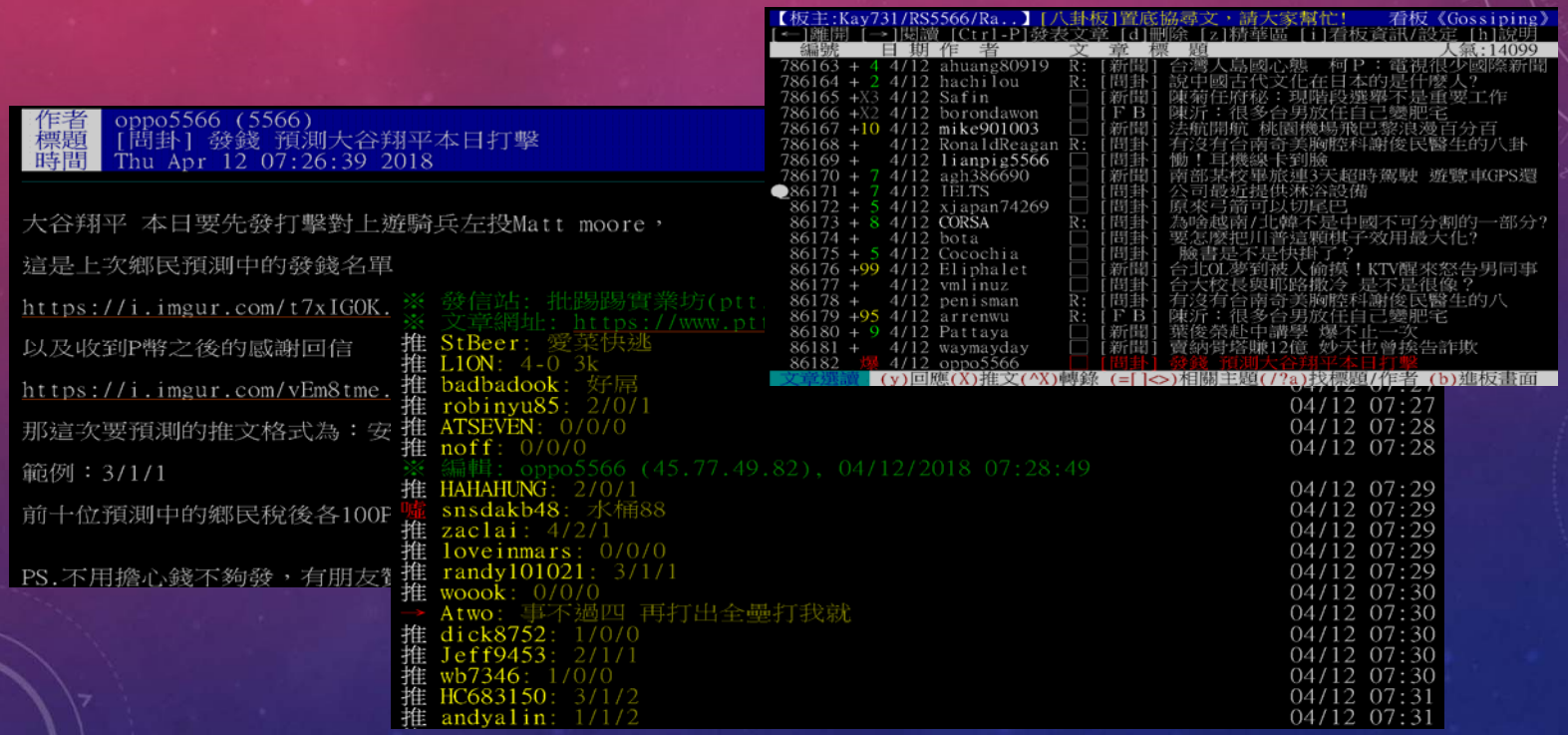
