

D33：反爬：登入授權模擬



簡報閱讀



範例與作業



問題討論

反爬：登入授權模擬 >

反爬：登入授權模擬

本日知識點目標 >

常見的反爬蟲機制有
些？ >

權限管理機制 >

登入有兩種實作方法 >



八

以 PTT 八卦版為例說明 >

利用 cookie/session 做登入 >

利用 tokens 做登入 >

重要知識點複習 >

Day 33 如何克服反制爬蟲的網站

反爬：登入授權模擬



出題教練：張維元



1

本日知識點目標



本日知識點目標

- 了解「登入權限機制」的反爬蟲機制
- 「登入權限機制」反爬蟲的因應策略

檢查 HTTP
標頭檔

驗證碼機制

登入權限機制

IP 黑/白名單

權限管理機制

大部分網站都有權限管理機制，使用上也會有登入/登出的機制。但由於爬蟲多半是基於 HTTP Request Response 一來一回的方式取資料。接下來我們將討論在爬蟲中要如何加上登入的做法。

登入有兩種實作方法

在開始講爬蟲登入之前，我們必須要知道現行的網站是如何做到登入這件事的。主要有兩種做法：



cookie/
session

token-
based

利用 cookie/session 做登入

cookie 是一種存放於瀏覽器的暫存空間，傳統的登入機制而會將驗證登入後的結果存在這裡，後續透過瀏覽器資料將 cookie 跟著 request 一起傳出去。所以 server 只要檢查 request 帶來的 cookie 是否存放正確的登入資訊，即可以判斷是否已登入過。

以 PTT 八卦版為例說明

當我們打開一個 PTT 八卦版，會直接跳到驗證的頁面：

本網站已依網站內容分級規定處理

警告：您即將進入之看板內容需滿十八歲方可瀏覽。

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。

我同意，我已年滿十八歲
進入

未滿十八歲或不同意本條款
離開

其實這個原理就跟「登入/權限」的做法很像，這個時候我們可以觀察瀏覽器所記錄的 cookie 資訊。

🔍

Elements

Console

Sources

Network

Performance

Memory

Application

Security

Audits

⋮

✕

🗑️ Clear storage

🔄 Filter

🚫 ✕

Storage

▶ 🗂️ Local Storage

▶ 🗂️ Session Storage

🗂️ IndexedDB

🗂️ Web SQL

▼ 🍪 Cookies

🌐 https://www.ptt.cc

Name	Value	Dom...	Path	Expir...	Size	Http...	Secure	Sam.
__cfduid	d5cbeecf757945ccf9e74fd0ff5f6bdb15...	.ptt.cc	/	2020...	51	✓		
_ga	GA1.2.1579753468.1574051640	.ptt.cc	/	2021...	30			
_gid	GA1.2.1174779314.1574051640	.ptt.cc	/	2019...	31			

Cache

這個時候當你按下同意（表示登入的行為），會發現 cookie 中多了一個 over18 = 1 的資料。

▶ Local Storage	_ga	GA1.2.1579753468.1574051640	.ptt.cc	/	2021...	30			
▶ Session Storage	_gat	1	.ptt.cc	/	2019...	5			
IndexedDB	_gid	GA1.2.1174779314.1574051640	.ptt.cc	/	2019...	31			
Web SQL	over18	1	www...	/	Sess...	7			
▼ Cookies									
🌐 https://www.ptt.cc									
Cache									

這樣的行為就像前面所提到的：「會在完成驗證行為之後，將資料記錄在瀏覽器當中」。因此爬蟲的做法，就是模仿「帶 Cookie 資訊」的行為。

利用 cookie/session 做登入

方法一

第一種做法，可以先模仿一個「登入」的請求，把這個請求的狀態保存，再接著發送第二次「取資料」的請求。

```
'from': '/bbs/Gossiping/index.html',
'yes': 'yes'
}

res = rs.post('https://www.ptt.cc/ask/over18', verify = False, data = payload)
res = rs.get('https://www.ptt.cc/bbs/Gossiping/index.html', verify = False)
soup = BeautifulSoup(res.text, 'html.parser')
print(soup.text)
```

