

網路爬蟲經驗分享

講師：郭智榮 (Jhih-Rong Guo)

E-mail：109B30612@mailst.cjcu.edu.tw

*Department of Computer Science and Information Engineering
Chang Jung Christian University, Tainan, Taiwan*

課程大綱

- 網路爬蟲課程簡介
- 靜態爬蟲
 - 批踢踢實業坊文章內容爬取
 - 批踢踢實業坊文章列表爬取
- 網路爬蟲的圖片抓取
- 動態爬蟲
 - Selenium的網頁操作
- 網站後台資料分析及爬取
 - 長榮大學課程查詢系統資料爬取
- 期末成品引導

網路爬蟲課程簡介

僅供「長榮大學軟體開發社 111-1 社團課程」使用

何謂網路爬蟲

➤ 網路爬蟲(Web crawler)或稱Spider

– 自動化蒐集網站資料的方式

– 靜態爬蟲

- 若網站不屬於動態更新，則使用靜態爬蟲能以較低的效能及較短的時間獲得資料
- 常使用的套件有 BeautifulSoup 並搭配 urllib.request 或 requests 套件索取網站資料

– 動態爬蟲

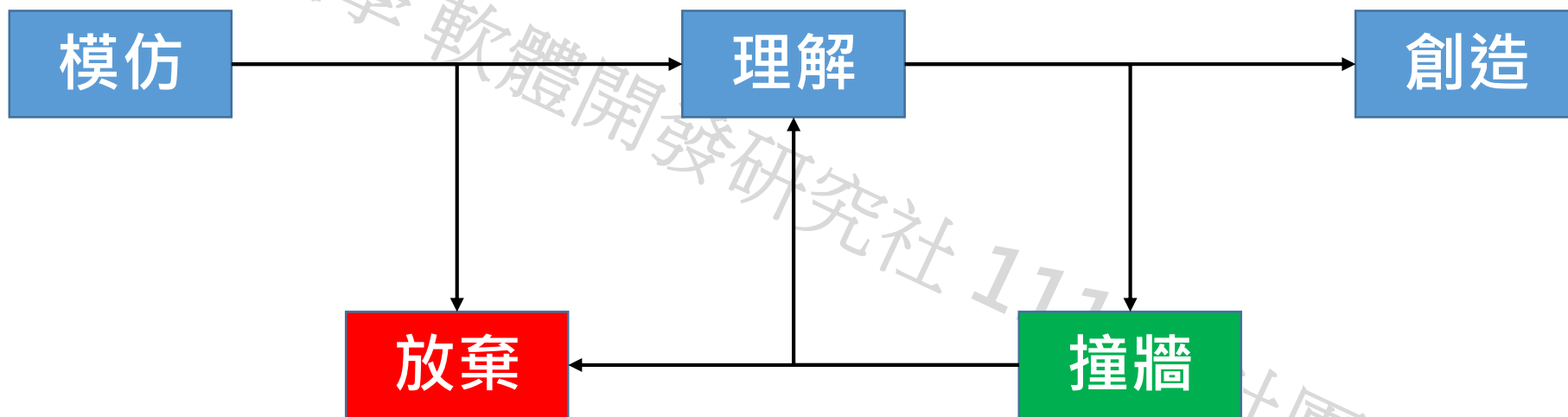
- 若需要的資料在網站中以動態更新，則需要使用動態爬蟲才能獲得資料
- 動態爬蟲能模擬一般使用者的操作，包含但不限於「選單選擇、畫面拉動、輸入文字、點擊」
- 常使用的套件有 WebDriver 並搭配特定瀏覽器，每種瀏覽器有各自的 Driver 檔需要安裝或呼叫

為何需要網路爬蟲

➤ 自動化蒐集資料

- 資料分析或多網站資料統整省去蒐集資料的時間
- 定時或單次到特定網站蒐集資料
 - 蒐集醫療諮詢資料，並訓練成醫療諮詢模型
 - 定時到股票網站中獲取最新的股價，分析股價趨勢或提供股價預測模型進行訓練
- 監看商品優惠、上架即時通知或自動訂購商品
 - 例如僅有 1 台 PS5，網路爬蟲能自動刷新頁面並查看是否上架或上架後自動購買

開始前要有的心理準備



靜態爬蟲 – Project1

批踢踢實業坊文章內容爬取

批踢踢實業坊文章內容爬取 – 題目敘述

➤ 題目要求

- 讓使用者可以自行輸入網址，請輸出提示訊息「請輸入批踢踢實業坊文章網址:」
- 抓取輸入的文章網址內的文章主文(不包含發文者資訊、留言或其他固定格式)

➤ 注意事項

- 批踢踢實業坊常有18+認證的按鈕

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

批踢踢實業坊文章內容爬取 – 輸入輸出

網頁畫面

作者 gstym (gstym) 看板 Gossiping
標題 [問卦] 台灣的銀行警示帳戶條件很鬆嗎?
時間 Tue Nov 15 17:51:50 2022

最近抓的這堆 都是怕人頭帳戶主人報警 所以乾脆直接綁架監禁一個禮拜左右
隔壁西台灣條件似乎嚴格很多 可以說是共產特色 這一兩年人頭帳戶也越來越貴 甚至很難買到了
西台灣只要被騙的人報案 沒多久銀行就鎖收款帳戶不能出款 根本不用等銀行在那慢慢查還等到一兩個禮拜才警示帳戶
台灣被騙的人都不會報案? 要等銀行過一個禮拜發現金流異常? 還是報案以後還要等警察+銀行慢慢查個一兩週才會鎖帳號?
這些流程速度484間接害到了被綁架的人?
烏卦謀?

Sent from JPTT on my iPhone

--
※ 發信站: 批踢踢實業坊(ptt.cc), 來自: 182.233.159.9 (臺灣)
※ 文章網址: <https://www.ptt.cc/bbs/Gossiping/M.1668505912.A.7D6.html>
→ cdmlin: 轉帳洗錢不一定是詐騙 1.34.229.213 11/15 17:53
金流異常銀行那裡一定馬上知道r 每個都被綁一週帳號才不能用
※ 編輯: gstym (182.233.159.9 臺灣), 11/15/2022 17:54:45
推 ghghfftjack: 上次警察告訴我 銀行說他們最近量 106.64.177.111 11/15 17:54
早點鎖帳戶至少還會放人 放了人才好抓被關的地點
→ ghghfftjack: 太大 要花些時 106.64.177.111 11/15 17:54
→ ah937609: 方便公司作業 1.200.45.109 11/15 17:54
→ ghghfftjack: 問 106.64.177.111 11/15 17:54
→ ghghfftjack: 具體花多久我不曉得 但配上最近的 106.64.177.111 11/15 17:55

輸出結果

請輸入批踢踢實業坊看板網址: <https://www.ptt.cc/bbs/Gossiping/M.1668505912.A.7D6.html>
最近抓的這堆 都是怕人頭帳戶主人報警 所以乾脆直接綁架監禁一個禮拜左右
隔壁西台灣條件似乎嚴格很多 可以說是共產特色 這一兩年人頭帳戶也越來越貴 甚至很難買到了
西台灣只要被騙的人報案 沒多久銀行就鎖收款帳戶不能出款 根本不用等銀行在那慢慢查還等到一兩個禮拜才警示帳戶
台灣被騙的人都不會報案? 要等銀行過一個禮拜發現金流異常? 還是報案以後還要等警察+銀行慢慢查個一兩週才會鎖帳號?
這些流程速度484間接害到了被綁架的人?
烏卦謀?

