課程名稱: 進階程式設計(一)

題目: Python 網頁資訊擷取

班級: 資工 1A

學號: 108021011

姓名: 蘇邑洋

日期: 2020/5/30

完成項目: (90%)

Youtube: https://youtu.be/Wpw38vjXg7g

執行過程與結果:分別有(60%、70%、80%、90%)的步驟

(60%)的部分: 用網路爬蟲到亞大資工系網頁擷取出專任教授（姓名+學歷+辦公室+分機+E-mail)。

首先，我先引導 Beautiful Soup 的模組進來

```
from bs4 import BeautifulSoup
```

再來 get 取得網頁連結和寫入 requests 的模組

```
import requests
url="https://csie.asia.edu.tw/faculty/professors"
r = requests.get(url)
```

再以以 Beautiful Soup 解析 HTML 的程式碼

```
soup = BeautifulSoup(r.text, 'lxml')
```

分析之後，再來就是取的你所要擷取的內容的標籤

```
tab_table = soup.find(attrs={'class': 'row contact-category'})
```

上圖，是先取的你要抓的資料的區域再如(下圖)


並找到區域裡面的一個範圍

```
professors = tab_table.find_all('div', class_='col-sm-12')
```


因為要寫入 csv 檔案所以一開始先寫一個 csv 的模組

```
import csv
```

然後要寫進去之前先開啟 csv 檔案

```
with open(csvfile, 'w+', newline='', encoding='big5') as fp:
```

，再來因為用 list 會產生對錯欄位的情形，因此我用字典，並直接使用 csv.DictWriter 直接將 dictionary 寫入 csv 檔案中，

```
field_names = ['姓名', '學歷', '辦公室', '分機', 'E-mail']
writer = csv.DictWriter(fp, field_names)
writer.writeheader()
```

設一個 for 的 range(professors)，並設一個空的字典

```
for professor in professors:
```

```
record_dict = {}
```

設一個 professor_name 找出姓名

```
professor_name = professor.find('h2', 'card-header').string
```

並直接用 dict[]=”” 的方式 直接對取 key 和 values

```
record_dict['姓名'] = professor_name
```

再來 find 找出 學歷 , 辦公室 , 分機 ,E-mail

```
card_descriptions = professor.find_all('p', class_='card-description')
```

一樣設一個變數然後 for xx in card_descriptions 取得區域內的範圍

```
for card_description in card_descriptions:
```

可以發現他們都是有:來分隔

學歷：美國南加州大學資訊工程博士

辦公室：I513

分機：1831

E-mail：arbee@asia.edu.tw

因此我們可以用 split(“:”)來分割成一個陣列

```
splited_text = card_description.text.split(':')
```

```
field_name = splited_text[0].strip()
field_value = splited_text[1].replace('\t', '').replace(' ', '')
```

分割完之後，再用 dict[]=”” 放進去 一開始設的空字典 record_dict={}，並寫進去 csv 檔案就好，

```
    record_dict[field_name] = field_value
writer.writerow(record_dict)
```

60%的部分結束。

(70%)的部分：把專任教授、專任副教授、專任助理教授儲存成一個 csv。

首先，因為要用到 3 個連結所以我用 for 迴圈，並 range 設定成 3，然後遞減到 1，如下圖

```
for i in range(3,0,-1):
```

，然後就是 i 是幾的時候用哪個 url，並生成 csv

```
if i==3:
    url="https://csie.asia.edu.tw/faculty/professors"
    csvfile = "108_CSIE_Faculty_專任教授.csv"
elif i==2:
    url="https://csie.asia.edu.tw/faculty/associate-professors"
    csvfile = "108_CSIE_Faculty_專任副教授.csv"
elif i==1:
    url="https://csie.asia.edu.tw/faculty/assistant-professors"
    csvfile = "108_CSIE_Faculty_專任助理教授.csv"
```