方法二

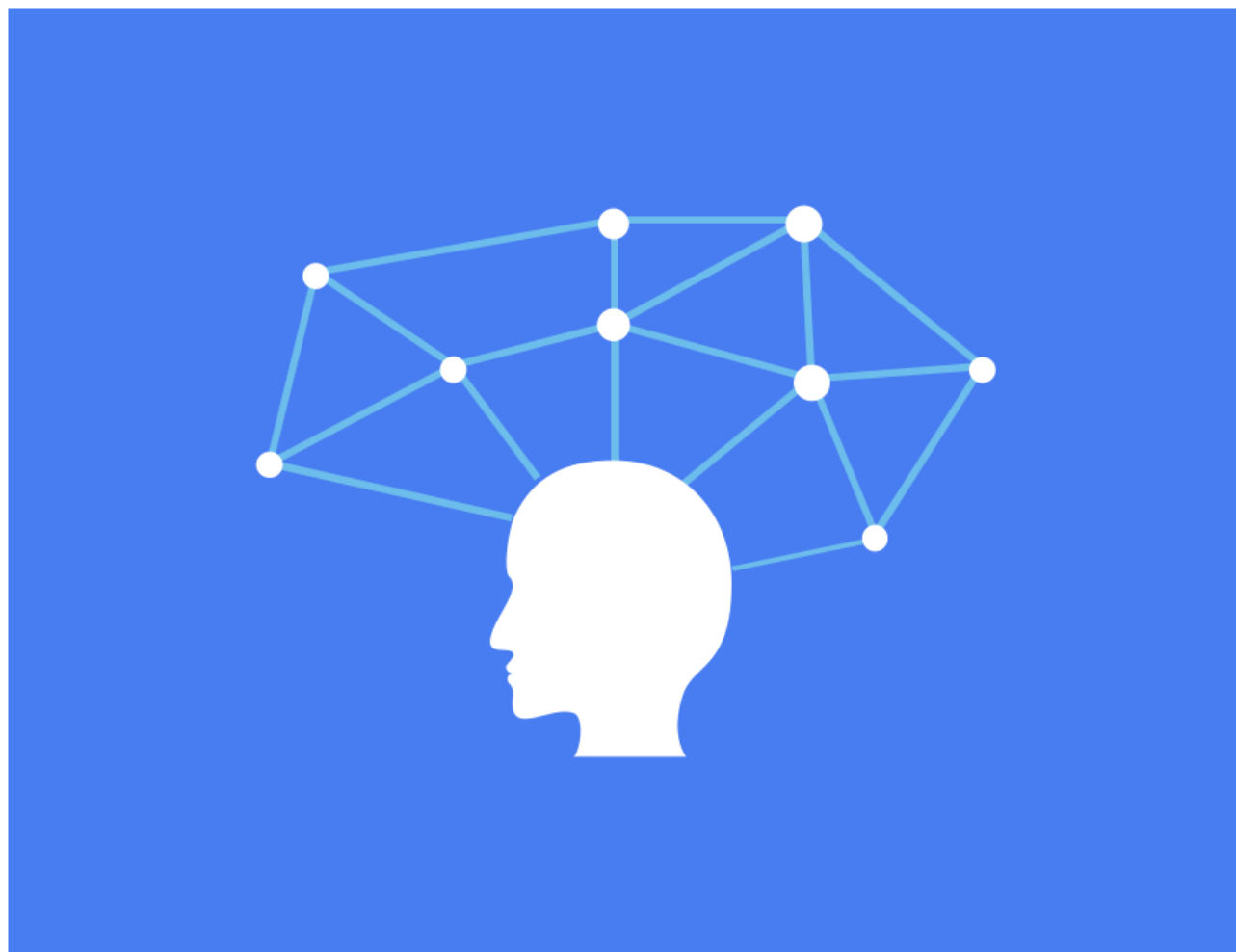
第二種做法，直接觀察瀏覽器記錄的資訊是什麼，將 cookie 帶在請求當中。

```
1 import requests
2 res = requests.get('https://www.ptt.cc/bbs/Gossiping/index.html', verify = False,
3 cookies={ 'over18': '1' })
soup = BeautifulSoup(res.text, 'html.parser')
print(soup.text)
```

利用 tokens 做登入

FB API 的用法，這裡就不示範了。

重要知識點複習



- 了解「登入權限機制」的反爬蟲機制

解題時間

解題時間 LET'S CRACK IT

Sample Code & 作業

開始解題



[下一步：閱讀範例與完成作業](#)