Sent from JPTT on my iPhone

批踢踢實業坊文章內容爬取 – 主程式

```
import requests
from bs4 import BeautifulSoup

url = input('請輸入批踢踢實業坊文章網址: ')
```

引入套件及輸入資訊

大致固定的框架

```
headers = {"cookie": "over18=1",
"User-Agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/107.0.0.0 Safari/537.36 Edg/107.0.1418.35"}

request = requests.get(url, headers=headers)
request.encoding = 'utf-8'

soup = BeautifulSoup(request.text, 'html.parser')
```

```
content_of_web = soup.find(id='main-content')
```

定位目標tag

```
content_of_web = content_of_web.text
content_of_web = content_of_web.split('\n')
content_of_web = content_of_web [1:]

content_of_target = []
for i, content in enumerate(content_of_web):
    if (content == ''):
        continue
    if (content == '--'):
        if ((content_of_web[i+1][0] == '※') and (content_of_web[i+2][0] == '※')):
            break
        content_of_target.append(content)

print("\n".join(content_of_target))
```

資料清理及輸出

批踢踢實業坊文章內容爬取 – 程式開頭¹

➤ 引入所需套件並讀取使用者輸入的網址

```
import requests  
from bs4 import BeautifulSoup
```

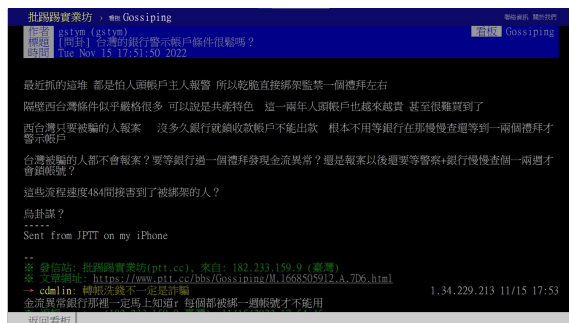
pip install bs4

```
url = input('請輸入批踢踢實業坊文章網址: ')
```

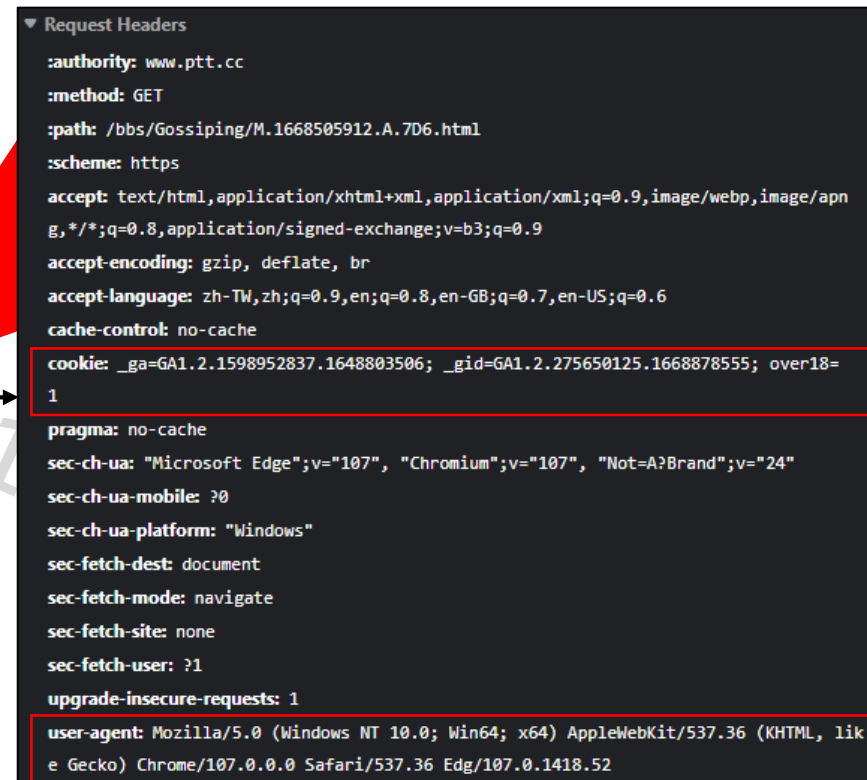
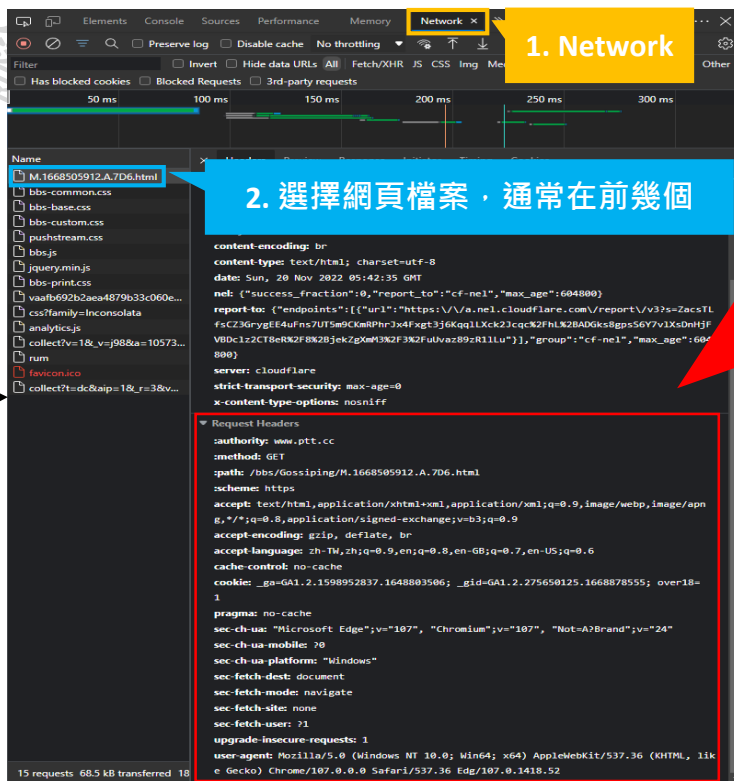
批踢踢實業坊文章內容爬取 – 設定Headers²

➤ 設定Headers資訊

```
headers = {"cookie": "over18=1",  
  "User-Agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/107.0.0.0 Safari/537.36 Edg/107.0.1418.35"}
```



+



批踢踢實業坊文章內容爬取－年齡驗證

僅供

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

使用

批踢踢實業坊文章內容爬取 – Cookies差異

先至不需驗證的看板紀錄必備的Cookies

Request Cookies ☐ show filtered out request cookies

Name	Value	Do...	Path	Ex...	Size	Http...	Se...	Sa...	Sa...	Par...	P...
_ga	GA1.2.738087403.1668894340	.pt...	/	20...	29						Me...
_gid	GA1.2.1912591072.16688943...	.pt...	/	20...	31						Me...
_gat	1	.pt...	/	20...	5						Me...