，然後為了整合分析我把 3 個 csv 合併起來一個新的 csv，而方法我則用 panda 一次讀取 3 個檔案並整合，就成功了，

```
import pandas
vms = pandas.read_csv('108_CSIE_Faculty_專任教授.csv', encoding="big5")
users = pandas.read_csv('108_CSIE_Faculty_專任副教授.csv', encoding="big5")
sas =  pandas.read_csv('108_CSIE_Faculty_專任助理教授.csv', encoding="big5")
merged_df = pandas.concat([vms, users,sas], axis = 1, join = 'outer')
merged_df.to_csv('108_CSIE_Faculty_蘇昌洋.csv', encoding="big5")
```

合併玩的結果，如下圖

| | 姓名 | 學歷 | 辦公室 | 分機 | E-mail | 姓名 | 學歷 | 辦公室 | 分機 | E-mail | 姓名 | 學歷 | 辦公室 | 分機 | E-mail |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 蔡進發 （ | 美國西北大 | 學電機工 | 1007 | president@ | 陳兆南(Ch | 長庚大學 | 8019 | 48019 | chencn@as | 呂咸甫(W | 國立交通 | 8030 | 48030 | weifu@asia.edu.tw |
| 1 | 陳艮弼(Ar | 美國南加州 | I513 | 1831 | arbee@asia | 陳瑞奇(Jui | 國立中興 | HB13 | 20013 | rikki@asia. | 楊偉儒(W | 國立交通 | HB33 | 1843 | wzyang@asia.edu.tw |
| 2 | 許健(Gene | 愛荷華大 | I406 | 1784 | g_sheu@as | 莊政宏(Ch | 中正大學 | 8035 | 48035 | chchuang@ | 關國裕(Ku | 國立成功 | HB65 | 20065 | kky@asia.edu.tw |
| 3 | 黃明祥(Mi | 國立交通 | I420 | 1864 | mshwang@ | 龔自民(Tz | 國立交通 | 8009 | 48009 | tlkung@asia.edu.tw | | | | | |
| 4 | 薛榮銀(Zo | 美國賓州 | H41G-4 | 1729 | zshael@as | 蔡志仁(Zh | 長庚大學 | 8029 | 48029 | ren@asia.edu.tw | | | | | |
| 5 | 許慶賢(Ch | 逢甲大學 | H602 | 6303 | robertchh@ | 朱學亭(Hs | 國立清華 | I511 | 1833 | htchu@asia.edu.tw | | | | | |
| 6 | 陳興忠(Hs | 國立中正 | 8015 | 48015 | cdma2000@ | 王經篤(Jin | 國立中正 | I517 | 1847 | jdwang@asia.edu.tw | | | | | |
| 7 | 施能義(Ne | 成功大學 | I624 | 1813 | shih@asia. | 周永振(Yu | 國立中正 | 8005 | 48005 | yungchen@gmail.com | | | | | |
| 8 | 陳永欽(Ye | 成功大學 | 8034 | 48034 | ycchenster | 何承遠(Ch | 交通大學 | I412 | 1852 | tommyho@asia.edu.tw | | | | | |
| 9 | 沈偉誌(W | 中正大學 | 8025 | 48025 | wcshen@g | 張剛鳴(Ka | 交通大學 | 資訊大樓 | 20003 | changkm@asia.edu.tw | | | | | |
| 10 | Tadao Mur | 伊利諾大學 | 厄巴納-香檳分校 | 電機工程博 | | 游瑞松(Ru | 國立中興 | 資訊大樓 | 20052 | rsyu@asia.edu.tw | | | | | |
| 11 | 王緒斌(Bi | 賓夕法尼亞州立大學 | 工業工程博士 | | | 吳俊賢(Ch | 中山大學 | 資訊大樓 | 20015 | chwu@asia.edu.tw | | | | | |
| 12 | 張文鐘(W | 卡內基美 | 資訊大樓 | 1820 | wtchang@: | 謝長傚(Ch | 成功大學 | 資訊大樓 | 20011 | cwhsien@asia.edu.tw | | | | | |
| 13 | 連耀南(Y | 普渡大學 | 資訊大樓 | 1811 | yaonanlien | 陳榮燊(Ru | 英國倫敦 | 二宿地下 | 48012 | rschen@asia.edu.tw | | | | | |
| 14 | 蕭進松(Ch | 清華大學 | 健康大樓 | 6310 | cshsiao@as | 葉榮輝(Rc | 中央大學 | 資訊大樓 | 20016 | rhyeh@asia.edu.tw | | | | | |
| 15 | | | | | | 柯腎愷(H | 大同大學 | 資訊大樓 | 20042 | hiko@asia.edu.tw | | | | | |

