網頁資料擷取: <http://ithelp.ithome.com.tw/question/10127973>

筆者用urllib抓網頁，urllib在python2/3版本的差異頗大，有些是傳回值的型別不一樣，有些是方法，成員改寫掉了，  
本文是用python3的版本。  
抓網頁中的table裏的td欄位裏的值，筆者是用知名的**BeautifulSoup庫**來進行

營收報表受IFRS影響，  
所以2013(102)年沒法用同一個網址來擷取data,   
iInfo先生分享的程式，只需要做一個微小的改變。  
不過，筆者用此來做個不同庫的DEMO, 可以做個比較。  
  
在<http://mops.twse.com.tw/mops/web/index>公開資訊觀測站的  
位置：  
**彙總報表／資訊揭露／每月營收／採用IFRSs前營業收入匯總表／每月營業收入統計彙總表  
102年後資料請至採用IFRSs後營業收入彙總查詢。**  
當你Key入102年，會出現：  
102年後資料請至採用IFRSs**後**營業收入彙總查詢。  
  
所以要去：採用IFRSs**後**營業收入匯總表，抓取資料。  
我們試抓１，２，８月來觀察網址（ＵＲＬ）的規律：  
http://mops.twse.com.tw/t21/sii/t21sc03\_102\_1\_0.html  
http://mops.twse.com.tw/t21/sii/t21sc03\_102\_2\_0.html  
http://mops.twse.com.tw/t21/sii/t21sc03\_102\_8\_0.html  
看一下  
採用IFRSs**前**營業收入匯總表的網址，  
http://mops.twse.com.tw/t21/sii/t21sc03\_101\_8.html  
所以只要把  
iInfo先生分享的程式，

[檢視原始檔複製到剪貼簿列印關於](http://ithelp.ithome.com.tw/question/10127973)

1. url = "http://mops.twse.com.tw/t21/" + stocktype + "/t21sc03\_" + pyROCYear
2. "\_" + str(k) + ".html"

url = "http://mops.twse.com.tw/t21/" + stocktype + "/t21sc03\_" + pyROCYear

+ "\_" + str(k) + ".html"

改成

[檢視原始檔複製到剪貼簿列印關於](http://ithelp.ithome.com.tw/question/10127973)

1. url = "http://mops.twse.com.tw/t21/" + stocktype + "/t21sc03\_" + pyROCYear
2. "\_" + str(k) + "\_0.html"

url = "http://mops.twse.com.tw/t21/" + stocktype + "/t21sc03\_" + pyROCYear

+ "\_" + str(k) + "\_0.html"

只差在**\_0**。  
－－－－－－－－－－－－－－－－－－－－－－－－－－－－－－－－－－－－－－－－－－－  
那就開始練習urllib＋BeautifulSoup  
進入練習之前：  
  
可以了解有很多個table, 我們要擷取的data就在眾多個table裏。  
**練習一：藉由urllib.request庫把網頁整個抓回來，並轉給BeautifulSoup，記得要decode('cp950')**

[檢視原始檔複製到剪貼簿列印關於](http://ithelp.ithome.com.tw/question/10127973)

1. import urllib.request
2. from bs4 import BeautifulSoup
3. url = 'http://mops.twse.com.tw/t21/sii/t21sc03\_102\_1\_0.html'
4. response = urllib.request.urlopen(url)
5. html = response.read()
6. sp = BeautifulSoup(html.decode('cp950'))
7. sp

import urllib.request

from bs4 import BeautifulSoup

url = 'http://mops.twse.com.tw/t21/sii/t21sc03\_102\_1\_0.html'

response = urllib.request.urlopen(url)

html = response.read()

sp = BeautifulSoup(html.decode('cp950'))

sp

**練習二：把多個table找出來**

[檢視原始檔複製到剪貼簿列印關於](http://ithelp.ithome.com.tw/question/10127973)

1. tbls=sp.find\_all('table', attrs={ 'width' : "100%" })
2. tbls
3. tbls[0]
4. tbls[1]

tbls=sp.find\_all('table', attrs={ 'width' : "100%" })

tbls

tbls[0]

tbls[1]

