



Big Data

大數據理論及實務應用

Created by 孫善堂 【小孫學堂】



Requests Methods Get or Post?

Created by 孫善堂 【小孫學堂】

What is the BeautifulSoup ?

1.網頁爬蟲工具庫

2.自動轉換為utf-8編碼

相關連結

採購統計

行政法人
相關採購資訊

加值服務
訂閱

優先採購

「購買原住民及身心障礙者
所提供之產品或勞務」
採購專區平台

法人團體
帳號申請

安裝程式
環境檢測

檢索設定 ☐ 同音 ☐ 容錯

查詢

項次	種類	機關名稱	標案案號 標案名稱	招標公 告日期
1	招標公告	臺南市體育處	TNS109-05 109年臺南亞 太國際棒球 訓練中心少 棒園區委託 管理勞務採 購案	109/07/17
2	招標公告	國防部空軍司令 部	EI10002P003 投影系統維 護案	109/07/17
3	招標公告	財團法人國際合 作發展基金會	ICDF-109-024 110年度國際 人力資源培 訓研習班計 畫會議服務	109/07/17
4	招標公告	台灣電力股份有 限公司通霄發電 廠	3700900073 109年度#1機 氣渦輪機輔 機大修工作 FI1080601CM 花蓮縣消防 局特種排班	109/07/17

Elements Console Sources Network Performance >> | ⚙️ | ✕

⏏️ | 🔍 | ☐ Preserve log ☐ Disable cache | Online | ⬆️ ⬇️ ⬆️ | ⚙️

Filter ☐ Hide data URLs

All | XHR JS CSS Img Media Font Doc WS Manifest Other ☐ Has blocked cookies

☐ Blocked Requests

50000 ms 100000 ms 150000 ms 200000 ms 250000 ms

Name × Headers Preview Response Initiator Timing >>

prms-searchBulletinClient...

▼ General

Request URL: https://web.pcc.gov.tw/prkms/prms-searchBulletinClient.do?opt=tps

Request Method: POST

Status Code: 200 OK

Remote Address: 61.57.42.137:443

Referrer Policy: no-referrer-when-downgrade

▼ Response Headers view source

Connection: Keep-Alive

Content-Encoding: gzip

Content-Language: zh-TW

content-length: 15657

Content-Type: text/html; charset=UTF-8

1 / 50 requests | 16.1 kB / 18.5

標籤GET(明信片)



