

Python Crawler Practice

Python 爬蟲實戰案例_PART II



Python Web Crawler

JavaScript 動態網頁擷取

- ◆ Selenium不只可以與HTML表單進行互動，還可以幫助我們從JavaScript產生的動態網頁擷取出所需資料。簡單地說，Selenium可以讓我們取得瀏覽器即時產生的HTML網頁內容



JavaScript 動態網頁擷取

擷取「Hahow好學校」的課程資訊

- ◆ 「Hahow 好學校」的課程資訊會公佈在其網站上，網址如下：

<http://hahow.in/courses>。

- ◆ 網頁的每一個方框是一門開課資訊，當我們檢視網頁的HTML原始碼時，HTML原始碼大部份是JavaScript程式碼，

根本看不到課程資訊的HTML標籤，因為網頁內容是使用 JavaScript動態產生的網頁。

```
<!doctype html><html lang="zh-TW"><head><meta charset="utf-8"><meta name="viewport" content="width=device-width,initial-scale=1"><title>Hahow 好學校 | 最有趣的線上課程平台 | 自學那些學校沒教的事</title><link rel="shortcut icon" href="https://hahow.in/favicon.ico"><link rel="apple-touch-icon" href="https://hahow.in/highres-icon.png"><link rel="apple-touch-icon" href="https://hahow.in/apple-touch-icon.png"><link rel="mask-icon" href="https://hahow.in/website_icon.svg" color="#eb5e00"><link rel="stylesheet" href="https://maxcdn.bootstrapcdn.com/bootstrap/3.3.7/css/bootstrap.min.css" integrity="sha384-BVYiiSIFeK1dGmJRAkycuHAHRg32OmUcww7on3RYdg4Va+PmSTsz/K68vbDEjh4u" crossorigin="anonymous"><script type="text/javascript">!function(){var n=window.analytics=window.analytics||[];if(!n.initialize)if(n.invoked)window.console&&console.error("Segment snippet included twice.");else{n.invoked=!0,n.methods=["trackSubmit","trackClick","trackLink","trackForm","pageview","identify","reset","group","track","ready","alias","debug","page","once","off","on"],n.factory=function(e){return function(){var t=Array.prototype.slice.call(arguments);return t.unshift(e),n.push(t),n}};for(var t=0;t<n.methods.length;t++){var e=n.methods[t];n[e]=n.factory(e)}n.load=function(t){var e=document.createElement("script");e.type="text/javascript",e.async=!0,e.src=("https:"==document.location.protocol?"https://":http://)+"cdn.segment.com/analytics.js/v1/"+t+"/analytics.min.js";var n=document.getElementsByTagName("script")[0];n.parentNode.insertBefore(e,n).n.SNIPPET_VERSION="4.0.0"}()}</script><script type="text/javascript">!function(e,t){var n=e.amplitude||{_q:[],_iq:{},i=t.createElement("script")};i.type="text/javascript",i.async=!0,i.src="https://cdn.amplitude.com/libs/amplitude-4.4.0-min.gz.js",i.onload=function(){e.amplitude.runQueuedFunctions?e.amplitude.runQueuedFunctions():console.log("[Amplitude] Error: could not load SDK");var r=t.getElementsByTagName("script")[0];function s(e,t){e.prototype[t]=function(){return this._q.push([t].concat(Array.prototype.slice.call(arguments,0))),this}}r.parentNode.insertBefore(i,r);for(var o=function(){return this._q=[],this},a=[{"add","append","clearAll","prepend","set","setOnce","unset"},u=0;u<a.length;u++)s(o,a[u]);n.Identify=o;for(var c=function(){return this._q=[],this},l=[{"setProductId","setQuantity","setPrice","setRevenueType","setEventProperties"},p=0;p<l.length;p++)s(c,l[p]);n.Revenue=c;var d=[{"init","logEvent","logRevenue","setUserId","setUserProperties","setOptOut","setVersionName","setDomain","setDeviceId","setGlobalUserProperties","identify","clearUserProperties","setGroup","logRevenueV2","regenerateDeviceId","logEventWithTimestamp","logEventWithGroups","setSessionId","resetSessionId"];function f(t){function e(e){t[e]=function(){t._q.push([e].concat(Array.prototype.slice.call(arguments,0)))}}}for(var n=0;n<d.length;n++)e(d[n]);n.getInstance=function(e){return e.e&&e.e.length?e.e:$default_instance}.toLowerCase(),n._iq={_q:[],_iq:{}},n._iq[e].amplitude=n;(window,document).init("od4a119f268ec1efe16f9178c2ea6f02")</script></head><div id="root"></div><script type="text/javascript" src="https://hahow.in/static/js/main.js?51b5328a"></script><div id="fb-root"></div><script>function(e,t,n){var o,c=e.getElementsByTagName(t)[0];e.getElementById(n)||((o=e.createElement(t)).id=n,o.src="//connect.facebook.net_zh_TW/sdk.js#xfbml=1&version=v2.9&appId=1287520694621477",c.parentNode.insertBefore(o,c))(document,"script","facebook-jssdk")</script><script src="//fast.wistia.com/assets/external/E-v1.js" defer="defer"></script><noscript id="deferred-styles"><link href="https://hahow.in/static/css/main.css?cd0c0947" rel="stylesheet"></noscript><script>[Element.prototype,Document.prototype,DocumentFragment.prototype].forEach(function(e){e.hasOwnProperty("prepend")||Object.defineProperty(e,"prepend",{configurable:!0,enumerable:!0,writable:!0,value:function(){var e=Array.prototype.slice.call(arguments),n=document.createDocumentFragment();e.forEach(function(e){var t=e instanceof Node;n.appendChild(t?e:document.createTextNode(String(e)))}),this.insertBefore(n,this.firstChild)}});var loadDeferredStyles=function(){var e=document.getElementById("deferred-styles"),t=document.createElement("div");t.innerHTML=e.textContent,document.head.prepend(t),e.parentElement.removeChild(e),raf>window.requestAnimationFrame||window.mozRequestAnimationFrame||window.webkitRequestAnimationFrame||window.msRequestAnimationFrame;raf(function(){window.setTimeout(loadDeferredStyles,0)});window.addEventListener("load",loadDeferredStyles)</script><script>function(t,h,e,j,s,n){t.hj=t.hj||[],t.hj.q||[].push(arguments)},t._hjSettings={hjid:301739,hjsv:6},s=h.getElementsByTagName("head")[0],(n=h.createElement("script")).async=1,n.src="https://static.hotjar.com/c/hotjar-".t._hjSettings.hjid+".js?sv=".t._hjSettings.hjsv,s.appendChild(n)(window,document)</script></body></html>
```