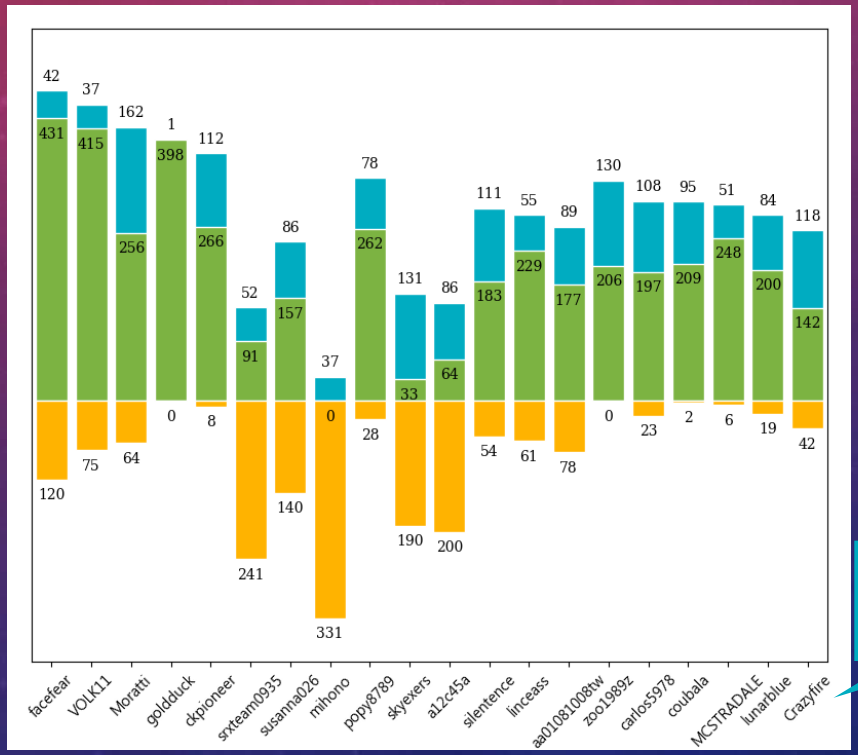
WEB CRAWLER