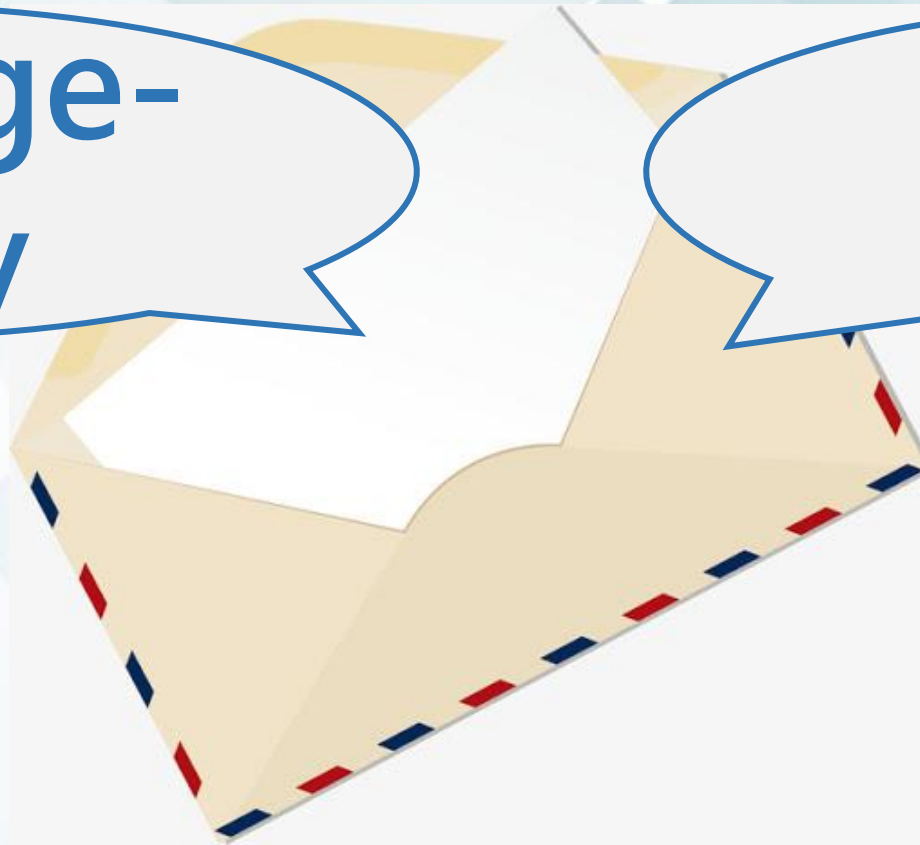
http-
header

使用 GET 的時候我們直接將要傳送的資料以 Query String (一種 Key/Value 的編碼方式) 加在我們要寄送的地址(URL)後面，然後交給郵差傳送。

POST(信封與信件)

message-
body

http-
header



使用 POST 的時候則是將寄送地址(URL)寫在信封上，另外將要傳送的資料寫在另一張信紙後，將信紙放到信封裡面，交給郵差傳送。

Form Data

檢索設定 ☐ 同音 ☐ 容錯

查詢

項次	種類	機關名稱	標案案號 標案名稱	招標公 告日期
1	招標公告	臺南市體育處	TNS109-05 109年臺南亞 太國際棒球 訓練中心少 棒園區委託 管理勞務採 購案	109/07/17
2	招標公告	國防部空軍司令 部	EI10002P003 投影系統維 護案	109/07/17
3	招標公告	財團法人國際合 作發展基金會	ICDF-109-024 110年度國際 人力資源培 訓研習班計 畫會議服務	109/07/17
4	招標公告	台灣電力股份有 限公司通霄發電 廠	3700900073 109年度#1機 氣渦輪機輔 機大修工作	109/07/17
			FI1080601CM 花蓮縣消防 局特種搜救	

Elements Console Sources Network Performance >> | ⚙️ ⋮ ✕

⬛ ⬛ 🔍 | ☐ Preserve log ☐ Disable cache | Online ⬇️ ⬆️ ⬇️ | ⚙️

Filter ☐ Hide data URLs

All XHR JS CSS Img Media Font Doc WS Manifest Other ☐ Has blocked cookies

☐ Blocked Requests

50000 ms 100000 ms 150000 ms 200000 ms 250000 ms

Name ✕ Headers Preview Response Initiator Timing >>

prms-searchBulletinClient... 103.116 Safari/537.36

▼ Query String Parameters view source view URL encoded

root: tps

▼ Form Data view source view URL encoded

tmpQuerySentence:

timeRange: 109/1/1-109/12/31

querySentence: 訓練

tenderStatusType: 招標

sortCol: TENDER_NOTICE_DATE

timeRangeTemp: 109/1/1-109/12/31

sym: on

itemPerPage: 10

1 / 50 requests | 16.1 kB / 18 B

使用POST抓取資料

Requests.post(URL,DATA)

```
url = 'https://web.pcc.gov.tw/prkms/prms-searchBulletinClient.do?root=tps'

data = {'tmpQuerySentence': None,
        'timeRange': '109/1/1-109/12/31',
        'querySentence': '訓練',
        'tenderStatusType': '招標',
        'sortCol': 'TENDER_NOTICE_DATE',
        'timeRangeTemp': '109/1/1-109/12/31',
        'sym': 'on',
        'itemPerPage': '100'}

res = requests.post(url,data)
soup = BeautifulSoup(res.text)|
```

DATA以dictionary形式讀入



網路爬蟲實作

政府電子採購招標清單

Created by 孫善堂 【小孫學堂】

requests method 分布

GET

政府電子採購網

English | 網站導覽 | 意見信箱 | 行動版

免費服務電話：0800-080512

首頁 > 全文檢索

全文檢索

招標查詢 | 決標查詢 | 全文檢索 | 公告日期查詢 | 機關名稱查詢 | 標的分類查詢 | 招標公告地圖查詢 | 財物出租查詢 | 財物變賣查詢 | 列印領標憑據