70%的部分結束。

(80%)的部分：選擇一個你喜歡的網站(說明動機與目的）設計 WebRobot 下載網頁資訊,並儲存成檔案 csv。

動機與目的：

我選擇 ptt 的 food 版，看我的身材就知道我愛吃，平常都會看有甚麼推薦好吃的，等到有出去遊玩的話，都會去朝聖一下。

跟前面一樣先用 request.get 和 BeautifulSoup 來取得網站許可和標籤，

```
r = requests.get(url)
soup = BeautifulSoup(r.text,"html.parser")
btn = soup.select('div.btn-group > a')
```

再來，取得 div 內 class 為 btn-group 下的 a 標籤，

(上)在第 3 個 Index

```
▼<div class="btn-group btn-group-paging">
  <a class="btn wide" href="/bbs/Food/index1.html">最舊</a>
  <a class="btn wide" href="/bbs/Food/index7010.html">‹ 上頁</a>
  <a class="btn wide disabled">下頁 ›</a>
  <a class="btn wide" href="/bbs/Food/index.html">最新</a>
```

並設一個 get_href 回傳給`def get_href(url):`

```
page = ptt_btn[3]['href']
nextpage = 'https://www.ptt.cc' + page
url = nextpage
get_all_href(url = url)
```

，回傳之後，也就和前面一樣找出標籤和節點

```
def get_href(url):
        r = requests.get(url)
        soup = BeautifulSoup(r.text, "html.parser")
        results = soup.select("div.title")
```

，然後設立一個開啟 csv 的檔案，要寫進去

```
with open(csvfile, 'w+', newline='', encoding='big5') as fp:
    writer = csv.writer(fp)
```

，並設一個 dict 然後 writer 進去

```
field_names = ['標題','網址']
```

```
writer.writerow(field_names)
```

再用 for 迴圈再 results 的區域找出文字檔 text 和網址連結，並用 dict[]=”” 的方法給予 values

```
for item in results:
    title = item.text.replace("\u6ca2","").replace("\t","").replace(" ","").strip()
    yes_item=item.select_one('a')
    b=yes_item.get('href')
    field_names[0]=title
    field_names[1]='https://www.ptt.cc'+b
```

然後設一個 判斷句，因為會有刪文情形則會 none，

所以要確認是否有值，才取 href，並 writer 進去 csv 裡面，如下圖。

```
if yes_item:
      print(field_names)
      writer.writerow(field_names)
```

80%的部分結束。

(90%)的部分：自行設計改進的功能。

首先，我更改進老師的 list 沒辦法取得正確欄位的空格，然後用 dict 的 key 和 value 來解決位子不對的問題，如下圖兩張的改變，

```
for One in ALL_Professors:
    OneRecordList = []
    print (One.find('h2','card-header').string)
    OneRecordList.append(One.find('h2','card-header').string)
    ALL_card_description = One.find_all("p",class_="card-description")

    for oneItem in ALL_card_description:
        OneRecordList.append(oneItem.text.split('：')[1].replace("\t","").replace(" ",""))
    print(OneRecordList)
    writer.writerow(OneRecordList)
```

```
for professor in professors:
    record_dict = {}
    professor_name = professor.find('h2', 'card-header').string
    record_dict['姓名'] = professor_name
    card_descriptions = professor.find_all('p', class_='card-description')
    for card_description in card_descriptions:
        splited_text = card_description.text.split('：')
        field_name = splited_text[0].strip()
        field_value = splited_text[1].replace('\t', '').replace(' ', '')
        record_dict[field_name] = field_value
    writer.writerow(record_dict)
```