進行年齡驗證並查看新增的Cookies

Request Cookies ☐ show filtered out request cookies

Name	Value	Do...	Path	Ex...	Size	Http...	Se...	Sa...	Sa...	Par...	P...
_ga	GA1.2.738087403.1668894340	.pt...	/	20...	29						Me...
_gid	GA1.2.1912591072.16688943...	.pt...	/	20...	31						Me...
_gat	1	.pt...	/	20...	5						Me...
over18	1	w...	/	Se...	7						Me...

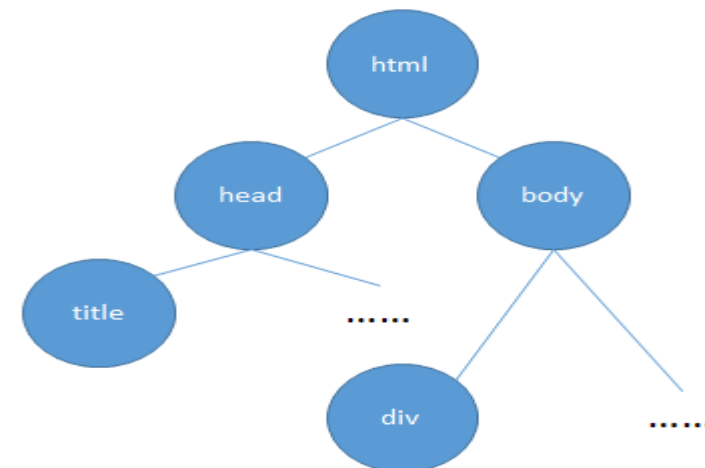
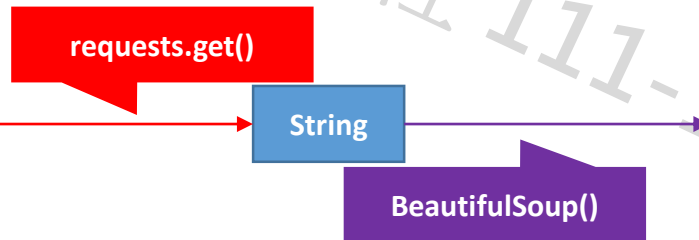
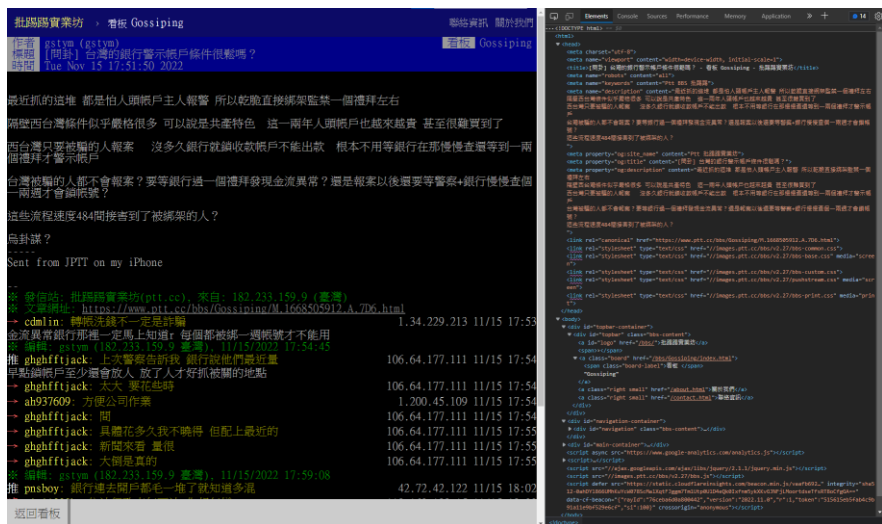
批踢踢實業坊文章內容爬取 – 獲取HTML原始碼³

➤ 發送網站要求並取得網站的原始碼資訊

```
request = requests.get(url, headers=headers)
request.encoding = 'utf-8'
```

➤ 以HTML的架構去解析獲取的原始碼

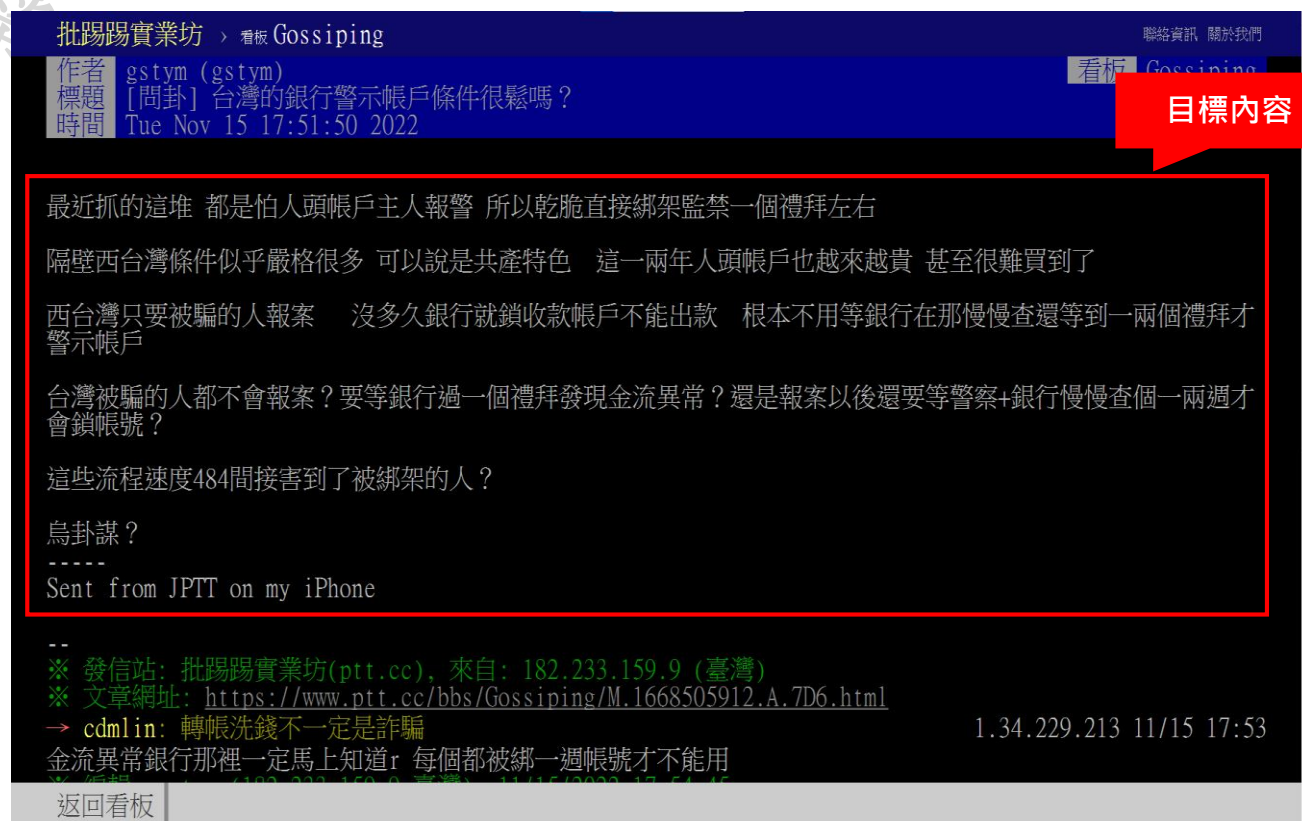
```
soup = BeautifulSoup(request.text, 'html.parser')
```