張家瑋 博士
新漢股份有限公司創新工業4.0中心顧問

PTT Introduction



Top 20 踴躍留言者分析的圖表實作



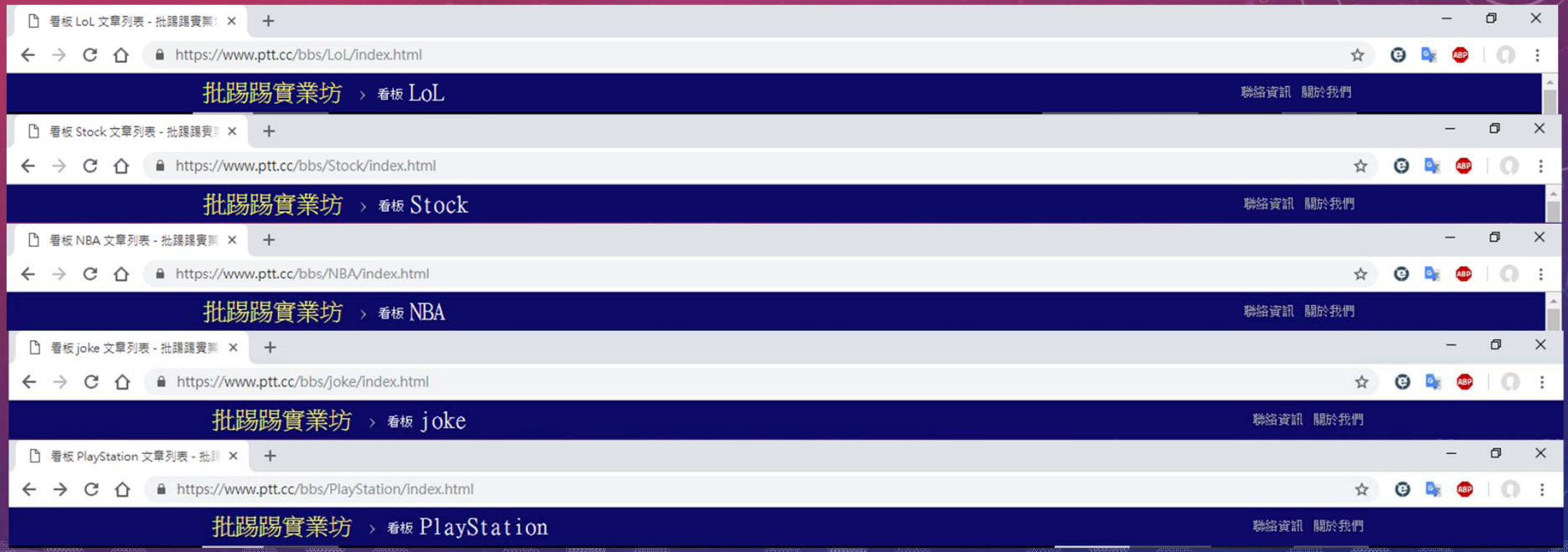
情緒辨識

輸入留言！偵測你的情緒狀態：罪有英德
罪有英德 ---你很不開心喔？
輸入留言！偵測你的情緒狀態：賴清德
賴清德 ---你很不開心喔？
輸入留言！偵測你的情緒狀態：蔡英文
蔡英文 ---感覺還不錯哦！
輸入留言！偵測你的情緒狀態：勞動部
勞動部 ---你很不開心喔？
輸入留言！偵測你的情緒狀態：勞基法
勞基法 ---你很不開心喔？
輸入留言！偵測你的情緒狀態：健保
健保 ---你很不開心喔？