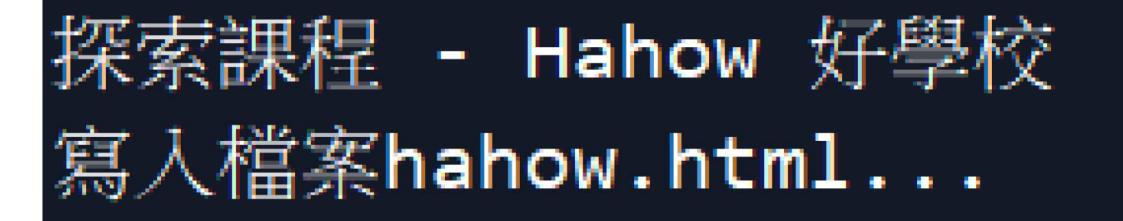
JavaScript 動態網頁擷取

儲存「HaHow」課程資訊的動態網頁

- ◆ 為了分析動態網頁內容，我們可以使用Selenium取得JavaScript產生的網頁內容，然後儲存成靜態網頁。
- ◆ 我們修改之前的程式，改儲存`https://hahow.in/courses`課程資料的網頁內容，程式碼載入HaHow網站的課程資訊後，

使用BeautifulSoup剖析儲存成`hahow.html`的HTML網頁檔案，執行結果如下

```
1 from selenium import webdriver
2 from bs4 import BeautifulSoup
3
4 driver = webdriver.Chrome("chromedriver")
5 driver.implicitly_wait(5)
6 driver.get("https://hahow.in/courses")
7 print(driver.title)
8 soup = BeautifulSoup(driver.page_source, "Lxml")
9 fp = open("hahow.html", "w", encoding="utf8")
10 fp.write(soup.prettify())
11 print("寫入檔案hahow.html...")
12 fp.close()
13 driver.quit()
```



探索課程 - HaHow 好學校
寫入檔案hahow.html...

