

結束測試

```
driver.close()
```

也可以使用 Firefox

https://www.mozilla.org/en-US/firefox/new/ (https://www.mozilla.org/en-US/firefox/new/)

下載驅動 Firefox 瀏覽器的 Driver

https://github.com/mozilla/geckodriver/releases
(https://github.com/mozilla/geckodriver/releases)

結束測試

```
driver.close()
```

測試

```
from selenium import webdriver

driver_path = "C:/Users/user/Downloads/geckodriver.exe" # 改為你的 driver 路徑
driver = webdriver.Firefox(executable_path = driver_path)
driver.get("http://www.python.org")
```

driver 的基本方法

- `.get()`: 移動至某一個 URL
- `.forward()`: 前一頁
- `.backward()`: 上一頁
- `.close()`: 關閉瀏覽器
- `.find_element(s)_by_css_selector`: 利用 CSS 選擇器擷取一個或多個元素
- `.find_element(s)_by_xpath`: 利用 XPath 選擇器擷取一個或多個元素

element 的基本方法

- `.send_keys()`:輸入文字
- `.clear()`:清空文字輸入框
- `.click()`:點擊

IMDB: La La Land

- 前往 IMDB (http://www.imdb.com/) 首頁
- 搜尋 'la la land'
- 點擊搜尋
- 點擊第一個搜尋結果
- 將評等與演員名單擷取下來

In [1]:

```
def get_movie_info():

    from selenium import webdriver
    from selenium.webdriver.common.keys import Keys
    import time

    driver_path = "C:/Users/user/Downloads/chromedriver.exe" # 改為你的 driver 路徑
    driver = webdriver.Chrome(driver_path)
    driver.get("http://www.imdb.com/")
    search_elem = driver.find_element_by_css_selector("#navbar-query")
    search_elem.send_keys('la la land')
    time.sleep(3)
    search_button_elem = driver.find_element_by_css_selector("#navbar-submit-butto
n .navbarSprite")
    search_button_elem.click()
    time.sleep(3)
    first_result_elem = driver.find_element_by_css_selector("#findSubHeader+ .find
Section .odd:nth-child(1) .result_text a")
    first_result_elem.click()
    time.sleep(3)
    rating_elem = driver.find_element_by_css_selector("strong span")
    rating = float(rating_elem.text)
    cast_elem = driver.find_elements_by_css_selector(".itemprop .itemprop")
    cast_list = [cast.text for cast in cast_elem]
    driver.close()
    return rating, cast_list
```

In [2]:
```
get_movie_info()
```

Out[2]:
```
(8.1,
 ['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',
  'Trevor Lissauer'])
```

隨堂練習：改用 XPath 選擇器 `get_movie_info()`

Yahoo! 奇摩股市：上市單日成交價排行

- 前往 Yahoo! 奇摩股市
- 點選**更多排行**
- 點選**上市行情類排行榜：單日成交價排行**
- 點選**列出前一百名排行**
- 將股票代號/名稱擷取下來

In [3]:

```
def get_top_100():
    from selenium import webdriver

    driver_path = "C:/Users/user/Downloads/chromedriver.exe" # 改為你的 driver 路徑
    driver = webdriver.Chrome(driver_path)
    driver.get("https://tw.finance.yahoo.com/")
    more_rank_elem = driver.find_element_by_css_selector('.yui-text-left .yui-text
-left table tr:nth-child(1) .stext div a')
    more_rank_elem.click()
    price_rank_elem = driver.find_element_by_css_selector('.yui-text-left+ .yui-te
xt-left tr:nth-child(5) a')
    price_rank_elem.click()
    top_100_elem = driver.find_element_by_css_selector('p a+ a')
    top_100_elem.click()
    ticker_name_elem = driver.find_elements_by_css_selector('.name')
    ticker_name = [tn.text for tn in ticker_name_elem]
    driver.close()
    return ticker_name
```

In [4]:
```
get_top_100()
```

Out[4]:
```
['3008 大立光',
 '6415 矽力-KY',
 '6409 旭隼',
 '1590 亞德客-KY',
 '6414 樺漢',
 '2723 美食-KY',
 '6452 康友-KY',
 '2059 川湖',
 '2327 國巨',
 '2207 和泰車',
 '5269 祥碩',
 '2049 上銀',
 '2474 可成',
 '2454 聯發科',
 '2231 為升',
 '8464 億豐',
 '3443 創意',
 '1476 儒鴻',
 '2912 統一超',
 '3406 玉晶光',
 '2357 華碩',
 '4943 康控-KY',
 '2227 裕日車',
 '3665 貿聯-KY',
 '2330 台積電',
 '8341 日友',
 '3413 京鼎',
 '2395 研華',
 '8454 富邦媒',
 '2939 凱羿-KY',
 '1707 葡萄王',
 '3533 嘉澤',
 '6456 GIS-KY',
 '2496 卓越',
```

'2439 美律',
'3130 一零四',
'8422 可寧衛',
'4551 智伸科',
'4763 材料-KY',
'9921 巨大',
'2239 英利-KY',
'2360 致茂',
'6504 南六',
'2707 晶華',
'8070 長華',
'1537 廣隆',
'2308 台達電',
'1256 鮮活果汁-KY',
'8114 振樺電',
'6451 訊芯-KY',
'2727 王品',
'4912 聯德控股-KY',
'1477 聚陽',
'9910 豐泰',
'4438 廣越',
'9914 美利達',
'4137 麗豐-KY',
'2731 雄獅',
'1536 和大',
'1723 中碳',
'8480 泰昇-KY',
'3450 聯鈞',
'3532 台勝科',
'1558 伸興',
'6271 同欣電',
'3034 聯詠',
'6591 動力-KY',
'4148 全宇生技-KY',
'6505 台塑化',
'6464 台數科',
'3026 禾伸堂',
'6533 晶心科',

```
'2379 瑞昱',
'9941 裕融',
'4739 康普',
'1338 廣華-KY',
'2929 淘帝-KY',
'2455 全新',
'4977 眾達-KY',
'2345 智邦',
'6269 台郡',
'3045 台灣大',
'2412 中華電',
'1326 台化',
'9938 百和',
'3661 世芯-KY',
'2228 劍麟',
'2492 華新科',
'5264 鎧勝-KY',
'3044 健鼎',
'4552 力達-KY',
'1301 台塑',
'3617 碩天',
'2456 奇力新',
'2383 台光電',
'1526 日馳',
'2317 鴻海',
'5871 中租-KY',
'3019 亞光',
'6230 超眾']
```

延伸閱讀

WebDriver API (http://selenium-python.readthedocs.io/api.html)