「今天來帶各位用PYTHON爬PTT」

做個友善的英雄聯盟玩家，一起支持G-REX
https://www.ptt.cc/bbs/LoL/index.html

逛PTT的過程中你會發現PTT的URL網址都有一個固定格式
https://www.ptt.cc/bbs/<看板名稱>/index.html



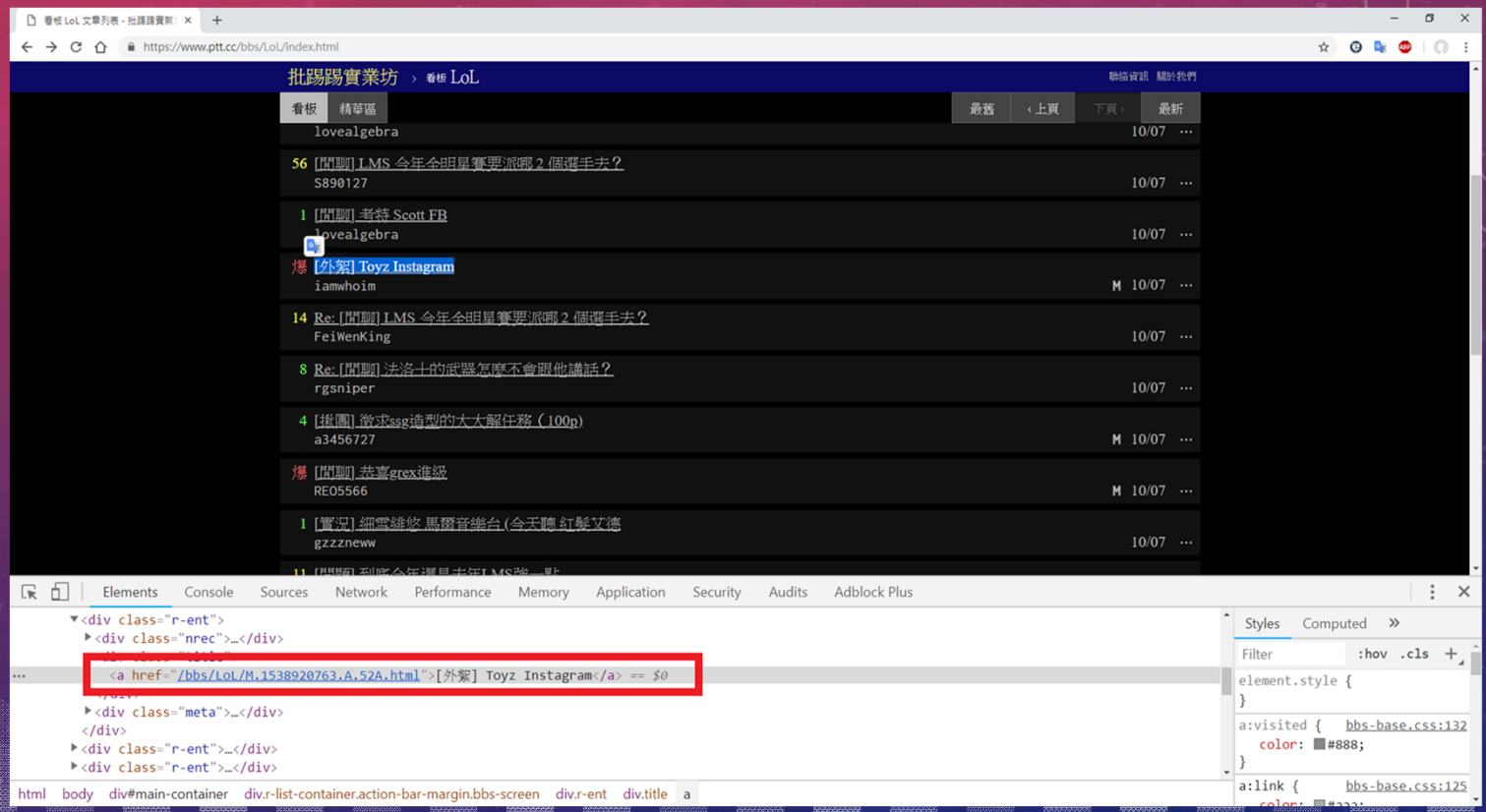
事前作業

- 開始爬蟲前，我們必須先安裝一些Python套件
- $ pip install requests
- $ pip install bs4

Get hrefs


https://www.ptt.cc/bbs/LoL/index.html
批踢踢實業坊 › 看板 LoL
聯絡資訊 關於我們
看板 精華區
最舊 ‹ 上頁 下頁 › 最新
搜尋文章…
11 [閒聊] 法洛士的武器怎麼不會跟他講話？
lycs0908 10/07
5 [閒聊] 今年跟S4，哪一年的台灣比較強？
wyner 10/07
5 [閒聊] S6和今年的MMD哪個比較難受？
bygamantou 10/07
24 [閒聊] GBM 推特
HaiTurtle 10/07
14 [閒聊] LMS
lovealgebra 10/07
56 [閒聊] LMS
S890127 10/07
1 [閒聊] 考特
lovealgebra 10/07
爆 [外絮] Toyz Instagram
iamwhoim M 10/07
14 Re: [閒聊] LMS 今年全明星賽要派哪2 個選手去？
FeiWenKing 10/07
在新分頁中開啟連結(T)
在新視窗中開啟連結(W)
在無痕式視窗中開啟連結(G)
另存連結為(K)...
複製連結網址(E)
IE Tab Options
檢查(N) Ctrl+Shift+I
https://www.ptt.cc/bbs/LoL/M.1538920763.A.52A.html
跟他講話？