，還有我運用了新的一個叫 dict 的 key 先寫入 csv

的方式，用運了 writer.writeheader()，把 key 先寫進去，之後寫入時直接對照填入 value

```
field_names = ['姓名', '學歷', '辦公室', '分機', 'E-mail']
writer = csv.DictWriter(fp, field_names)
writer.writeheader()
```

。之後 70%部分的時候，我用 for 迴圈一次產生 3 個 csv 檔案，如下圖

```
for i in range(3,0,-1):
    if i==3:
        url="https://csie.asia.edu.tw/faculty/professors"
        csvfile = "108_CSIE_Faculty_專任教授.csv"
    elif i==2:
        url="https://csie.asia.edu.tw/faculty/associate-professors"
        csvfile = "108_CSIE_Faculty_專任副教授.csv"
    elif i==1:
        url="https://csie.asia.edu.tw/faculty/assistant-professors"
        csvfile = "108_CSIE_Faculty_專任助理教授.csv"
```

，這時的我發現這樣好單調沒啥意思，於是去網路是找到一個可以直接 3 個合併再一起一個新的 csv 檔案，那神奇的東西就是 panda

```
import pandas
```

```
vms = pandas.read_csv('108_CSIE_Faculty_專任教授.csv', encoding="big5")
users = pandas.read_csv('108_CSIE_Faculty_專任副教授.csv', encoding="big5")
sas =  pandas.read_csv('108_CSIE_Faculty_專任助理教授.csv', encoding="big5")
merged_df = pandas.concat([vms, users,sas], axis = 1, join = 'outer')
merged_df.to_csv('108_CSIE_Faculty_蘇昌洋.csv', encoding="big5")
```

後面加,encoding=”big5”才不會產生亂碼，之後就可以直接產生新的 csv，下圖展示

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 姓名 | 學歷 | 辦公室 | 分機 | E-mail | 姓名 | 學歷 | 辦公室 | 分機 | E-mail | 姓名 | 學歷 | 辦公室 | 分機 | E-mail | |
| 0 | | 蔡進發 （ | 美國西北大學 | 電機工 | 1007 | president@ | 陳兆南(Ch | 長庚大學 | 8019 | 48019 | chencn@as | 呂威甫(W | 國立交通 | 8030 | 48030 | weifu@asia.edu.tw | |
| 1 | | 陳艮弼(Ar | 美國南加州 | I513 | 1831 | arbee@asia | 陳瑞奇(Jui | 國立中興 | HB13 | 20013 | rikki@asia. | 楊偉儒(W | 國立交通 | HB33 | 1843 | wzyang@asia.edu.tw | |
| 2 | | 許健(Gene | 愛荷華大 | I406 | 1784 | g_sheu@as | 莊政宏(Ch | 中正大學 | 8035 | 48035 | chchuang@ | 關國裕(Ku | 國立成功 | HB65 | 20065 | kky@asia.edu.tw | |
| 3 | | 黃明祥(Mi | 國立交通 | I420 | 1864 | mshwang@ | 龔自良(Tz | 國立交通 | 8009 | 48009 | tlkung@asia.edu.tw | | | | | | |
| 4 | | 薛榮銀(Zo | 美國賓州 | H41G-4 | 1729 | zshael@as | 蔡志仁(Zh | 長庚大學 | 8029 | 48029 | ren@asia.edu.tw | | | | | | |
| 5 | | 許慶賢(Ch | 逢甲大學 | H602 | 6303 | robertchh@ | 朱學亭(Hs | 國立清華 | I511 | 1833 | htchu@asia.edu.tw | | | | | | |
| 6 | | 陳興忠(Hs | 國立中正 | 8015 | 48015 | cdma2000@ | 王經篤(Jin | 國立中正 | I517 | 1847 | jdwang@asia.edu.tw | | | | | | |
| 7 | | 施能義(Ne | 成功大學 | I624 | 1813 | shih@asia. | 周永振(Yu | 國立中正 | 8005 | 48005 | yungchen@gmail.com | | | | | | |
| 8 | | 陳永欽(Ye | 成功大學 | 8034 | 48034 | ycchenster | 何承遠(Ch | 交通大學 | I412 | 1852 | tommyho@asia.edu.tw | | | | | | |
| 9 | | 沈偉誌(W | 中正大學 | 8025 | 48025 | wcshen@g | 張剛鳴(Ka | 交通大學 | 資訊大樓 | 20003 | changkm@asia.edu.tw | | | | | | |
| 10 | | Tadao Mur | 伊利諾大學 | 厄巴納-香檳 | 分校電機工程博 | | 游瑞松(Ru | 國立中興 | 資訊大樓 | 20052 | rsyu@asia.edu.tw | | | | | | |
| 11 | | 王緒斌(Bo | 賓夕法尼亞州立大學 | 工業工程博士 | | | 吳俊賢(Ch | 中山大學 | 資訊大樓 | 20015 | chwu@asia.edu.tw | | | | | | |
| 12 | | 張文鐘(W | 卡內基美 | 資訊大樓 | 1820 | wtchang@ | 謝長倭(Ch | 成功大學 | 資訊大樓 | 20011 | cwhsien@asia.edu.tw | | | | | | |
| 13 | | 連耀南(Y | 普渡大學 | 資訊大樓 | 1811 | yaonanlien | 陳榮淼(Ru | 英國倫敦 | 二宿地下 | 48012 | rschen@asia.edu.tw | | | | | | |
| 14 | | 蕭進松(Ch | 清華大學 | 健康大樓 | 6310 | cshsiao@as | 葉榮輝(Rc | 中央大學 | 資訊大樓 | 20016 | rhyeh@asia.edu.tw | | | | | | |
| 15 | | | | | | | 柯賢儒(Hs | 大同大學 | 資訊大樓 | 20042 | hjko@asia.edu.tw | | | | | | |