批踢踢實業坊文章內容爬取 – 定位tag⁴

➤ 透過id定位目標tag

```
content_of_web = soup.find(id='main-content')
```



批踢踢實業坊 > 看板 Gossiping

作者 gstym (gstym) 看板 Gossiping
標題 [問卦] 台灣的銀行警示帳戶條件很鬆嗎?
時間 Tue Nov 15 17:51:50 2022

目標內容

最近抓的這堆 都是怕人頭帳戶主人報警 所以乾脆直接綁架監禁一個禮拜左右
隔壁西台灣條件似乎嚴格很多 可以說是共產特色 這一兩年人頭帳戶也越來越貴 甚至很難買到了
西台灣只要被騙的人報案 沒多久銀行就鎖收款帳戶不能出款 根本不用等銀行在那慢慢查還等到一兩個禮拜才
警示帳戶
台灣被騙的人都不會報案? 要等銀行過一個禮拜發現金流異常? 還是報案以後還要等警察+銀行慢慢查個一兩週才
會鎖帳號?
這些流程速度484間接害到了被綁架的人?
烏卦謀?

Sent from JPTT on my iPhone

--
※ 發信站: 批踢踢實業坊(ptt.cc), 來自: 182.233.159.9 (臺灣)
※ 文章網址: <https://www.ptt.cc/bbs/Gossiping/M.1668505912.A.7D6.html>
→ cdm1in: 轉帳洗錢不一定是詐騙 1.34.229.213 11/15 17:53
金流異常銀行那裡一定馬上知道r 每個都被綁一週帳號才不能用
※ 編輯: 182.233.159.9 (臺灣), 11/15/2022 17:51:45

返回看板

點擊前往文章

批踢踢實業坊文章內容爬取 – 如何定位目標

2. 開啟網頁元素檢查工具

1. Element

3-2. 藍色透明框定位到的tag原始碼

4. 在定位的元素內找到目標內容

3. 將滑鼠放到要定位的元素上

3-1. 藍色透明框定位到的tag資訊

點擊前往文章

批踢踢實業坊文章內容爬取 – 元素內容獲取⁵

➤ 獲取目標tag的文字

```
content_of_web = content_of_web.text
```

批踢踢實業坊 > 看板 Gossiping 聯絡資訊 關於我們

作者 gstym (gstym) 看板 Gossiping
標題 [問卦] 台灣的銀行警示帳戶條件很鬆嗎?
時間 Tue Nov 15 17:51:50 2022

最近抓的這堆 都是怕人頭帳戶主人報警 所以乾脆直接綁架監禁一個禮拜左右

隔壁西台灣條件似乎嚴格很多 可以說是共產特色 這一兩年人頭帳戶也越來越貴 甚至很難買到了

西台灣只要被騙的人報案 沒多久銀行就鎖收款帳戶不能出款 根本不用等銀行在那慢慢查還等到一兩個禮拜才警示帳戶

台灣被騙的人都不會報案? 要等銀行過一個禮拜發現金流異常? 還是報案以後還要等警察+銀行慢慢查個一兩週才會鎖帳號?

這些流程速度484間接害到了被綁架的人?

烏卦謀?

Sent from JPTT on my iPhone

--
※ 發信站: 批踢踢實業坊(ptt.cc), 來自: 182.233.159.9 (臺灣)
※ 文章網址: <https://www.ptt.cc/bbs/Gossiping/M.1668505912.A.7D6.html>
→ cdm1in: 轉帳洗錢不一定是詐騙 1.34.229.213 11/15 17:53
金流異常銀行那裡一定馬上知道r 每個都被綁一週帳號才不能用

返回看板

'作者gstym (gstym)看板Gossiping標題[問卦] 台灣的銀行警示帳戶條件很鬆嗎? 時間Tue Nov 15 17:51:50 2022\n最近抓的這堆 都是怕人頭帳戶主人報警 所以乾脆直接綁架監禁一個禮拜左右\n隔壁西台灣條件似乎嚴格很多 可以說是共產特色 這一兩年人頭帳戶也越來越貴 甚至很難買到了\n西台灣只要被騙的人報案 沒多久銀行就鎖收款帳戶不能出款 根本不用等銀行在那慢慢查還等到一兩個禮拜才警示帳戶\n台灣被騙的人都不會報案? 要等銀行過一個禮拜發現金流異常? 還是報案以後還要等警察+銀行慢慢查個一兩週才會鎖帳號? \n這些流程速度484間接害到了被綁架的人? \n烏卦謀? \n-----\nSent from JPTT on my iPhone\n\n※ 發信站: 批踢踢實業坊(ptt.cc), 來自: 182.233.159.9 (臺灣) 文章網址: https://www.ptt.cc/bbs/Gossiping/M.1668505912.A.7D6.html\n→ cdm1in: 轉帳洗錢不一定是詐騙 1.34.229.213 11/15 17:53\n金流異常銀行那裡一定馬上知道r 每個都被綁一週帳號才不能用\n編輯: gstym (182.233.159.9 臺灣), 11/15/2022 17:54:45\n推 ghghfftjack: 上次警察告訴我 銀行說他們最近量 106.64.177.111 11/15 17:54\n早點鎖帳戶至少還會放人 放了人才好抓被關的地點\n→ ghghfftjack: 太大 要花些時 106.64.177.111 11/15 17:54\n→ ah937609: 方便公司作業 1.200.45.109 11/15 17:54\n→ ghghfftjack: 間 106.64.177.111 11/15 17:54\n→ ghghfftjack: 具體花多久我不曉得 但配上最近的 106.64.177.111 11/15 17:55\n→ ghghfftjack: 新聞來看 量很 106.64.177.111 11/15 17:55\n→ ghghfftjack: 大倒是真的 106.64.177.111 11/15 17:55\n※ 編輯: gstym (182.233.159.9 臺灣), 11/15/2022 17:59:08\n推 pnsboy: 銀行連去開戶都毛一堆了就知道多混 42.72.42.122 11/15 18:02\n噓 mimi1020b: 依法行政查無不法 你想怎樣 118.160.133.15 11/15 18:20\n→ mimi1020b: 你想讓銀行給你杯這個郭喔 直接凍結 118.160.133.15 11/15 18:20\n→ mimi1020b: 然後被告 還要賠錢是嘛 啊我就依法行 118.160.133.15 11/15 18:20\n→ mimi1020b: 這就是程序時間 每天這麼多人要處理 118.160.133.15 11/15 18:21\n→ mimi1020b: 就你特殊喔z 118.160.133.15 11/15 18:21\n推 rickieyang: 西台灣鎖帳號同時也增加銀行收入 當 203.222.17.36 11/15 22:19\n→ rickieyang: 然鎖的快 連沒犯罪行為都一起鎖! 203.222.17.36 11/15 22:19'

