


PTT 網路爬蟲教學

吳智鴻

國立臺中教育大學 數位內容科技學系

2019/12/18



建立虛擬環境

建立一個虛擬環境，以便安裝合適的套件

Step#1 建立虛擬環境 & 指定python=3.5版本（相容性較高）

```
conda create -n py35 python=3.5 jupyter numpy matplotlib bs4
```

Step#2 啟動虛擬環境

```
source activate py35
```

Step#3 退出虛擬環境

```
deactivate py35
```

需要套件

(1) requests套件(抓取網頁資料用，anaconda 預設會安裝)

安裝方式

```
pip install requests
```

(2) BeautifulSoup (解析網頁用)

安裝方式

```
pip install bs4
```

分析ptt的看板網址（可以設定自己有興趣的版）

ptt.cc/bbs/**movie**/index.html

看板名稱

批踢踢實業坊 > 看板 **movie** 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...

[評論] 聞天祥談 / 茱蒂嘉蘭 MyA11	12/23 ...
2 [影評] 聞天祥評 / 燃燒女子的畫像 MyA11	12/23 ...

分析文章標題 (按Ctrl+U)

看板	精華區	最舊	< 上頁	下頁 >	最新
搜尋文章...					
	[評論] 聞天祥談 / 茱蒂嘉蘭 MyAll	12/23	...		
2	[影評] 聞天祥評 / 燃燒女子的畫像 MyAll	12/23	...		
1	[請益] 找片 <山田孝之的痛苦與榮耀> aiweisen	12/23	...		
11	[新聞] 小島秀夫選出 2019 他最愛的 5 部電影 hahaha0204	12/23	...		

```
119 <div class="r-ent">
120 <div class="nrec"><span class="hl f2">1</span></div>
121 <div class="title">
122
123 <a href="/bbs/movie/M.1577104167.A.CC3.html">[請益] 找片 &lt;山田孝之的痛苦與榮耀&gt;</a>
124
125 </div>
126 <div class="meta">
127 <div class="author">aiweisen</div>
128 <div class="article-menu">
129
130 <div class="trigger">&#x22ef;</div>
131 <div class="dropdown">
132 <div class="item"><a href="/bbs/movie/search?
q=thread%3A5B%E8%AB%8B%E7%9B%8A%5D&#43;%E6%89%BE%E7%89%87&#43;%3C%E5%B1%B1%E7%94%B0%E5%AD%9D%E4%B9%8B%E7%9A%84%E7%97%9B%E8%8B%A6%E8%
88%87%E6%A6%AE%E8%80%80%3E">搜尋同標題文章</a></div>
```

分析HTML結構 (這個步驟是關鍵)

```
<div class= "title" >
```

文章標題 被放在 href 裡面

.....

```
</div>
```

分析結果


原始網頁內容

```
119     <div class="r-ent">
120         <div class="nrec"><span class="hl f2">1</span></div>
121         <div class="title">
122             <a href="/bbs/movie/M.1577104167.A.CC3.html">[請益] 找片 &lt;山田孝之的痛苦與榮耀&gt;</a>
123         </div>
124     </div>
125     <div class="meta">
126         <div class="author">aiweisen</div>
127         <div class="article-menu">
128             <div class="trigger">&#x22ef;</div>
129             <div class="dropdown">
130                 <div class="item"><a href="/bbs/movie/search?
131 q=thread%3A%5B%E8%AB%8B%E7%9B%8A%5D%#43;%E6%89%BE%E7%89%87%#43;%3C%E5%B1%B1%E7%94%B0%E5%AD%9D%E4%B9%8B%E7%9A%84%E7%97%9B%E8%8B%A6%E8%
132 88%87%E6%A6%AE%E8%80%80%3E">搜尋同標題文章</a></div>
```

文章內容 (在href=XXX)

文章內文被放在 `` 裡面

```
119 <div class="r-ent">
120   <div class="nrec"><span class="hl f2">1</span></div>
121   <div class="title">
122     <a href="/bbs/movie/M.1577104167.A.CC3.html">[請益] 找片 &lt;山田孝之的痛苦與榮耀&gt;</a>
123   </div>
124   <div class="meta">
125     <div class="author">aiweisen</div>
126     <div class="article-menu">
127       <div class="trigger">&#x22ef;</div>
128       <div class="dropdown">
129         <div class="item"><a href="/bbs/movie/search?
130 q=thread%3A%5B%E8%AB%8B%E7%9B%8A%5D&#43;%E6%89%BE%E7%89%87&#43;%3C%E5%B1%B1%E7%94%B0%E5%AD%9D%E4%B9%8B%E7%9A%84%E7%97%9B%E8%8B%A6%E8%
131 88%87%E6%A6%AE%E8%80%80%3E">搜尋同標題文章</a></div>
```