全文查詢 教育 ☐ 從結果中查詢

標案種類 ☒ 招標 ☐ 決標 ☐ 公開閱覽及公開徵求 ☐ 政府採購預告

排序欄位 招標公告日期

查詢範圍

<input checked="" type="radio"/> 109年1至12月	<input type="radio"/> 108年1至12月	<input type="radio"/> 107年1至12月	<input type="radio"/> 106年1至12月
<input type="radio"/> 105年1至12月	<input type="radio"/> 104年1至12月	<input type="radio"/> 103年1至12月	<input type="radio"/> 102年1至12月
<input type="radio"/> 101年1至12月	<input type="radio"/> 100年1至12月	<input type="radio"/> 99年1至12月	<input type="radio"/> 98年1至12月
<input type="radio"/> 97年1至12月	<input type="radio"/> 96年1至12月	<input type="radio"/> 95年1至12月	<input type="radio"/> 94年1至12月
<input type="radio"/> 93年1至12月	<input type="radio"/> 92年1至12月	<input type="radio"/> 91年1至12月	<input type="radio"/> 90年1至12月
<input type="radio"/> 89年1至12月	<input type="radio"/> 88年1至12月		

檢索設定 ☐ 同音 ☐ 容錯 每頁筆數 10

查詢

項次	種類	機關名稱	標案案號 標案名稱	招標公告日期	決標或無法 決標公告	截止投 標日期	公開閱覽/徵 求 起訖日期	預告公 告日期
1	招標公告	臺中市沙鹿區竹林國民小學	10921 竹林國小109學年度六年級戶外教育	109/09/28		109/10/05		
2	招標公告	財團法人原住民族文化事業基金會	109080 109年度原住民民族廣播電臺地方分臺節目製播暨主持-教育文化類家庭議題節目	109/09/28		109/10/06		
			109037					

POST

html階層觀察與物件抓取

知識管理

下載專區

相關連結

採購統計

行政法人
相關採購資訊

加值服務
訂閱

優先採購

「購買原住民及身心障礙者
所提供之產品或勞務」
採購專區平台

法人團體
帳號申請

安裝程式
環境檢測

查詢範圍

○ 101年1至12月

○ 97年1至12月

○ 93年1至12月

○ 89年1至12月

○ 100年1至12月

○ 96年1至12月

○ 92年1至12月

○ 88年1至12月

檢索設定

☐ 同音

☐ 容錯

查

項次	種類	機關名稱	標案案號 標案名稱	招標公 告日期
1	招標公告	臺南市體育處	TNS109-05 109年臺南亞 太國際棒球 訓練中心少 棒園區委託 管理勞務採 購案	109/07/...
2	招標公告	國防部空軍司令部	E110002P003 投影系統維 護案	109/07/...
3	招標公告	財團法人國際合作發展基金會	ICDF-109-024 110年度國際 人力資源培 訓研習班計 畫會議服務	109/07/...
		台灣電力股份有	3700900073 109年度#1機	

Elements Console Sources Network Performance >>

cellspacing="1" bgcolor="#FFFFFF">

<tbody>

<tr>

<th id="addThId_j" style="display: none;">

</th>

<td headers="addThId_j">

<table cellspacing="0" cellpadding="3" style="width:100%;background-color:#FFFFFF" id="searchResult">

<thead>...</thead>

<tbody>

<tr class="odd" style="background: rgb(255, 219, 166); color: red;">

<td style="text-align:center">

1

</td> == \$0

<td class="T12" style="width:10%;text-align:left;min-width:60px;">招標公告</td>

<td class="T12" style="width:18%;text-align:left">臺南市體育處</td>

><td class="T12" style="text-align:left">

tbody tr td table tbody tr td table tbody tr td #searchResult tbody tr td

Styles Event Listeners DOM Breakpoints Properties Accessibility

Filter :hov .cls +

element.style {
text-align: center;

border -

html階層觀察與物件抓取

```
'tenderStatusType': '招標',  
'sortCol': 'TENDER_NOTICE_DATE',  
'timeRangeTemp': '109/1/1-109/12/31',  
'sym': 'on',  
'itemPerPage': '100'}  
  
res = requests.post(url,data)  
soup = BeautifulSoup(res.text)  
  
print(soup.select('tbody')[0].select('tr')[0].select('td')[3].a.div.string)
```