批踢踢實業坊文章內容爬取 – 資料整理⁶

➤ 將字串轉為陣列

```
content_of_web = content_of_web.split('\n')
```

'作者gstym (gstym)看板Gossiping標題[問卦] 台灣的銀行警示帳戶條件很鬆嗎?時間Tue Nov 15 17:51:50 2022\n\n最近抓的這堆 都是怕人頭帳戶主人報警 所以乾脆直接綁架監禁一個禮拜左右\n\n隔壁西台灣條件似乎嚴格很多 可以說是共產特色 這一兩年人頭帳戶也越來越貴 甚至很難買到了\n\n西台灣只要被騙的人報案 沒多久銀行就鎖收款帳戶不能出款 根本不用等銀行在那慢慢查還等到一兩個禮拜才警示帳戶\n\n台灣被騙的人都不會報案?要等銀行過一個禮拜發現金流異常?還是報案以後還要等警察+銀行慢慢查個一兩週才會鎖帳號?\n\n這些流程速度484間接害到了被綁架的人?\n\n烏卦謀?\n\n-----\n\nSent from JPTT on my iPhone\n\n--*\n\n※ 發信站: 批踢踢實業坊(ptt.cc), 來自: 182.233.159.9 (臺灣)\n\n※ 文章網址: https://www.ptt.cc/bbs/Gossiping/M.1668505912.A.7D6.html\n\n→ cdmlin: 轉帳洗錢不一定是詐騙 1.34.229.213 11/15 17:53\n\n金流異常銀行那裡一定馬上知道r 每個都被綁一週帳號才不能用\n\n※ 編輯: gstym (182.233.159.9 臺灣), 11/15/2022 17:54:45\n\n推 ghghfftjack: 上次警察告訴我 銀行說他們最近量 106.64.177.111 11/15 17:54\n\n早點鎖帳戶至少還會放人 放了人才好抓被關的地點\n\n→ ghghfftjack: 太大 要花些時 106.64.177.111 11/15 17:54\n\n→ ah937609: 方便公司作業 1.200.45.109 11/15 17:54\n\n→ ghghfftjack: 間 106.64.177.111 11/15 17:54\n\n→ ghghfftjack: 具體花多久我不曉得 但配上最近的 106.64.177.111 11/15 17:55\n\n→ ghghfftjack: 新聞來看 量很 106.64.177.111 11/15 17:55\n\n→ ghghfftjack: 大倒是真的 106.64.177.111 11/15 17:55\n\n※ 編輯: gstym (182.233.159.9 臺灣), 11/15/2022 17:59:08\n\n推 pnsboy: 銀行連去開戶都毛一堆了就知道多混 42.72.42.122 11/15 18:02\n\n噓 mimi1020b: 依法行政查無不法 你想怎樣 118.160.133.15 11/15 18:20\n\n→ mimi1020b: 你想讓銀行給你杯這個郭喔 直接凍結 118.160.133.15 11/15 18:20\n\n→ mimi1020b: 然後被告 還要賠錢是嘛 啊我就依法行 118.160.133.15 11/15 18:20\n\n→ mimi1020b: 這就是程序時間 每天這麼多人要處理 118.160.133.15 11/15 18:21\n\n→ mimi1020b: 就你特殊喔z 118.160.133.15 11/15 18:21\n\n推 rickieyang: 西台灣鎖帳號同時也增加銀行收入, 當 203.222.17.36 11/15 22:19\n\n→ rickieyang: 然鎖的快, 連沒犯罪行為都一起鎖! 203.222.17.36 11/15 22:19\n\n'

```
[ '作者gstym (gstym)看板Gos...51:50 2022', '', '最近抓的這堆 都是怕人頭帳戶主人報警 所...綁...  
> special variables  
> function variables  
00: '作者gstym (gstym)看板Gossiping標題[問卦] 台灣的銀行警示帳戶條件很鬆嗎?時  
01: ''  
02: '最近抓的這堆 都是怕人頭帳戶主人報警 所以乾脆直接綁架監禁一個禮拜左右'  
03: ''  
04: '隔壁西台灣條件似乎嚴格很多 可以說是共產特色 這一兩年人頭帳戶也越來越貴 甚  
05: ''  
06: '西台灣只要被騙的人報案 沒多久銀行就鎖收款帳戶不能出款 根本不用等銀行在那  
07: ''  
08: '台灣被騙的人都不會報案?要等銀行過一個禮拜發現金流異常?還是報案以後還要等  
09: ''  
10: '這些流程速度484間接害到了被綁架的人?'  
11: ''  
12: '烏卦謀?'  
13: '-----'  
14: 'Sent from JPTT on my iPhone'  
15: ''  
16: '--'  
17: '※ 發信站: 批踢踢實業坊(ptt.cc), 來自: 182.233.159.9 (臺灣)'  
18: '※ 文章網址: https://www.ptt.cc/bbs/Gossiping/M.1668505912.A.7D6.html'  
19: ''  
Hold Alt key to switch to editor language hover
```

批踢踢實業坊文章內容爬取 – 資料清理⁷

➤ 將發文者資訊清除

```
content_of_web = content_of_web[1:]
```

```
[ '作者gstym (gstym)看板Gos...51:50 2022', '', '最近抓的這堆 都是怕人頭帳戶主人報警 所...綁架監禁一個禮拜左右', '', '隔壁西台灣條件似乎_> special variables> function variables00: '作者gstym (gstym)看板Gossiping標題[問卦] 台灣的銀行警示帳戶條件很鬆嗎? 時01: ''02: '最近抓的這堆 都是怕人頭帳戶主人報警 所以乾脆直接綁架監禁一個禮拜左右'03: ''04: '隔壁西台灣條件似乎嚴格很多 可以說是共產特色 這一兩年人頭帳戶也越來越貴 甚05: ''06: '西台灣只要被騙的人報案 沒多久銀行就鎖收款帳戶不能出款 根本不用等銀行在那07: ''08: '台灣被騙的人都不會報案? 要等銀行過一個禮拜發現金流異常? 還是報案以後還要等09: ''10: '這些流程速度484間接害到了被綁架的人?'11: ''12: '烏卦謀?'13: '-----'14: 'Sent from JPTT on my iPhone'15: ''16: '--'17: '※ 發信站: 批踢踢實業坊(ptt.cc), 來自: 182.233.159.9 (臺灣)'
```

Hold Alt key to switch to editor language hover

```
[ '', '最近抓的這堆 都是怕人頭帳戶主人報警 所...綁架監禁一個禮拜左右', '', '隔壁西台灣條件似乎_> special variables> function variables00: ''01: '最近抓的這堆 都是怕人頭帳戶主人報警 所以乾脆直接綁架監禁一個禮拜左右'02: ''03: '隔壁西台灣條件似乎嚴格很多 可以說是共產特色 這一兩年人頭帳戶也越來越貴 甚04: ''05: '西台灣只要被騙的人報案 沒多久銀行就鎖收款帳戶不能出款 根本不用等銀行在那06: ''07: '台灣被騙的人都不會報案? 要等銀行過一個禮拜發現金流異常? 還是報案以後還要等08: ''09: '這些流程速度484間接害到了被綁架的人?'10: ''11: '烏卦謀?'12: '-----'13: 'Sent from JPTT on my iPhone'14: ''15: '--'16: '※ 發信站: 批踢踢實業坊(ptt.cc), 來自: 182.233.159.9 (臺灣)'17: '※ 文章網址: https://www.ptt.cc/bbs/Gossiping/M.1668505912.A.7D6.html'18: '※ 文章標題: [問卦] 台灣的銀行警示帳戶條件很鬆嗎? 時
```