虛擬環境操作指令介紹 (Anaconda)

Step #1 建立虛擬環境 & 指定python版本

```
conda create -n py35 python==3.5 jupyter
```

```
conda create -n 虛擬環境名稱 python==版本
```

Step #2 啟動虛擬環境

```
conda source activate py35
```

Step #3 退出虛擬環境

```
conda deactivate py35
```

Step #4 刪除虛擬環境 (避免佔用硬碟空間)

```
conda remove -n py35
```


Prg: ptt_movie#1

```
# ptt_movie 電影版爬蟲
import requests
from bs4 import BeautifulSoup
```

```
article_href = []
r = requests.get("https://www.ptt.cc/bbs/movie/index.html") #指定要抓取的版網址
```

```
soup = BeautifulSoup(r.text, "html.parser")
results = soup.select("div.title")
print(results)
```

指定抓取
div.title部分

Results變數是一個list，
裡面把該頁的div class="title"元素都取出來且
裡面包覆著<a>標籤

Results
結果

```
<div class="title">
<a href="/bbs/movie/M.1577103729.A.B44.html">[評論] 聞天祥談 / 茱蒂嘉蘭</a>
</div>, <div class="title">
<a href="/bbs/movie/M.1577104082.A.F6B.html">[影評] 聞天祥評 / 燃燒女子的畫像</a>
</div>, <div class="title">
<a href="/bbs/movie/M.1577104167.A.CC3.html">[請益] 找片 &lt;山田孝之的痛苦與榮耀&gt;</a>
</div>, <div class="title">
<a href="/bbs/movie/M.1577104449.A.D5A.html">[新聞] 小島秀夫選出 2019 他最愛的 5 部電影</a>
</div>, <div class="title">
<a href="/bbs/movie/M.1577104844.A.846.html">[討論] Myvideo 2019電影排行榜 Top1無限之戰(雷</a>
</div>, <div class="title">
<a href="/bbs/movie/M.1577105464.A.286.html">[新聞] 他這樣看電影創作 張善政：不要摻入意識</a>
</div>, <div class="title">
<a href="/bbs/movie/M.1577105579.A.8A9.html">[討論] 超粒方：為什麼你該看星際大戰? </a>
</div>, <div class="title">
<a href="/bbs/movie/M.1577108008.A.45C.html">[討論] 魔鬼終結者4未來救贖 如果這樣拍</a>
</div>, <div class="title">
<a href="/bbs/movie/M.1577108916.A.654.html">[討論] 不討喜的米老鼠版星際大戰主要角色個性...</a>
</div>, <div class="title">
<a href="/bbs/movie/M.1559611458.A.DCA.html">[公告] 板規 2019/08/24</a>
</div>, <div class="title">
<a href="/bbs/movie/M.1565855832.A.0A7.html">[公告] 板規新增每日發文上限規定</a>
</div>, <div class="title">
<a href="/bbs/movie/M.1574587497.A.388.html">Fw: [公告] 請使用安全的連線方式連線本站</a>
</div>, <div class="title">
<a href="/bbs/movie/M.1576565795.A.547.html">[公告] 獎季發文限制放寬</a>
</div>, <div class="title">
<a href="/bbs/movie/M.1577083238.A.F44.html">[公告] 請注意發文格式 標題不要爆雷</a>
</div>
```

把文章連結印出來 取出該頁所有的連結

加上把results的list印出來的程式

```
# 把list印出來
for item in results:
    item_href = item.select_one("a").get("href") # 取出 a href得料
    article_href.append(item_href)
print(article_href)
```

```
['/bbs/movie/M.1577103729.A.B44.html', '/bbs/movie/M.1577104082.A.F6B.html', '/bbs/movie/M.1577104167.A.CC3.html', '/bbs/movie/M.1577104449.A.D5A.html', '/bbs/movie/M.1577104844.A.846.html', '/bbs/movie/M.1577105464.A.286.html', '/bbs/movie/M.1577105579.A.8A9.html', '/bbs/movie/M.1577108008.A.45C.html', '/bbs/movie/M.1577108916.A.654.html', '/bbs/movie/M.1559611458.A.DCA.html', '/bbs/movie/M.1565855832.A.0A7.html', '/bbs/movie/M.1574587497.A.388.html', '/bbs/movie/M.1576565795.A.547.html', '/bbs/movie/M.1577083238.A.F44.html']
```