**練習三：換找table的屬性，上一練習找到不需要的table, 多個屬性時，用字典(Dictionaries)型別來表示，字典型別對筆者來說，算是不太習慣，在這裏，key/value, 是非常恰當的用途！**

[檢視原始檔複製到剪貼簿列印關於](http://ithelp.ithome.com.tw/question/10127973)

1. tblh=sp.find\_all('table', attrs={ 'border' : '0','width' : '100%' })
2. tblh[0]

tblh=sp.find\_all('table', attrs={ 'border' : '0','width' : '100%' })

tblh[0]

在語義上，可以發現find\_all，find的差異，非常直覺  
**練習四：發現要抓取的data裏是，table裏再包一個table**

[檢視原始檔複製到剪貼簿列印關於](http://ithelp.ithome.com.tw/question/10127973)

1. tbl=tblh[0].find('table', attrs={ 'bordercolor' : "#FF6600" })
2. tbl

tbl=tblh[0].find('table', attrs={ 'bordercolor' : "#FF6600" })

tbl

**練習五：營收數字的table裏，把每一列（tr）抓出來**

[檢視原始檔複製到剪貼簿列印關於](http://ithelp.ithome.com.tw/question/10127973)

1. trs=tbl.find\_all('tr')
2. trs[0],trs[1],trs[2]

trs=tbl.find\_all('tr')

trs[0],trs[1],trs[2]

看了資料，知道trs[0],trs[1]是中文欄位說明，所以要塞資料庫(sqlite3)時，從第trs[2]開始解析欄位值  
**練習六：list的長度，用於抓取最後一列**

[檢視原始檔複製到剪貼簿列印關於](http://ithelp.ithome.com.tw/question/10127973)

1. ln=len(trs)
2. ln
3. trs[ln-1],trs[ln-2],trs[ln-3]

ln=len(trs)

ln

trs[ln-1],trs[ln-2],trs[ln-3]

用len(List)知道這個table裏有幾列（tr），最後一列是ln-1，是合計列，也不用寫入資料庫(sqlite3)  
**練習七：從每一列裏，得到每一欄（td）值, 做到此，大家都知道用find\_all('td')吧**

[檢視原始檔複製到剪貼簿列印關於](http://ithelp.ithome.com.tw/question/10127973)

1. tds=trs[2].find\_all('td')
2. tds[0]
3. tds[0].get\_text(),tds[1].get\_text(),tds[2].get\_text(),tds[3].get\_text(),tds[4].get\_text()
4. output:
5. ('1101', '台泥', '              9,134,465', '不適用', '              7,125,036')

tds=trs[2].find\_all('td')

tds[0]

tds[0].get\_text(),tds[1].get\_text(),tds[2].get\_text(),tds[3].get\_text(),tds[4].get\_text()

output:

('1101', '台泥', ' 9,134,465', '不適用', ' 7,125,036')

**練習八：把數字欄位的空白去掉，千分位，十萬分位的逗點去掉**

[檢視原始檔複製到剪貼簿列印關於](http://ithelp.ithome.com.tw/question/10127973)

1. td2=tds[2].get\_text()
2. td2=td2.strip().replace(",", "")
3. td2
4. output:
5. '9134465'

td2=tds[2].get\_text()

td2=td2.strip().replace(",", "")

td2

output:

'9134465'

**練習八：把數字欄位裏“不適用”，換成０**

[檢視原始檔複製到剪貼簿列印關於](http://ithelp.ithome.com.tw/question/10127973)

1. td3=tds[3].get\_text()
2. if 　td3 == (u'不適用')　:
3. td3='0'
4. td3
5. output:
6. if 　td3 == (u'不適用')　:
7. ^
8. SyntaxError: invalid character in identifier

td3=tds[3].get\_text()

if 　td3 == (u'不適用')　:

td3='0'

td3

output:

if 　td3 == (u'不適用')　:

　 ^

SyntaxError: invalid character in identifier

糗了，真的'不適用'  
肚子餓，頭昏，晚一點再找一下解法  
  
**小結：**互動式作業環境，對於解析網頁特別適用，之後再用迴圈把片斷的程式碼，  
組合起來，就能批次作業了。

**啟動web瀏覽器並指定網址**

import webbrowser

c = webbrowser.open('http://mops.twse.com.tw/mops/web/index')

**\*python3.4.1指定瀏覽器需要用其他方式，很麻煩!!!**