Hold Alt key to switch to editor language hover

批踢踢實業坊文章內容爬取 – 資料篩選⁸

➤ 資料篩選演算法

```
content_of_target = []
```

```
for i, content in enumerate(content_of_web):
```

依序將content_of_web的元素放入content，並給予i目前正在第幾個元素(0~(len(content_of_web)-1))

```
    if (content == ''):  
        continue
```

遇到陣列元素內是空字串則跳到下一個陣列元素

```
    if (content == '--'):  
        if ((content_of_web[i+1][0] == '※') and (content_of_web[i+2][0] == '※')):  
            break
```

若遇到「--」則判斷後兩個陣列元素內第一個字是否為「※」
若是則判斷為遇到結尾，因此需要跳出迴圈

```
    content_of_target.append(content)
```

執行到此可以確認為目標內容，因此放入陣列content_of_target中

```
print("\n".join(content_of_target))
```

批踢踢實業坊文章內容爬取 – 輸出方式

content_of_target

"\n".join(content_of_target)

```
[ '最近抓的這堆 都是怕人頭帳戶主人報警 所...綁架監禁一個禮拜左右', '隔壁西台灣條件似乎嚴格很多 ...  
> special variables  
> function variables  
0: '最近抓的這堆 都是怕人頭帳戶主人報警 所以乾脆直接綁架監禁一個禮拜左右'  
1: '隔壁西台灣條件似乎嚴格很多 可以說是共產特色 這一兩年人頭帳戶也越來越貴 甚至  
2: '西台灣只要被騙的人報案 沒多久銀行就鎖收款帳戶不能出款 根本不用等銀行在那  
3: '台灣被騙的人都不會報案?要等銀行過一個禮拜發現金流異常?還是報案以後還要等警  
4: '這些流程速度484間接害到了被綁架的人?'  
5: '烏卦謀?'  
6: '-----'  
7: 'Sent from JPTT on my iPhone'  
len(): 8
```

Hold Alt key to switch to editor language hover

最近抓的這堆 都是怕人頭帳戶主人報警 所以乾脆直接綁架監禁一個禮拜左右

隔壁西台灣條件似乎嚴格很多 可以說是共產特色 這一兩年人頭帳戶也越來越貴 甚至很難買到了

西台灣只要被騙的人報案 沒多久銀行就鎖收款帳戶不能出款 根本不用等銀行在那慢慢查還等到一兩個禮拜才警示帳戶

台灣被騙的人都不會報案?要等銀行過一個禮拜發現金流異常?還是報案以後還要等警察+銀行慢慢查個一兩週才會鎖帳號?

這些流程速度484間接害到了被綁架的人?

烏卦謀?

Sent from JPTT on my iPhone

靜態爬蟲 – Project2

批踢踢實業坊文章列表爬取

批踢踢實業坊文章列表爬取 – 題目敘述

➤ 題目要求

- 讓使用者可以自行輸入網址，請輸出提示訊息「請輸入批踢踢實業坊看板網址:」
- 抓取輸入的看板網址內的所有文章標題及連結

➤ 注意事項

- 批踢踢實業坊常有18+認證的按鈕
- 所抓取的網址可能會缺少前綴(如:https://www.ptt.cc/)，請自行補上
- 文章被刪除時需要另外處理，請參考「批踢踢實業坊文章列表爬取 – 輸入輸出」

批踢踢實業坊文章列表爬取 – 輸入輸出

網頁畫面

2	[本文已被刪除] [melissalewis]	11/15
8	[問卦] 萬年里長落選後會去哪? strmo722	11/15 ...
1	[新聞] 拜習會長談3小時！拜登重申一中政策不變 Caress	11/15 ...
	[新聞] 批選舉成人格毀滅戰 陳時中：北市民眼睛 lycppt	11/15 ...
12	[問卦] 台積電自前低漲了近百點?? s820912gmail	11/15 ...
3	Re: [問卦] 美國沒有賣雞屁股嗎? arnold3	11/15 ...
1	[問卦] 台灣巴肥特的台gg賣了沒? kink1999	11/15 ...
9	[問卦] 修機車請特休一天會很瞎嗎? Mopack22926	11/15 ...
3	[問卦] 正妹喜歡下棋，大家可以嗎? moshenisshit	11/15 ...
2	[問卦] Rap是唱rap還是唸rap jackeman	11/15 ...
4	[問卦] 不打仗+波克夏加持 該allin台積了吧? werqq	11/15 ...
5	[問卦] locklock水壺真的能耐熱100°C嗎 mjj90138	11/15 ...
35	[新聞] 巴菲特買進台積電ADR逾41億美元 罕見押注 NTKingsman	11/15 ...
7	[新聞] 護士空姐制服任選! 人夫用感情吸爆小三做 ha3810996	11/15 ...
5	[問卦] 現在恐、危、怕等侵台新聞哪家最多? engineer1	11/15 ...
9	Re: [新聞] 北市封關民調！蔣萬安36%領先、黃珊珊27% hurtmind	11/15 ...
11	[問卦] 龜仙人為什麼可以參加力之大會?? zzyyxx77	11/15 ...
7	Re: [新聞] 巴菲特買進台積電ADR逾41億美元 罕見押注 tamama000	11/15 ...
	[問卦] 包養正妹 leon771170	11/15 ...
	[新聞] 否認低薪高報詐領！高虹安扯「前立委吳 math520	11/15 ...

輸出結果

請輸入批踢踢實業坊看板網址: <https://www.ptt.cc/bbs/Gossiping/index38997.html>

第1篇文章: (本文已被刪除) [melissalewis] - 文章已被刪除

第2篇文章: [問卦] 萬年里長落選後會去哪? - <https://www.ptt.cc/bbs/Gossiping/M.1668474251.A.E47.html>

第3篇文章: [新聞] 拜習會長談3小時！拜登重申一中政策不變 - <https://www.ptt.cc/bbs/Gossiping/M.1668474311.A.DCE.html>

第4篇文章: [新聞] 批選舉成人格毀滅戰 陳時中：北市民眼睛 - <https://www.ptt.cc/bbs/Gossiping/M.1668474311.A.1D8.html>

第5篇文章: [問卦] 台積電自前低漲了近百點?? - <https://www.ptt.cc/bbs/Gossiping/M.1668474507.A.827.html>

第6篇文章: Re: [問卦] 美國沒有賣雞屁股嗎? - <https://www.ptt.cc/bbs/Gossiping/M.1668474579.A.9C3.html>

第7篇文章: [問卦] 台灣巴肥特的台gg賣了沒? - <https://www.ptt.cc/bbs/Gossiping/M.1668474656.A.B5A.html>

第8篇文章: [問卦] 修機車請特休一天會很瞎嗎? - <https://www.ptt.cc/bbs/Gossiping/M.1668474713.A.068.html>

第9篇文章: [問卦] 正妹喜歡下棋，大家可以嗎? - <https://www.ptt.cc/bbs/Gossiping/M.1668474770.A.73F.html>

第10篇文章: [問卦] Rap是唱rap還是唸rap - <https://www.ptt.cc/bbs/Gossiping/M.1668474801.A.EBA.html>