取出上下一頁資料

上頁

在上頁處
按下右鍵
選檢查

取得div內class為btn-group下的a
標籤
回傳的結果可以看到要的「上」
在第3個Index

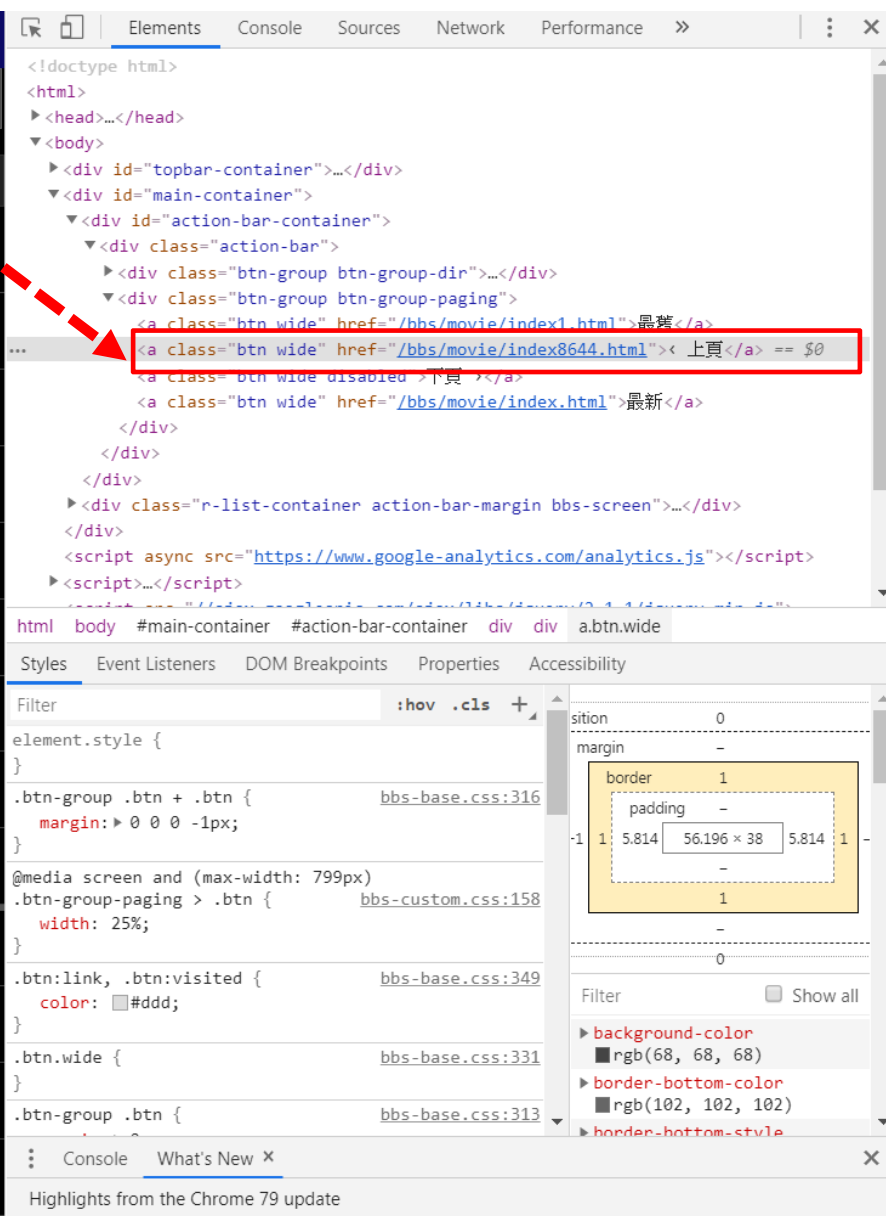


批踢踢實業坊 > 看板 movie 聯絡資訊 關於我們

看板 精華區 最舊 < 上頁 下頁 > 最新

搜尋文章...