JavaScript 動態網頁擷取

分析Hahow課程資訊的靜態網頁內容

- ◆ 在成功將Hahow網站的課程資訊儲存成hahow.html網頁檔案後，這是一份靜態網頁，我們可以啟動 Chrome開啟HTML網頁檔案和使用開發人員工具來分析網頁內容(當Chrome開啟本機網頁檔案時，我們無法使用Selector Gadget和XPath Helper工具)。
- ◆ 由於Hahow網站會隨時更新內容，所以您看到的畫面會有所出入。



JavaScript 動態網頁擷取

分析Hahow課程資訊的靜態網頁內容

- ◆ 請選擇Elements標籤前方箭頭，可以在左方網頁選取HTML元素，以此例是選方框中的課程名稱，其HTML標籤如下，課程名稱是

標籤，class屬性值有title、marg-t-20 marg-b-10，在分析後可知每一門課程名稱都是標籤，選取所有課程名稱的CSS選擇器如下所示。



The screenshot shows a course card for "超人氣吸睛表達力 - 職場必學表達" on the Hahow platform. The card includes a thumbnail, the course title, a brief description, a price of NT\$ 1280, and a progress bar indicating 20/30 people have funded it. On the right, the developer tools' element inspector is open, highlighting the course title with a blue box. The element details show the HTML structure:

```
▼ <a href="/courses/551198a738239d1000577870">
  ▼ <div class="text pad-r1-15">
    <h4 class="title marg-t-20 marg-b-10">
      超人氣吸睛表達力一職場必學表達課！
    </h4> == $0
  ▶ <div class="meta-description">...</div>
```

JavaScript 動態網頁擷取

Selenium 的JavaScript動態網頁擷取

- ◆ 現在我們可以建立Python程式擷取JavaScript動態產生的網頁內容，即取出<https://hahow.in/courses>網頁的所有課程名稱，程式碼載入課程網頁後，呼叫`find_elements_by_css_selector()`函數取出所有課程名稱的HTML元素`<h4>`，然後用`for`迴圈一一顯示課程名稱。

```
1 from selenium import webdriver
2
3 driver = webdriver.Chrome("chromedriver")
4 driver.implicitly_wait(8)
5 url = "https://hahow.in/courses"
6 driver.get(url)
7
8 items = driver.find_elements_by_css_selector("h4.title")
9
10 for item in items:
11     print(item.text)
12
13 driver.quit()
```

超人氣吸睛表達力－職場必學表達課！
為什麼我買的股票都不會漲？888機器人教你
提升工程師的科技力！AWS 雲端網站建置
角色動起來：3ds Max 角色骨架與權重
一滑就看完！條漫繪製攻略
用 Python 理財：打造自己的 AI 股票理專
淺談設計營運：給設計團隊的版本控制技巧
打造被動收入：建立人生主導權的系統化做法
3D基礎進階渲染 ■ Vray for Sketchup3.x
TALK SMART！打造更有深度的英語口說
學會 9 種場合適用妝容－小資女孩化妝術
行銷必上文案課：受眾溝通與表達
時間管理心法：升級人生作業系統
工作疲累的救星：放鬆筋膜不求人！
Python 入門特訓 - 基礎實作到證照攻略
百萬 YouTuber 阿滴－攻心剪輯術！
火頭工做麵包：用科學方法學做健康麵包
存零股學理財 - 累積你人生的第一張股王
啟動你的工匠魂：打造第一個皮件作品！

Selenium 換頁擷取 - 爬取樂透網站

◆ 台灣彩卷是一著名的彩卷網站，我們要使用Selenium自動擷取每一頁的樂透資料，網址如下：

http://www.taiwanlottery.com.tw/index_new.aspx

The screenshot shows the homepage of the Taiwan Lottery website. At the top, there's a navigation bar with the logo '中國信託金控 台灣彩券' on the left and a red heart logo with the text '公益彩券 做公益 積功德' on the right. Below the navigation bar, the main content area features several large lottery result displays:

- 威力彩**: Current jackpot estimate is \$0545020903.
- 大樂透**: Current jackpot estimate is \$0167153037.

On the left side, there's a sidebar for '刮刮樂' (Scratch-off) with sections for '熱賣中' (Hot Sale) and '預告' (Forecast). It lists four scratch-off games:

- 66大順: Maximum prize 600,000元.
- 好運輪盤: Maximum prize 200,000元.
- 超級麻將: Total prize over 900 million元.

In the center, there's a banner for the '57彩券王' (57 Lottery King) show, which airs every Monday to Saturday at 8:30 PM on 57 East森財經新聞. It also features a '最新消息 HOT NEWS' section with links to various news articles.

On the right side, there are two prominent promotional banners:

- A yellow banner for '威力彩' with the text '未滿十八歲者 不得購買或兌領彩券' (Those under 18 years old shall not purchase or cash in lottery tickets).
- A pink banner for '公益彩券 讓愛成真' (Lottery for Good Causes, Let Love Come True) with the text '感人影片 點我馬上看' (Watch感動影片) and a large '55個 55,000元' (55 winners, 55,000元).