109年臺南亞太國際棒球訓練中心少棒園區委託管理勞務採購案

以for迴圈針對各標案

```
for case in soup.select('tbody')[0].select('tr'):
    print(case.select('td')[3].a.div.string)|
```

109年臺南亞太國際棒球訓練中心少棒園區委託管理勞務採購案
投影系統維護案

110年度國際人力資源培訓研習班計畫會議服務

109年度#1機氣渦輪機輔機大修工作

花蓮縣消防局特種搜救隊訓練基地建置新建統包工程

運動地墊及拳擊訓練擂台採購案

109年度新建餐廳3樓休閒中心購置遊戲機1批案

檢驗科設備、藥劑科設備、病理科設備、轉譯中心設備、護理科設備及氣體設備等10項
超音波焊接(熱熔)機組乙項

臺中市西屯區協和國民小學幼兒園搬遷修繕工程委託技術服務勞務採購

找到其他資料欄位

```
print(soup.select('tbody')[0].select('tr')[0].select('td')[2].string)
print(soup.select('tbody')[0].select('tr')[0].select('td')[3].a.div.string)
print(soup.select('tbody')[0].select('tr')[0].select('td')[6].string)
print('https://web.pcc.gov.tw'+soup.select('tbody')[0].select('tr')[0].select('a')[0]['href'])
```

臺南市體育處

109年臺南亞太國際棒球訓練中心少棒園區委託管理勞務採購案

109/07/27

https://web.pcc.gov.tw/tps/tpam/main/tps/tpam/tpam_tender_detail.do?searchMode=common&scope=F&primaryKey=53204344

製作excel報表

```
wb = openpyxl.Workbook()
```

```
ws = wb.active
```

```
ws['A1'] = '招標機關'
```

```
ws['B1'] = '標案名稱'
```

```
ws['C1'] = '截止投標'
```

```
ws['D1'] = '標案種類'
```

```
ws['E1'] = '預算金額'
```

```
ws['F1'] = '網址'
```

```
ws['I1'] = '聯絡人'
```

```
ws['G1'] = '連絡電話'
```

```
wb.save('招標清單'+str(datetime.date.today())+'.xlsx')
```

```
print('報表製作完成')
```

填入excel欄位

```
ws['E1'] = '預算金額'  
ws['F1'] = '網址'  
ws['I1'] = '聯絡人'  
ws['G1'] = '連絡電話'
```

```
for item in soup.select('tbody')[0].select('tr'):  
    ws.append([item.select('td')[2].string,\  
               item.select('td')[3].a.div.string,\  
               item.select('td')[6].string,\  
               '',\  
               '',\  
               'https://web.pcc.gov.tw'+item.select('a')[0]['href'],\  
               '',\  
               ''])
```

```
wb.save('招標清單'+str(datetime.date.today())+'.xlsx')
```

```
print('報表製作完成')
```