[評論] 聞天祥談 / 茱蒂嘉蘭	MyAll	12/23	...
2 [影評] 聞天祥評 / 燃燒女子的畫像	MyAll	12/23	...
1 [請益] 找片 <山田孝之的痛苦與榮耀>	aiweisen	12/23	...
11 [新聞] 小島秀夫選出 2019 他最愛的 5 部電影	hahaha0204	12/23	...
1 [討論] Myvideo 2019電影排行榜 Top1無限之戰(雷	Alyssachao	12/23	...
X1 [新聞] 他這樣看電影創作 張善政：不要摻入意識	sister4949	12/23	...
[討論] 超粒方:為什麼你該看星際大戰?	kem0606	12/23	...
3 [討論] 魔鬼終結者4未來救贖 如果這樣拍	bth060104	12/23	...
[討論] 不討喜的米老鼠版星際大戰主要角色個性..	AisinGiuro	12/23	...
26 [公告] 板規 2019/08/24	ckshchen	6/04	...
3 [公告] 板規新增每日發文上限規定	hhwang	M 8/15	...
1 Fw: [公告] 請使用安全的連線方式連線本站	kai3368	11/24	...
[公告] 獎季發文限制放寬	hhwang	12/17	...



Elements Console Sources Network Performance

```
<!doctype html>
<html>
<head>...</head>
<body>
  <div id="topbar-container">...</div>
  <div id="main-container">
    <div id="action-bar-container">
      <div class="action-bar">
        <div class="btn-group btn-group-dir">...</div>
        <div class="btn-group btn-group-paging">
          <a class="btn wide" href="/bbs/movie/index1.html">最舊</a>
          <a class="btn wide" href="/bbs/movie/index8644.html">< 上頁</a> == $0
          <a class="btn wide disabled">下頁</a>
          <a class="btn wide" href="/bbs/movie/index.html">最新</a>
        </div>
      </div>
    </div>
  </div>
  <div class="r-list-container action-bar-margin bbs-screen">...</div>
</div>
<script async src="https://www.google-analytics.com/analytics.js"></script>
<script>...</script>
```

html body #main-container #action-bar-container div div a.btn.wide

Styles Event Listeners DOM Breakpoints Properties Accessibility

Filter :hov .cls +

```
element.style {
}
.btn-group .btn + .btn {
  margin: 0 0 0 -1px;
}
@media screen and (max-width: 799px)
.btn-group-paging > .btn {
  width: 25%;
}
.btn:link, .btn:visited {
  color: #ddd;
}
.btn.wide {
}
.btn-group .btn {
```

position 0
margin -
border 1
padding -
1 5.814 56.196 x 38 5.814 1
background-color rgb(68, 68, 68)
border-bottom-color rgb(102, 102, 102)
border-bottom-style

Console What's New x

Highlights from the Chrome 79 update

Prg: ptt_movie#2

抓取n頁的連結

```
# ptt_movie #2 電影版爬蟲  
# 取出上下一頁資料
```

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

```
# 抓取 n 頁資料  
n = 10  
url="https://www.ptt.cc/bbs/movie/index.html"  
for page in range(1,n+1):  
    r = requests.get(url)  
    soup = BeautifulSoup(r.text,"html.parser")  
    btn = soup.select('div.btn-group > a')  
    up_page_href = btn[3]['href']  
    next_page_url = 'https://www.ptt.cc' + up_page_href  
    url = next_page_url  
    print(url)
```

完整程式碼

Prg: ptt_movie#3 完整程式

抓取n頁的連結，並加上抓取每頁的標題

```
# prg#3
# ptt_movie #3 電影版爬蟲
# 取出上下一頁資料
# 完整程式版本 加入get_all_href
```

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

```
url="https://www.ptt.cc/bbs/movie/index.html" #ptt看板網址
n = 2 #需要抓取幾頁資料
```

```
# 抓取該頁所有的標題
def get_all_href(url):
    r = requests.get(url)
    soup = BeautifulSoup(r.text, "html.parser")
    results = soup.select("div.title")
    for item in results:
        a_item = item.select_one("a")
        title = item.text
        if a_item:
            print(title, 'https://www.ptt.cc'+ a_item.get('href'))
```

```
# 抓取前n頁所有的連結
for page in range(1, n+1):
    r = requests.get(url)
    soup = BeautifulSoup(r.text, "html.parser")
    btn = soup.select('div.btn-group > a')
    up_page_href = btn[3]['href']
    next_page_url = 'https://www.ptt.cc' + up_page_href
    url = next_page_url
    get_all_href(url = url)
```