Selenium 換頁擷取 - 大樂透

```
1 from bs4 import BeautifulSoup as bs
2 from selenium import webdriver
3 from selenium.webdriver.support.ui import Select
4 from matplotlib import pyplot as plt
5 import numpy as np
6 import pandas as pd
7
8
9 #用以儲存所爬到的大樂透號碼
10 lotto_list=[]
11
12 driver=webdriver.Chrome(r'C:\Users\Alex\Desktop\chromedriver_win32\chromedriver')
13 driver.get('http://www.taiwanlottery.com.tw/Lotto/Lotto649/history.aspx')
14
15
16 #勾選要以年月查詢的選項
17 driver.find_element_by_id('Lotto649Control_history_radYM').click()
```

Selenium 換頁擷取 - 大樂透

```
21 while True:
22     select_year= Select(driver.find_element_by_id('Lotto649Control_history_dropYear'))
23     year=input('請輸入你要找的樂透年份(國曆103年開始)\t')
24     print('請稍候~~~')
25     select_year.select_by_value(year)
26     for i in range(12):
27         #找出選擇月份的標籤
28         select_month=Select(driver.find_element_by_id('Lotto649Control_history_dropMonth'))
29         select_month.select_by_value(str(i+1))
30
31         #點擊「查詢」按鈕
32         driver.find_element_by_id('Lotto649Control_history_btnSubmit').click()
33
34         #抓取網頁內容
35         html=driver.page_source
36         soup=bs(html, 'html.parser')
37
38         #數網頁中有多少個table
39         table_count=len(soup.findAll('table',{'class':'td_hm'}))
40         #針對每一個table抓取樂透號碼並加入串列
41         for i in range(table_count):
42             for j in range(1,7):
43                 temp=soup.find('span', {'id':'Lotto649Control_history_dlQuery_No' + str(j) + '_'+str(i)})
44                 lotto_list.append(int(temp.text))
45         check=input('還要繼續嗎？繼續請輸入Y\t')
46         if check.upper()!='Y':
47             print('已結束，謝謝！')
48         break
```

JavaScript 動態網頁擷取

使用Selenium 擷取下一頁資料

- ◆ 如果爬取的資料是有很多分頁的表格資料，而且每一頁分頁都是JavaScript動態產生的網頁內容，Selenium可以模擬按下一頁切換表格的分頁，和抓取下一頁表格資料。
- ◆ 例如NBA官網球員以得分排序的統計資料是分頁的HTML表格(其11頁)，其網址如下：
<http://stats.nba.com/players/traditional/?sort=PTS&dir=-1>。
- ◆ 在分頁HTML表格按右下方箭頭鈕，就會使用JavaScript程式碼切換至下一頁的分頁表格，請使用Chrome開發人員工具取得HTML表格的CSS選擇器和>鈕的XPath表達式。

JavaScript 動態網頁擷取

使用Selenium 擷取下一頁資料

- ◆ 以下我們將使用Selenium爬取所有NBA球員每場賽事平均的統計資料，
- ◆ 程式是使用Selenium自動按表格的下一頁鈕，因為是分頁的HTML表格，所以直接使用Pandas套件的read_html()函數來爬取HTML表格資料。
- ◆ 首先匯入相關模組和套件，程式碼匯入Pandas套件和time模組，然後載入NBA統計資料的網頁後，使用下方while迴圈取全部11頁分頁的HTML表格資料。

```
1 from selenium import webdriver
2 from bs4 import BeautifulSoup
3 import pandas as pd
4 import time
5
6 driver = webdriver.Chrome("chromedriver")
7 driver.implicitly_wait(8)
8 driver.get("http://stats.nba.com/players/traditional/?sort=PTS&dir=-1")
```

JavaScript 動態網頁擷取

使用Selenium 擷取下一頁資料

- ◆ 以下的 pages_reaminging 變數判斷是否還有下一頁，while迴圈在使用Beautiful Soup剖析HTML網頁後使用select_one()函數取得表格標籤，然後呼叫Pandas的read_html()函數，傳回值是所有表格資料的清單(有可能不只一個表格)，然後呼叫df[0].to_csv()函數將取得的第一個表格資料寫成CSV檔案，檔名加上page_num變數的頁碼。

```
1 from selenium import webdriver
2 from bs4 import BeautifulSoup
3 import pandas as pd
4 import time
5
6 driver = webdriver.Chrome("chromedriver")
7 driver.implicitly_wait(8)
8 driver.get("http://stats.nba.com/players/traditional/?sort=PTS&dir=-1")
9
10 pages_remaining = True
11 page_num = 1
12 while pages_remaining:
13     # 使用Beautiful Soup剖析HTML網頁
14     soup = BeautifulSoup(driver.page_source, "lxml")
15     table = soup.select_one("body > main > div.stats-container_inner > \
16                             div > div.row > div > div > nba-stat-table > \
17                             div.nba-stat-table > div.nba-stat-table_overflow > table")
18     df = pd.read_html(str(table))
19     # print(df[0].to_csv())
20     df[0].to_csv("ALL_players_stats" + str(page_num) + ".csv")
21     print("儲存頁面:", page_num)
```