第11篇文章: [問卦] 不打仗+波克夏加持 該allin台積了吧? - <https://www.ptt.cc/bbs/Gossiping/M.1668474826.A.B92.html>

第12篇文章: [問卦] locklock水壺真的能耐熱100°C嗎 - <https://www.ptt.cc/bbs/Gossiping/M.1668474901.A.44A.html>

第13篇文章: [新聞] 巴菲特買進台積電ADR逾41億美元 罕見押注 - <https://www.ptt.cc/bbs/Gossiping/M.1668474917.A.095.html>

第14篇文章: [新聞] 護士空姐制服任選! 人夫用感情吸爆小三做 - <https://www.ptt.cc/bbs/Gossiping/M.1668475026.A.C52.html>

第15篇文章: [問卦] 現在恐、危、怕等侵台新聞哪家最多? - <https://www.ptt.cc/bbs/Gossiping/M.1668475140.A.E8F.html>

第16篇文章: Re: [新聞] 北市封關民調！蔣萬安36%領先、黃珊珊27% - <https://www.ptt.cc/bbs/Gossiping/M.1668475179.A.512.html>

第17篇文章: [問卦] 龜仙人為什麼可以參加力之大會?? - <https://www.ptt.cc/bbs/Gossiping/M.1668475195.A.094.html>

第18篇文章: Re: [新聞] 巴菲特買進台積電ADR逾41億美元 罕見押注 - <https://www.ptt.cc/bbs/Gossiping/M.1668475379.A.89B.html>

第19篇文章: [問卦] 包養正妹 - <https://www.ptt.cc/bbs/Gossiping/M.1668475477.A.F3E.html>

第20篇文章: [新聞] 否認低薪高報詐領！高虹安扯「前立委吳

批踢踢實業坊文章列表爬取 – 主程式

```
import requests
from bs4 import BeautifulSoup

url = input('請輸入批踢踢實業坊看板網址: ')
```

引入套件及輸入資訊

大致固定的框架

```
headers = {"cookie": "over18=1",
"User-Agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/107.0.0.0 Safari/537.36 Edg/107.0.1418.35"}

request = requests.get(url, headers=headers)
request.encoding = 'utf-8'

soup = BeautifulSoup(request.text, 'html.parser')
```

```
titles = soup.find_all(class_='title')
```

定位目標tag

```
for i, title in enumerate(titles):
    try:
        link = title.find('a')['href']

        if ('https' not in link):
            link = f'https://www.ptt.cc/{link}'

    except:
        link = '文章已被刪除'

    finally:
        title = title.text.replace('\n', '').replace('\t', '')

        print(f'第{i+1}篇文章: {title} - {link}')
```

資料整理及輸出

批踢踢實業坊文章內容爬取 – 程式開頭¹

➤ 以下資訊請參考「批踢踢實業坊文章內容爬取」的課程內容

```
import requests
from bs4 import BeautifulSoup

url = input('請輸入批踢踢實業坊看板網址: ')

headers = {"cookie": "over18=1",
"User-Agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/107.0.0.0 Safari/537.36 Edg/107.0.1418.35"}

request = requests.get(url, headers=headers)
request.encoding = 'utf-8'

soup = BeautifulSoup(request.text, 'html.parser')
```

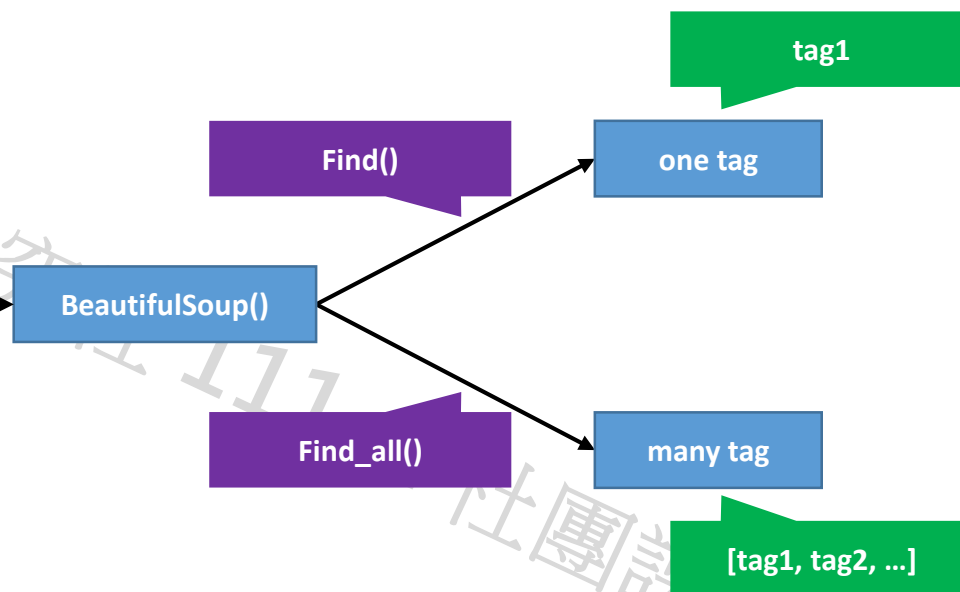
批踢踢實業坊文章內容爬取 – 定位tag²

➤ 以**Find_all**搭配class屬性的方式進行定位tag

```
titles = soup.find_all(class_='title')
```

2 (本文已被刪除) [melissalewis]	11/15
8 [問卦] 萬年里長落選後會去哪?	11/15 ...
1 [新聞] 拜習會長談3小時! 拜登重申一中政策不變	11/15 ...
[新聞] 批選舉成人格毀滅戰 陳時中:北市民眼睛	11/15 ...
12 [問卦] 台積電自前低漲了近百點??	11/15 ...
3 Re: [問卦] 美國沒有賣雞屁股嗎?	11/15 ...
1 [問卦] 台灣巴肥特的台gg賣了沒?	11/15 ...
9 [問卦] 修機車請特休一天會很騷嗎?	11/15 ...
3 [問卦] 正妹喜歡下棋, 大家可以嗎?	11/15 ...
2 [問卦] Rap是唱rap還是唸rap	11/15 ...
4 [問卦] 不打仗+波克夏加持 該allin台積電了吧?	11/15 ...
5 [問卦] locklock水壺真的能耐熱100°C嗎	11/15 ...
35 [新聞] 巴菲特買進台積電ADR逾41億美元 罕見押注	11/15 ...
7 [新聞] 護士空姐制服任選! 人夫用感情吸爆小三做	11/15 ...
5 [問卦] 現在恐、危、怕等侵台新聞哪家最多?	11/15 ...
9 Re: [新聞] 北市封關民調! 蔣萬安36%領先、黃珊珊27%	11/15 ...
11 [問卦] 龜仙人為什麼可以參加力之大會??	11/15 ...
7 Re: [新聞] 巴菲特買進台積電ADR逾41億美元 罕見押注	11/15 ...
[問卦] 包養正妹	11/15 ...
[新聞] 否認低薪高報詐領! 高虹安批「前立委與	11/15 ...