Re: [好雷] 葉問4 弘揚國術發揚光大

<https://www.ptt.cc/bbs/movie/M.1577084192.A.C41.html>

[好雷] 葉問以收尾作來說 算是收的不錯

<https://www.ptt.cc/bbs/movie/M.1577085551.A.682.html>

[雷] 《貓》:年度最佳恐怖/邪教電影候選人

<https://www.ptt.cc/bbs/movie/M.1577087346.A.3FA.html>

[新聞] 楊德昌《一一》當年台灣未上映 Netflix

<https://www.ptt.cc/bbs/movie/M.1577089971.A.D87.html>

[討論] 魔戒有辦法重啟嗎

<https://www.ptt.cc/bbs/movie/M.1577090068.A.1DA.html>

Re: [新聞] 原力助攻 天行者的崛起輕鬆登北美票房龍頭

<https://www.ptt.cc/bbs/movie/M.1577090447.A.FBA.html>

[討論] 簡單一句話介紹昆崙大獸

Ptt_movie4

抓取文章內容

完整程式碼

Prg: ptt_movie#4 完整程式

抓取n頁的連結，並加上抓取每頁的內文

加上 get_article_content 抓取文章內文

```
def get_article_content(article_url):  
    r = requests.get(article_url)  
    soup = BeautifulSoup(r.text, "lxml")  
    results = soup.select('span.article-meta-value')  
    if results:  
        print('作者:', results[0].text)  
        print('看板:', results[1].text)  
        print('標題:', results[2].text)  
        print('時間:', results[3].text)
```

內文放在span class= "article-meta-value" >裡面

```
55 <div id="main-container">  
56     <div id="main-content" class="bbs-screen bbs-content"><div class="article-metaline"><span class="art  
value">a2337548 (santa91)</span></div><div class="article-metaline-right"><span class="article-meta-tag"  
</div><div class="article-metaline"><span class="article-meta-tag">標題</span><span class="article-meta-v  
class="article-metaline"><span class="article-meta-tag">時間</span><span class="article-meta-value">Mon C  
57  
58  
59 趁著重新上映，昨天去看了海上花  
60 雖然一直長鏡頭很過癮，但有時候會突然疲乏  
61 想請問劇情中的問題  
62
```

程式碼

```
# prg#4
# ptt_movie #4 電影版爬蟲
# 取出上下一頁資料
# 完整程式版本 加入get_all_href 顯示下一頁 抓取文章內容
```

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

```
url="https://www.ptt.cc/bbs/movie/index.html" #ptt看板網址
n = 2 #需要抓取幾頁資料
```

```
def get_article_content(article_url):
    r = requests.get(article_url)
    soup = BeautifulSoup(r.text, "html.parser")
    results = soup.select('span.article-meta-value')
    if results:
        print('作者：', results[0].text)
        print('看板：', results[1].text)
        print('標題：', results[2].text)
        print('時間：', results[3].text)
        print('-----')
```

```
# 抓取該頁所有的標題
```

```
def get_all_href(url):
    r = requests.get(url)
    soup = BeautifulSoup(r.text, "html.parser")
    results = soup.select("div.title")
    for item in results:
        a_item = item.select_one("a")
        title = item.text
        if a_item:
            #原先程式
            #print(title, 'https://www.ptt.cc'+ a_item.get('href'))
            #改成呼叫 get_article_content去抓取內容
            get_article_content(article_url='https://www.ptt.cc'+ a_item.get('href'))
    print('----- Next Page -----')
```

```
# 抓取前n頁所有的連結
```

```
for page in range(1, n+1):
    r = requests.get(url)
    soup = BeautifulSoup(r.text, "html.parser")
    btn = soup.select('div.btn-group > a')
    up_page_href = btn[3]['href']
    next_page_url = 'https://www.ptt.cc' + up_page_href
    url = next_page_url
    get_all_href(url = url)
```

作者： dakkk (我是牛我反芻)
看板： movie
標題： Re: [好雷] 葉問4 弘揚國術發揚光大
時間： Mon Dec 23 14:56:30 2019

作者： purue (purue)
看板： movie
標題： [好雷] 葉問以收尾作來說 算是收的不錯
時間： Mon Dec 23 15:19:09 2019