Get hrefs

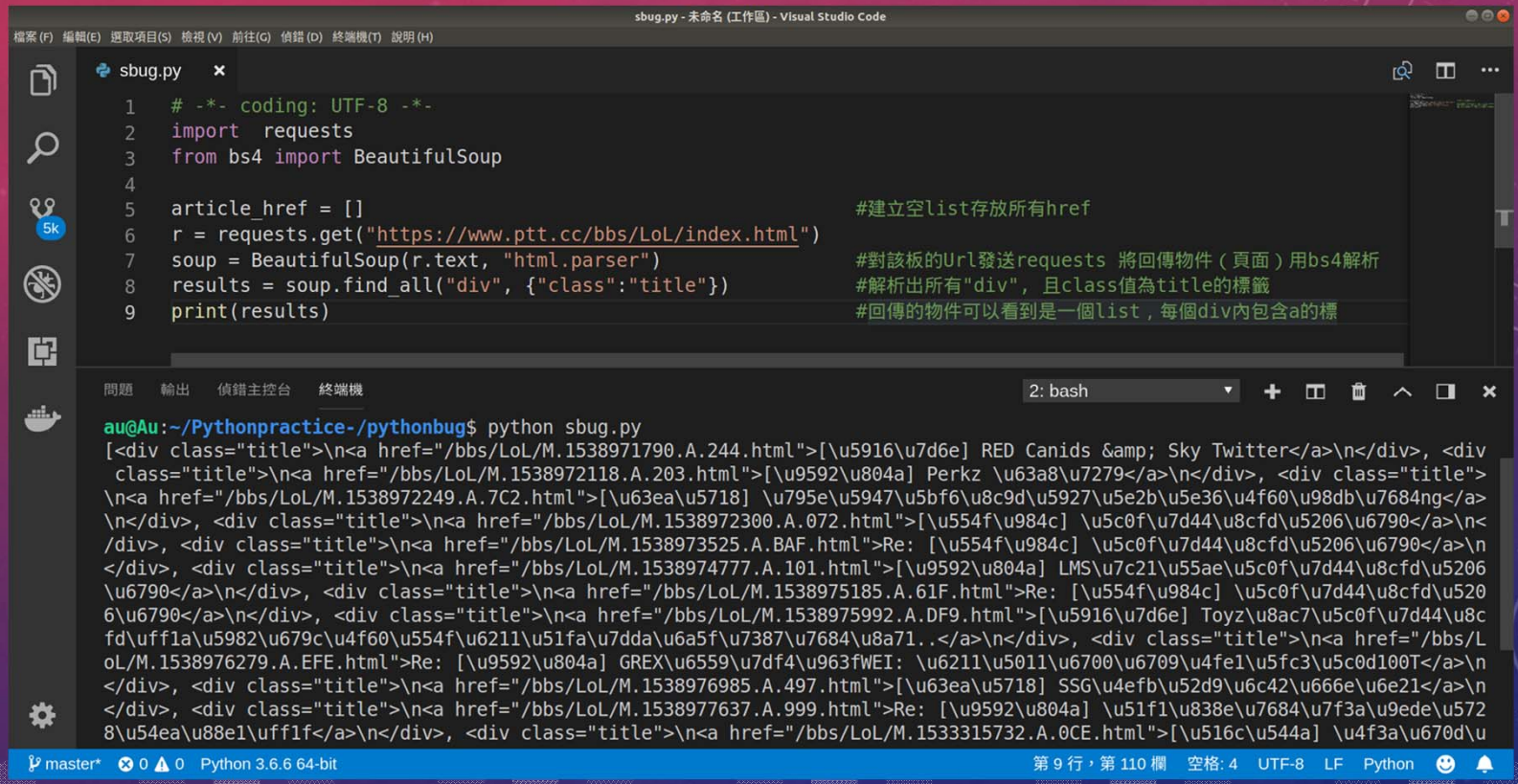

https://www.ptt.cc/bbs/LoL/index.html
批踢踢實業坊 › 看板 LoL
看板 精華區
最舊 ‹ 上頁 下頁 › 最新
56 [閒聊] LMS 今年全明星賽要派哪 2 個選手去？
S890127 10/07
1 [閒聊] 考特 Scott FB
lovealgebra 10/07
爆 [外絮] Toyz Instagram
iamwhoim M 10/07
14 Re: [閒聊] LMS 今年全明星賽要派哪 2 個選手去？
FeiWenKing 10/07
8 Re: [閒聊] 法洛士的武器怎麼不會跟他講話？
rgsniper 10/07
4 [推圖] 徵求ssg造型的太太解任務（100p）
a3456727 M 10/07
爆 [閒聊] 恭喜grex進級
REO5566 M 10/07
1 [實況] 細雪緋悠 馬爾音樂台 (今天聽 紅髮艾德
gzzzneww 10/07
Elements Console Sources Network Performance Memory Application Security Audits Adblock Plus