報表製作完成

A1 招標機關					
A	B	C	D	E	F
1	招標機關	標案名稱	截止投標	標案種類	預算金額 網址
2	臺南市體育處	109年臺南亞太國際棒球訓練中心少棒園區委託管理勞務採購案	109/07/27		https://web.pcc.gov.tw/tps/tpam/main/tps/tpam/tpam_tender_detail.do?searchMode=
3	國防部空軍司令部	投影系統維護案	109/07/28		https://web.pcc.gov.tw/tps/tpam/main/tps/tpam/tpam_tender_detail.do?searchMode=
4	財團法人國際合作發展基金會	110年度國際人力資源培訓研習班計畫會議服務	109/08/03		https://web.pcc.gov.tw/tps/tpam/main/tps/tpam/tpam_tender_detail.do?searchMode=
5	台灣電力股份有限公司通霄發電廠	109年度#1機氣渦輪機輔機大修工作	109/07/28		https://web.pcc.gov.tw/tps/tpam/main/tps/tpam/tpam_tender_detail.do?searchMode=
6	花蓮縣消防局	花蓮縣消防局特種搜救隊訓練基地建置新建統包工程	109/07/22		https://web.pcc.gov.tw/tps/tpam/main/tps/tpam/tpam_tender_detail.do?searchMode=
7	新北市立鶯歌高級工商職業學校	運動地墊及拳擊訓練擂台採購案	109/07/21		https://web.pcc.gov.tw/tps/tpam/main/tps/tpam/tpam_tender_detail.do?searchMode=
8	國家運動訓練中心	109年度新建餐廳3樓休閒中心購置遊戲機1批案	109/07/23		https://web.pcc.gov.tw/tps/tpam/main/tps/tpam/tpam_tender_detail.do?searchMode=
9	衛生福利部桃園醫院	檢驗科設備、藥劑科設備、病理科設備、轉譯中心設備、護理科設備及氣體設備等10項	109/07/30		https://web.pcc.gov.tw/tps/tpam/main/tps/tpam/tpam_tender_detail.do?searchMode=
10	國防部軍備局生產製造中心	超音波焊接(熱熔)機組乙項	109/07/22		https://web.pcc.gov.tw/tps/tpam/main/tps/tpam/tpam_tender_detail.do?searchMode=
11	臺中市西屯區協和國民小學	臺中市西屯區協和國民小學幼兒園搬遷修繕工程委託技術服務勞務採購	109/07/23		https://web.pcc.gov.tw/tps/tpam/main/tps/tpam/tpam_tender_detail.do?searchMode=
12	新北市立聯合醫院	銀光活力中心健身器材-訓練設備採購1批	109/07/22		https://web.pcc.gov.tw/tps/tpam/main/tps/tpam/tpam_tender_detail.do?searchMode=
13	勞動部勞動力發展署中彰投分署	第46屆國際技能競賽選手訓練設備-電動設備工具組採購案	109/07/27		https://web.pcc.gov.tw/tps/tpam/main/tps/tpam/tpam_tender_detail.do?searchMode=
14	臺北大眾捷運股份有限公司	電聯車錄影主機採購	109/07/30		https://web.pcc.gov.tw/tps/tpam/main/tps/tpam/tpam_tender_detail.do?searchMode=
15	國立白河高級商工職業學校	多元體適能訓練教學空間活化工程	109/07/24		https://web.pcc.gov.tw/tps/tpam/main/tps/tpam/tpam_tender_detail.do?searchMode=
16	陸軍軍官學校	訓練用模型槍採購案	109/07/27		https://web.pcc.gov.tw/tps/tpam/main/tps/tpam/tpam_tender_detail.do?searchMode=
17	彰化縣立二林高級中學	109年度充實體育器材設備暨基層運動選手訓練站改善訓練環境及器材設備採購	109/07/28		https://web.pcc.gov.tw/tps/tpam/main/tps/tpam/tpam_tender_detail.do?searchMode=
18	苗栗縣立維真國民中學	維真國中109年度基層運動選手訓練站器材設備採購	109/07/21		https://web.pcc.gov.tw/tps/tpam/main/tps/tpam/tpam_tender_detail.do?searchMode=
19	基隆市立正濱國民中學	109年基層運動選手訓練站改善訓練環境及器材設備	109/07/28		https://web.pcc.gov.tw/tps/tpam/main/tps/tpam/tpam_tender_detail.do?searchMode=
20	誠正中學	109年機車修護技能訓練材料開口合約採購案	109/07/22		https://web.pcc.gov.tw/tps/tpam/main/tps/tpam/tpam_tender_detail.do?searchMode=
21	台灣電力股份有限公司興達發電廠	複1機核心組件升級改善附屬設備大修工作	109/08/18		https://web.pcc.gov.tw/tps/tpam/main/tps/tpam/tpam_tender_detail.do?searchMode=
22	誠正中學	109年汽車修護技能訓練材料開口合約採購案	109/07/22		https://web.pcc.gov.tw/tps/tpam/main/tps/tpam/tpam_tender_detail.do?searchMode=
23	內政部警政署	建置互動式情境模擬射擊訓練靶場10套案	109/08/18		https://web.pcc.gov.tw/tps/tpam/main/tps/tpam/tpam_tender_detail.do?searchMode=
24	臺灣土地銀行股份有限公司	第三方服務供應商辦理紐約分行查核作業	109/07/22		https://web.pcc.gov.tw/tps/tpam/main/tps/tpam/tpam_tender_detail.do?searchMode=
25	衛生福利部桃園醫院	眼科、急診、心臟內科、心臟外科、護理科設備7項	109/07/22		https://web.pcc.gov.tw/tps/tpam/main/tps/tpam/tpam_tender_detail.do?searchMode=