作者： sunny1991225 (桑妮)
看板： movie
標題： [雷] 《貓》:年度最佳恐怖/邪教電影候選人
時間： Mon Dec 23 15:49:04 2019

作者： CYKONGG (CYKONGG)
看板： movie
標題： [新聞] 楊德昌《一一》當年台灣未上映 Netflix
時間： Mon Dec 23 16:32:49 2019

Ptt_movie5

抓取每個文章中留言內容 完整程式碼

留待練習

分析網頁結構

上網搜尋

參考文獻：

<http://wcck2017.blogspot.com/2017/06/crawler-ptt-movie.html>

結果如下

作者： dakkk (我是牛我反芻)

看板： movie

標題： Re: [好雷] 葉問4 弘揚國術發揚光大

時間： Mon Dec 23 14:56:30 2019

messages = 就是啟蒙概念吧

messages = 李小龍就是看不過傳武太迂，才創截拳道

messages = 他把詠春發揮在寸拳上了吧

messages = 裡面劇情幾乎都是編的

messages = 一大堆葉問作品只有大飛哥那部最接近歷史

messages = 笑死除了葉問每個師傅上台撐不了10，秒

messages = 截拳道和詠春的關係已經很低了。

messages = 提他太小，不然羅師傅就幹掉了

解答

prg5

```
# prg#5
# ptt_movie #5 電影版爬蟲
# 取出上下一頁資料
# 完整程式版本 加入get_all_href 顯示下一頁 抓取文章內容 抓取內文

import requests
from bs4 import BeautifulSoup
```

```
作者： dakkk (我是牛我反芻)
看板： movie
標題： Re: [好雷] 葉問4 弘揚國術發揚光大
時間： Mon Dec 23 14:56:30 2019
```

```
-----
messages = 就是啟蒙概念吧
```

```
-----
messages = 李小龍就是看不過傳武太迂，才創截拳道
```

```
-----
messages = 他把詠春發揮在寸拳上了吧
```

```
-----
messages = 裡面劇情幾乎都是編的
```

```
-----
messages = 一大堆葉問作品只有大飛哥那部最接近歷史
```

```
-----
messages = 笑死除了葉問每個師傅上台撐不了10，秒
```

```
-----
messages = 截拳道和詠春的關係已經很低了。
```

```
-----
messages = 場地太小 不然羅師傅就該撐很久了
```

```
url="https://www.ptt.cc/bbs/movie/index.html" #ptt看板網址
n = 2 #需要抓取幾頁資料

def get_article_content(article_url):
    r = requests.get(article_url)
    soup = BeautifulSoup(r.text, "html.parser")
    results = soup.select('span.article-meta-value')
    if results:
        print('作者：', results[0].text)
        print('看板：', results[1].text)
        print('標題：', results[2].text)
        print('時間：', results[3].text)
        print('-----')

def get_article_content_inside(article_url):
    r = requests.get(article_url)
    soup = BeautifulSoup(r.text, "html.parser")
    articles = soup.find_all('div', 'push')
    for article in articles:
        # 去掉冒號和左右的空白
        messages = article.find('span', 'f3 push-content').getText().replace(':', '').strip()
        print('messages = ', messages)
        print('-----')

# 抓取該頁所有的標題
def get_all_href(url):
    r = requests.get(url)
    soup = BeautifulSoup(r.text, "html.parser")
    results = soup.select("div.title")
    for item in results:
        a_item = item.select_one("a")
        title = item.text
        if a_item:
            # 原先程式
            # print(title, 'https://www.ptt.cc'+ a_item.get('href'))
            # 改成呼叫 get_article_content 去抓取內容
            get_article_content(article_url='https://www.ptt.cc'+ a_item.get('href'))
            get_article_content_inside(article_url='https://www.ptt.cc'+ a_item.get('href'))
    print('----- Next Page -----')

# 抓取前頁所有的連結
for page in range(1, n+1):
    r = requests.get(url)
    soup = BeautifulSoup(r.text, "html.parser")
    btn = soup.select('div.btn-group > a')
    up_page_href = btn[3]['href']
    next_page_url = 'https://www.ptt.cc' + up_page_href
    url = next_page_url
    get_all_href(url = url)
```

參考來源

- [1] <https://ithelp.ithome.com.tw/articles/10204709>
 - [2] <https://ithelp.ithome.com.tw/articles/10205022>
 - [3] <http://wcck2017.blogspot.com/2017/06/crawler-ptt-movie.html>
- 