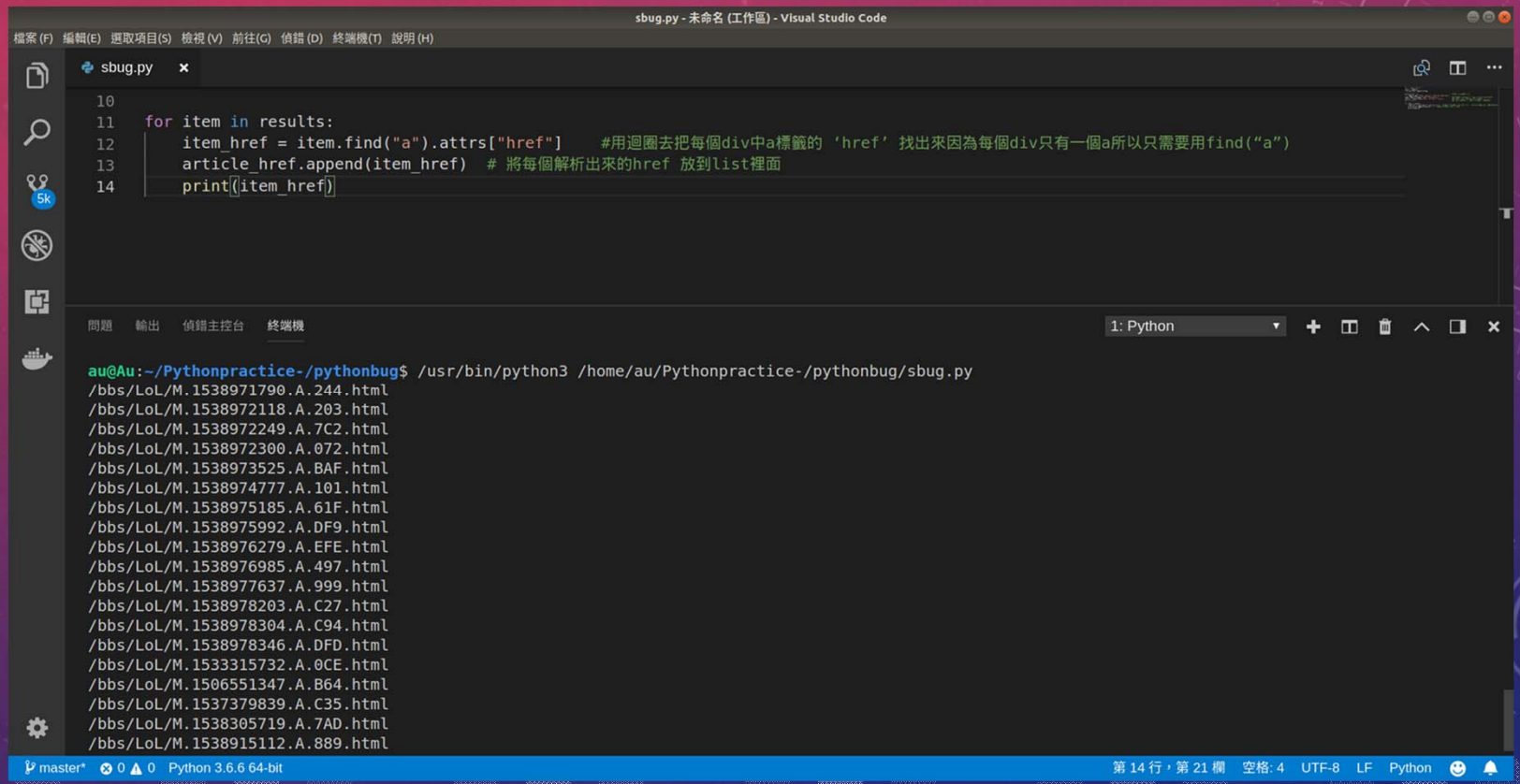
…
[外絮] Toyz Instagram == $0
…

…
…
Styles Computed
Filter :hov .cls
element.style {
}
a:visited { bbs-base.css:132
color: #888;
}
a:link { bbs-base.css:125
html body div#main-container div.r-list-container.action-bar-margin.bbs-screen div.r-ent div.title a