JavaScript 動態網頁擷取

使用Selenium 擷取下一頁資料

- ◆ 下方try/catch例外處理，處理沒有找到HTML元素的例外，我們是在try程式區塊處理Selenium模擬按下一頁鈕，當使用find_element_by_xpath()函數取得按鈕a元素後(沒有找到a元素即丟出例外)，呼叫click()函數模擬按下按鈕。
- ◆ 在等待5秒切換至下一頁後，即可繼續執行while迴圈擷取下一頁HTML表格資料。
- ◆ except程式區塊是當例外發生時，即按鈕元素不存在(表示沒有下一頁)，因為NBA網站的下一頁鈕不會消失，所以是使用if/else條件判斷是否已擷取11頁。

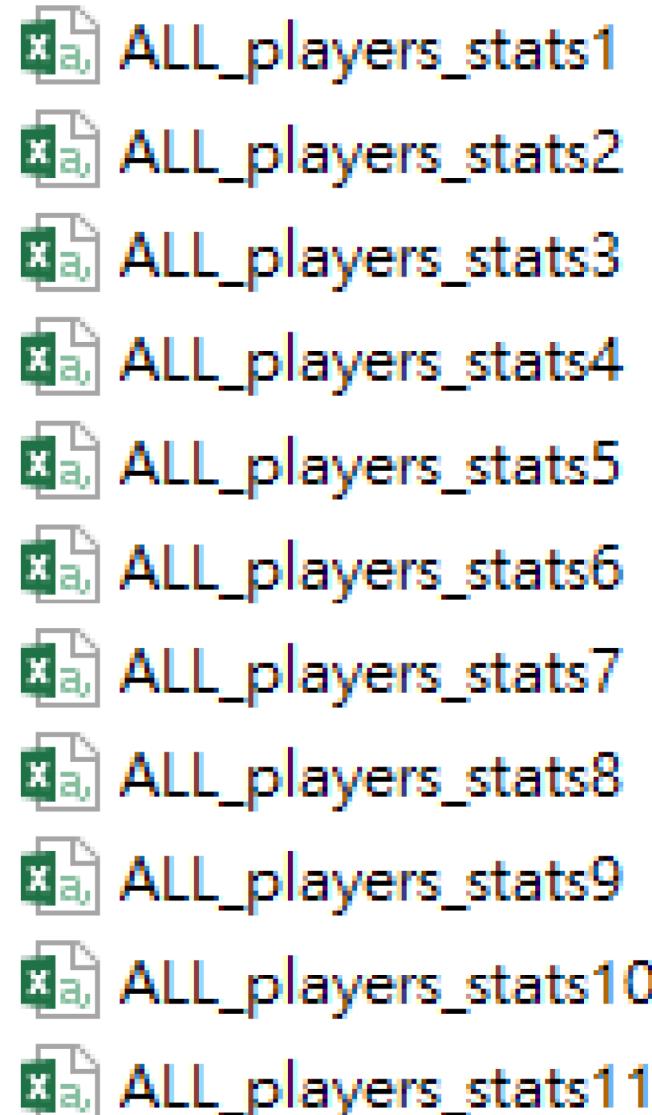
```
22     try:  
23         # 自動按下一頁按鈕  
24         next_link = driver.find_element_by_xpath('/html/body/main/div[2]/div/div[2]\  
25             /div/div/nba-stat-table/div[3]/div/div/a[2]')  
26         next_link.click()  
27         time.sleep(5)  
28         if page_num < 11:  
29             page_num = page_num + 1  
30         else:  
31             pages_remaining = False  
32     except Exception:  
33         pages_remaining = False  
34  
35 driver.close()
```

```
儲存頁面: 1  
儲存頁面: 2  
儲存頁面: 3  
儲存頁面: 4  
儲存頁面: 5  
儲存頁面: 6  
儲存頁面: 7  
儲存頁面: 8  
儲存頁面: 9  
儲存頁面: 10  
儲存頁面: 11
```

JavaScript 動態網頁擷取

使用Selenium 擷取下一頁資料

- ◆ 執行結果如下。在Python程式的同一目錄，可以看到新增共11個CSV檔案。



Q & A