點擊前往文章



批踢踢實業坊文章內容爬取 – 如何定位多個相同目標

[illegible]

[點擊前往文章](#)

批踢踢實業坊文章內容爬取 – 資料篩選³

➤ 資料篩選演算法

```
for i, title in enumerate(titles):  
    try:
```

```
        link = title.find('a')['href']
```

如果文章已刪除，再找tag <a>會變成None，導致TypeError

```
        if ('https' not in link):  
            link = f'https://www.ptt.cc/{link}'
```

網址沒有前綴，因此要補上該網站的前綴

```
    except:
```

```
        link = '文章已被刪除'
```

若沒有link就要自己補，否則print(link)會Error

```
    finally:
```

```
        title = title.text.replace('\n', '').replace('\t', '')
```

抓取標題文字並處理多餘符號(\t為一個tab鍵)

```
    print(f'第{i+1}篇文章: {title} - {link}')
```

<div><a>text</div>

Find('a')

<a>text

Exception has occurred: TypeError ×
'NoneType' object is not subscriptable
File "D:\Program\CJCU\CJCU_Club\111-1\
link = title.find('a')['href']

```
try:  
    num1 = 10  
    num2 = 1/0  
except:  
    num1 = 0  
    num2 = -1  
finally:  
    print(num1, num2)
```

Exception has occurred: ZeroDivisionError
division by zero

File "D:\Program\CJCU\CJCU_Club\hi
num2 = 1/0

running

0 -1

靜態爬蟲 – Project3

課堂練習

課堂練習 – 題目敘述

➤ 題目要求

- 讓使用者可以自行輸入網址，請輸出提示訊息「請輸入批踢踢實業坊看板網址:」
- 抓取輸入的看板網址內所有文章標題及連結，並延伸抓取每篇文章的內容
- 按照「課堂練習 – 輸入輸出」規格，將相關訊息輸出

➤ 注意事項

- 批踢踢實業坊常有18+認證的按鈕
- 所抓取的網址可能會缺少前綴(如:https://www.ptt.cc/)，請自行補上
- 文章被刪除時需要另外處理，請參考「課堂練習 – 輸入輸出」

課堂練習 – 輸入輸出

網頁畫面1

2 [本文已被刪除] [melissalewis]	11/15	...
8 [問卦] 萬年里長落選後會去哪？ strmf22	11/15	...
1 [新聞] 拜習會長談3小時！拜登重申一中政策不變 Caress	11/15	...
[新聞] 批選舉成人格毀滅戰 陳時中：北市民眼瞞 lycpot	11/15	...
12 [問卦] 台積電自前低漲了近百點？ ?? s828912gmail	11/15	...
3 Re: [問卦] 美國沒有賣雅尼股嗎？ arnold3	11/15	...
1 [問卦] 台灣巴肥特的台gg賣了沒？ klnki999	11/15	...
9 [問卦] 修機車請特休一天會很噁嗎？ Mopack22926	11/15	...
3 [問卦] 正妹喜歡下棋，大家可以嗎？ moshenisshit	11/15	...
2 [問卦] Rap是唱rap還是唸rap Jackeman	11/15	...
4 [問卦] 不打仗+波克夏加持 該allin台積電了吧？ werqq	11/15	...
5 [問卦] locklock水壺真的能耐熱100°C嗎 mjj98138	11/15	...
35 [新聞] 巴菲特買進台積電ADR逾41億美元 罕見押注 NTKingsman	11/15	...
7 [新聞] 護士空姐制服任選1人夫用感情吸爆小三做 ha3818996	11/15	...
5 [問卦] 現在疫、危、怕等侵台新聞專家最多？ engineer1	11/15	...
9 Re: [新聞] 北市封關民調：蔣萬安36%領先、黃珊珊27% hurtalnd	11/15	...
11 [問卦] 龜仙人為什麼可以參加力之大會?? zzyyxx77	11/15	...
7 Re: [新聞] 巴菲特買進台積電ADR逾41億美元 罕見押注 tanana088	11/15	...
[問卦] 包養正妹 leon771170	11/15	...
[新聞] 否認低薪高報詐領！高虹安批「前立委吳 math528	11/15	...

網頁畫面2

strmf22 (海綿寶哥)

看板

Gossiping

【問卦】萬年里長落選後會去哪？

Tue Nov 15 09:04:09 2022

現在一堆年輕正妹出來選里長

假設她們都選上然後當了四個好了

如果現在26歲加上16年

那時都已經42了，沒有里長當又沒有其他工作經驗

那時候他們都要做什麼啊？

※ 發信站: 批踢踢實業坊(ptt.cc), 來自: 114.136.139.131 (臺灣)

※ 文章網址: <https://www.ptt.cc/bbs/Gossiping/M.1668474251.A.E47.html>

→ tzonren: 里長選完通常就會往更高的爬拉

117.56.248.91 11/15 09:04

→ tzonren: 蔡B八

117.56.248.91 11/15 09:04

推 sd09090: 回去工作

1.200.19.11 11/15 09:05

→ charlie01: 反正都過得比你你好 (無誤)

220.143.191.64 11/15 09:05

→ neillisme: 撈4屆都能退休了

61.216.64.210 11/15 09:05

噓 ymb: 早賺飽了，可以過你十輩子的積蓄

1.200.58.16 11/15 09:05

→ ePaper: 上去選議員了阿還在基層幹嘛

114.36.2.195 11/15 09:05

推 dearl33: 他們有錢不需擔心

61.228.78.180 11/15 09:05

推 SKiii: 移民享清福了

1.175.109.245 11/15 09:06

輸出結果

請輸入批踢踢實業坊看板網址: https://www.ptt.cc/bbs/Gossiping/index38997.html			
第1篇文章: (本文已被刪除) [melissalewis] - 文章已被刪除			
文章已被刪除			

第2篇文章: 【問卦】萬年里長落選後會去哪？ - https://www.ptt.cc/bbs/Gossiping/M.1668474251.A.E47.html			
現在一堆年輕正妹出來選里長			
假設她們都選上然後當了四屆好了			
如果現在26歲加上16年			
那時都已經42了，沒有里長當又沒有其他工作經驗			
那時候他們都要做什麼啊？			

第3篇文章: 【新聞】拜習會長談3小時！拜登重申一中政策不變 - https://www.ptt.cc/bbs/Gossiping/M.1668474311.A.DCE.html			
1. 媒體來源:			
聯合報			
2. 記者署名:			
陳章廷			
3. 完整新聞標題:			
拜習會長談3小時！拜登重申一中政策不變 反對大陸脅迫台灣			
4. 完整新聞內容:			
美國總統拜登14日與中國大陸國家主席習近平會談時表示，美國的一中政策不變，反對片面改變現狀，維護台海和平穩定符合世界利益。			
拜登也對中國脅迫台灣、日益挑釁行為、破壞台海和更廣泛區域和平穩定、危害全球繁榮之舉，提出美方異議。			
5. 完整新聞連結 (或短網址)需放媒體原始連結，不可用轉載媒體連結:			
https://udn.com/news/story/6811/6764928			
6. 備註:			
..			
正港台灣人的電視台--蕃薯台			

第4篇文章: 【新聞】批選舉成人格毀滅戰 陳時中：北市民眼瞞 - https://www.ptt.cc/bbs/Gossiping/M.1668474311.A.108.html			
馬偉熙			

Thanks for listening