二層爬蟲架構

```
for item in soup.select('tbody')[0].select('tr'):
    res2 = requests.get('https://web.pcc.gov.tw'+item.select('a')[0]['href'])
    soup2 = BeautifulSoup(res2.text)
    ws.append([item.select('td')[2].string,\
               item.select('td')[3].a.div.string,\
               item.select('td')[6].string,\
               '\',\
               soup2.select('table')[0].select('tr')[57].td.text,\
               'https://web.pcc.gov.tw'+item.select('a')[0]['href'],\
               '\',\
               '''])
    delay = random.randint(10,20)
    print('已抓取一筆資料，等待'+str(delay)+'秒')
    time.sleep(delay)
```

二層爬蟲去除多餘字元

```
print(soup2.select('table')[0].select('tr')[i].td.text.replace('\t','').replace('\n','').replace('\r',''))
```

常見多餘字元	
\t	TAB
\n	空行
\r	回車
	半形、全形空白

二層爬蟲判斷架構

```
res2 = requests.get('https://web.pcc.gov.tw'+soup.select('tbody')[0].select('tr')[3].select('td')[3].a['href'])
soup2 = BeautifulSoup(res2.text)
i=40
while i<60:
    i+=1
    if soup2.select('table')[0].select('tr')[i].th.text.replace('\t','').replace('\n','').replace('\r','') == '聯絡人':
        print(soup2.select('table')[0].select('tr')[i].td.text.replace('\t','').replace('\n','').replace('\r',''))
    if soup2.select('table')[0].select('tr')[i].th.text.replace('\t','').replace('\n','').replace('\r','') == '聯絡電話':
        print(soup2.select('table')[0].select('tr')[i].td.text.replace('\t','').replace('\n','').replace('\r',''))
    if soup2.select('table')[0].select('tr')[i].th.text.replace('\t','').replace('\n','').replace('\r','') == '預算金額':
        print(soup2.select('table')[0].select('tr')[i].td.text.replace('\t','').replace('\n','').replace('\r',''))
```


加time.sleep控制流量！

```
for item in soup.select('tbody')[0].select('tr'):
    res2 = requests.get('https://web.pcc.gov.tw'+item.select('a')[0]['href'])
    soup2 = BeautifulSoup(res2.text)
    ws.append([item.select('td')[2].string,\
               item.select('td')[3].a.div.string,\
               item.select('td')[6].string,\
               '\',\
               soup2.select('table')[0].select('tr')[57].td.text,\
               'https://web.pcc.gov.tw'+item.select('a')[0]['href'],\
               '\',\
               '''])
    delay = random.randint(10,20)
    print('已抓取一筆資料，等待'+str(delay)+'秒')
    time.sleep(delay)
```

抓取過快則需手動驗證

IndexError

Traceback (most recent call last)

<ipython-input-7-224f217611c4> in <module>

41 item.select('td')[6].string,\

42 '',\

---> 43 soup2.select('table')[0].select('tr')[57].td.text,\

44 'https://web.pcc.gov.tw'+item.select('a')[0]['href'],\

45 '',\

IndexError: list index out of range



註 如何使用

為預防惡意程式針對本系統進行大量查詢致影響系統服務品質，並確保資訊安全，請於B區挑選與A區相同之撲克牌後送出。

註：◎B區被點擊之撲克牌會出現紫色粗框，表示該撲克牌已被挑選，若欲取消挑選，請再點擊該撲克牌1次，紫色粗框會消失。



確認送出