Get hrefs

點進來可以發現每篇文章的網址跟剛剛選取的Href是一樣的


sbug.py - 未命名 (工作區) - Visual Studio Code
檔案 (F) 編輯(E) 選取項目(S) 檢視 (V) 前往(G) 偵錯 (D) 終端機(T) 說明 (H)
sbug.py
1 # -*- coding: UTF-8 -*-
2 import requests
3 from bs4 import BeautifulSoup
4
5 article_href = [] #建立空list存放所有href
6 r = requests.get("https://www.ptt.cc/bbs/LoL/index.html")
7 soup = BeautifulSoup(r.text, "html.parser") #對該板的Url發送requests 將回傳物件（頁面）用bs4解析
8 results = soup.find_all("div", {"class":"title"}) #解析出所有"div", 且class值為title的標籤
9 print(results) #回傳的物件可以看到是一個list，每個div內包含a的標
問題 輸出 偵錯主控台 終端機
2: bash
au@Au:~/Pythonpractice-/pythonbug$ python sbug.py
[\n[\u5916\u7d6e] RED Canids & Sky Twitter\n , \n[\u9592\u804a] Perkz \u63a8\u7279\n , \n[\u63ea\u5718] \u795e\u5947\u5bf6\u8c9d\u5927\u5e2b\u5e36\u4f60\u98db\u7684ng\n , \n[\u554f\u984c] \u5c0f\u7d44\u8cfd\u5206\u6790\n , \nRe: [\u554f\u984c] \u5c0f\u7d44\u8cfd\u5206\u6790\n , \n[\u9592\u804a] LMS\u7c21\u55ae\u5c0f\u7d44\u8cfd\u5206\u6790\n , \nRe: [\u554f\u984c] \u5c0f\u7d44\u8cfd\u5206\u6790\n , \n[\u5916\u7d6e] Toyz\u8ac7\u5c0f\u7d44\u8cfd\uff1a\u5982\u679c\u4f60\u554f\u6211\u51fa\u7dda\u6a5f\u7387\u7684\u8a71..\n , \nRe: [\u9592\u804a] GREX\u6559\u7df4\u963fWEI: \u6211\u5011\u6700\u6709\u4fe1\u5fc3\u5c0d100T\n , \n[\u63ea\u5718] SSG\u4efb\u52d9\u6c42\u666e\u6e21\n , \nRe: [\u9592\u804a] \u51f1\u838e\u7684\u7f3a\u9ede\u5728\u54ea\u88e1\uff1f\n , \n<a href="/bbs/LoL/M.1533315732.A.0CE.html">[\u516c\u544a] \u4f3a\u670d\u
master* 0 0 Python 3.6.6 64-bit
第 9 行，第 110 欄 空格: 4 UTF-8 LF Python


sbug.py - 未命名 (工作區) - Visual Studio Code
檔案 (F) 編輯(E) 選取項目(S) 檢視 (V) 前往(G) 偵錯 (D) 終端機(T) 說明 (H)
sbug.py
10
11 for item in results:
12 item_href = item.find("a").attrs["href"] #用迴圈去把每個div中a標籤的 'href' 找出來因為每個div只有一個a所以只需要用find("a")
13 article_href.append(item_href) # 將每個解析出來的href 放到list裡面
14 print(item_href)
問題 輸出 偵錯主控台 終端機
1: Python
au@Au:~/Pythonpractice-/pythonbug$ /usr/bin/python3 /home/au/Pythonpractice-/pythonbug/sbug.py
/bbs/LoL/M.1538971790.A.244.html
/bbs/LoL/M.1538972118.A.203.html
/bbs/LoL/M.1538972249.A.7C2.html
/bbs/LoL/M.1538972300.A.072.html
/bbs/LoL/M.1538973525.A.BAF.html
/bbs/LoL/M.1538974777.A.101.html
/bbs/LoL/M.1538975185.A.61F.html
/bbs/LoL/M.1538975992.A.DF9.html
/bbs/LoL/M.1538976279.A.EFE.html
/bbs/LoL/M.1538976985.A.497.html
/bbs/LoL/M.1538977637.A.999.html
/bbs/LoL/M.1538978203.A.C27.html
/bbs/LoL/M.1538978304.A.C94.html
/bbs/LoL/M.1538978346.A.DFD.html
/bbs/LoL/M.1533315732.A.0CE.html
/bbs/LoL/M.1506551347.A.B64.html
/bbs/LoL/M.1537379839.A.C35.html
/bbs/LoL/M.1538305719.A.7AD.html
/bbs/LoL/M.1538915112.A.889.html
master* 0 0 Python 3.6.6 64-bit
第 14 行，第 21 欄 空格: 4 UTF-8 LF Python