，至於，爬自己喜歡的網站的同時則是學會了 def 回傳，其實就跟 java 很像，沒啥改變。

90%的部分結束。

## 討論與問題:

這次的功課也是上次的放狗題，我跟同學因為沒做出來，而回去馬上討論，但因為我們都是爬蟲菜鳥，對 python 也沒有任何經驗。因此，我們翻閱了許多書籍，查閱了許多網路資料，學到了 python 的字典應用和回傳之類的東西，而終於做出來 60%的部分，至於 70%呢，我很快就想出用 for 迴圈就好，因為讀取多個連結就和讀取資料一樣，繼上次期中專題的經驗，很快就做出來，然後我因為覺得太單調，一個個用出的 csv 檔案沒啥分析解果，因此我用 panda 直接合併成一個 csv，打開就可以直接用 excel 分

析，多麼方便的模組，還好有學了一下。至於 80%的部分，則是有參考一些資料，剛好又找到自己興趣類的方向，則直接爬蟲，然後又發現到了 java 的回傳其實就是到 python 則變 def，感覺 python 用了 def 會更簡易許多，而不用像 java 一樣寫了一大串密密麻麻的 code，而讓自己看的都覺得疲累。則到 90%的最後部分，則是把發現的一些新大陸而加以改變到自己的程式碼，然後再稍微修改和創新，就成功用出比自己一開始想出的更理想的程式。

## 心得:

做這個期中專題，讓我學到了許多新的程式應用和一些技巧。我跟同學一起討論並解決問題，解決問題的當下，心理是非常舒暢的。畢竟，我跟這些同學，都是對 python 沒有任何的觀念，當初學 java 也是，都是互相討論和自己翻書查閱網路資料，才呈現自己努力的結果，其實有先學的人都是有優勢的，而我們也只能從零開始慢慢學，積少成多，總有一天也是會變成程式高手的。因此，每一次做的程式碼我都會上

傳到 github 裡面，以後可能實習或面試會用到，這時就可以拿你大學四年所學到的展現給面試官看，所以我現在應該多買一些程式的書或看網路來學習更多的技巧，而不是只學上課的。持續下去，我相信有一天可以展現出成果的，還有學程式不求快，只求精，這樣才是我覺得會進步的方法。