Parser Article

可以看到「作者」、「看板」、「標題」、「時間」資料
都放在

```
▼<div class="article-metaline">
    <span class="article-meta-tag">作者</span>
    <span class="article-meta-value">coolplus (cool)</span>
  </div>
▼<div class="article-metaline-right">
    <span class="article-meta-tag">看板</span>
    <span class="article-meta-value">LoL</span>
  </div>
▼<div class="article-metaline">
    <span class="article-meta-tag">標題</span>
    <span class="article-meta-value">[閒聊] GREX教練阿WEI: 我們最有信心對100T</span>
  </div>
▼<div class="article-metaline">
    <span class="article-meta-tag">時間</span>
    <span class="article-meta-value">Mon Oct  8 11:27:35 2018</span>
```

Parser Article

```
author = soup.select('span.article-meta-value')[0].text #作者
board = soup.select('span.article-meta-value')[1].text  #看板
title = soup.select('span.article-meta-value')[2].text  #標題
time = soup.select('span.article-meta-value')[3].text   #時間
push_tag = soup.select('span.push-tag')                 #推文
push_userid = soup.select('span.push-userid')           #推文id
push_content = soup.select('span.push-content')         #推文內容
push_ipdatetime = soup.select('span.push-ipdatetime')   #推文時間
print('作者:', author)
print(board,' 看版')
print('標題:', title)
print('時間:', time)
push_list_len = len(push_tag)                           #計算推文筆數
count = 0
while (count < push_list_len):
    print (push_tag[count].text + push_userid[count].text + push_content[count].text + push_ipdatetime[count].text)
    count = count+1
print("======================分隔線=======================")
```

Parser Article

處理內文的資料

```
content = soup.find(id="main-content").text                #content 文章內文
target_content = u'※ 發信站: 批踢踢實業坊(ptt.cc),'
content = content.split(target_content)                    #去除掉 target_content
content = content[0].split(time)
main_content = content[1].replace('--','  ')               #去除掉文末 --
```

Output

將剛才所整理好的所有資料用print打印出來

```
print('作者:', author)
print(board,' 版')
print('標題:', title)
print('時間:', time)
print('內文:', main_content)

push_list_len = len(push_tag)                          #計算推文筆數
count = 0

while (count < push_list_len):                         #利用迴圈印出所有推文
    print (push_tag[count].text + push_userid[count].text + push_content[count].text + push_ipdatetime[count].text)
    count = count+1
```

輸出結果

```
=======================分隔線=======================
作者: airiguodala
LoL  版
標題: [公告] LoL 板 開始舉辦樂透!
時間: Sun Oct  7 20:25:11 2018
內文:
特!
請到 LoL 板 按'f'參與樂透!

16:00 KT vs TL 17:00 EDG vs MAD 猜TL跟MAD誰撐的比較久(比較慢輸)
1.TL
2.MAD
3.TL/MAD贏       開獎順序:3>2=1


因為間隔都很短 又不像LMS這麼會修 開一次混合的試試看

如果下注狀況不好之後就不開惹


一張 100 Ptt幣 (平民級)
樂透結束時間: 10/10/2018 16:05:17 Wed
※編輯: airiguodala (114.36.7.34), 10/07/2018 20:26:08
推 ilove640    : 特!  10/07 20:25
→ ray221740718: 特!  10/07 20:25
→ asd860079   : 特! 10/07 20:25
推 dinosaur8484: 特! ! ! ! ! ! ! ! ! ! !  10/07 20:25
推 soalzelance : 特!  10/07 20:26
推 Ball0427    : 特!  10/07 20:26
推 hkr91511208 : 3 10/07 20:26
推 Znps        : 特! 10/07 20:26
推 marginalFeng: https://i.imgur.com/RANbro2.jpg 10/07 20:26
推 tw15        : 三是什麼啊 10/07 20:26
```