

中學生的Python趣學程 資料爬取與分析 Foundation of Al

跟著創新廣告和新創CEO
一起從創意發想,程式素養,創客修為到新創精神
玩轉程式趣學程





目次

Chapter	01	網頁資料擷取和轉換	. 1
1-1	網址	解析	. 1
		網頁內容	
		1 requests 套件取得網頁原始碼	
		2 BeautifulSoup 套件解析網頁	
	T-7-7	∠ DEGUUTUISOUP 云 下門小 刻 見	. /

封面改編圖像來源: 攝圖網



Chapter 01 網頁資料擷取和轉換

之前我們已玩過 Python 的基本語法,選擇判斷,回圈,資料型態,檔案和異常處理... 等.而這些都是接下來要玩網頁資料爬取和分析的基礎,而數據資料又是 AI 人工智慧的燃料,且網頁爬蟲資料分析其中重要環節就是相關 Pyhton 套件的運用,這些套件多數也應用於 AI. 可見它的重要性!

這篇網頁爬蟲資料分析主要以實戰應用為主,程式碼的數量和複雜度較之前的練習都提高許多,尤其對於國中生難度不小! 我們會以程式案例拆解的方式,降低難度,一步一步達到目標! 加油啦!

Python 網頁資料爬取過程一般如下:

- 1. 存取網站取得網頁內容(原始碼)
- 2. 解析網頁取得資料內容
- 3. 資料處理與展示 這章先來實作前兩項.

1-1 網址解析

我們擷取資料主要有取得 OpenData(開放資料)的檔案或抓取網頁中的內容資料.不管哪一種,這些資料檔案和網頁內容都在網站的網頁中,都需要讓程式能早找到網站,那程式就需要知道網址解析網址.

要解析網址可用 Python 中 urllib 套件的 urlparse 函式, 會傳回元組型態的 ParseResult 物件, 物件屬件中可取得網址的各種參數.

我們以台灣銀行告牌匯率網頁為例,

網址為 => https://rate.bot.com.tw/xrt?Lang=zh-TW





(ch1-1 exchangeRateTable.py)

```
8 from urllib.parse import urlparse
9 url = 'https://rate.bot.com.tw/xrt?Lang=zh-TW'
10 xrt = urlparse(url)
11 print(xrt)
```

In [1]: runfile('G:/Course/Python_DataAnalysis/ch1-1_exchangeRateTable.py', wdir='G:/
Course/Python_DataAnalysis')
ParseResult(scheme='https', netloc='rate.bot.com.tw', path='/xrt', params='',
query='Lang=Zh-TW', fragment='')

列	程式			說明	
8	from urllib.parse import	由urllib套	件的 par	se 模組引入 urlparse 方法	
	urlparse				
9	url =	將台灣銀行	告牌匯	率網頁的網址字串	
	'https://rate.bot.com.tw/	'https://rate.bot.com.tw/xrt?Lang=zh-TW'			
	xrt?Lang=zh-TW'	指派給變數 url			
10	xrt = urlparse(url)	用引入的 u	ırlparse	方法解析變數 url 的網址,	
		會傳回 Pars	seResult	的元祖物件, 元祖元素是	
		網址的各項	屬性		
11	print(xrt)	輸出 Parse	Result 的	元祖物件:	
		ParseResult	(scheme	='https',	
		netloc='rate	.bot.con	n.tw', path='/xrt',	
		params=", o	query='La	ang=zh-TW', fragment='')	
		ParseResult 元祖物件屬性說明:			
		Parsenesuit	人工作出7 97		
		屬性	索引值	回傳值	
		scheme	0	scheme 通訊協定	
		netloc	1	網站域名	
		path	2	路徑	
		params	3	url 查詢參數 params 字串	
		query	4	query 查詢字串,GET 參數	
		fragment	5	框架名稱	
		port	無	通訊埠	

繼續 ch1-1 輸出各屬性

```
8 from urllib.parse import urlparse
9 url = 'https://rate.bot.com.tw/xrt?Lang=zh-TW'
10 xrt = urlparse(url)
11 print(xrt)
12 print()
13 print('scheme =', xrt[0])
14 print('netloc =', xrt[1])
15 print('partm =', xrt[2])
16 print('params =', xrt[3])
17 print('query =', xrt[4])
18 print('fagment =', xrt[5])
19 print('port =', xrt.port)
In [5]: runfile('6:/Course/Python_DataAnalysis/chl-1_exchangeRateTable.py', wd:
Course/Python_DataAnalysis')
ParseResult(scheme='https', netloc='rate.bot.com.tw', path='/xrt', params='', query='Lang=zh-TW', fragment='')

scheme = https
netloc = rate.bot.com.tw
path = /xrt
params =
query = Lang=zh-TW
fragment =
port = None
```

Cocoding-Youth www.941coding.com

列	程式	說明
13	print('scheme =', xrt[0])	輸出'shceme'字串和元祖索引值為 0 的第一
		個元素=>通訊協定: scheme = https
14	print('netloc =', xrt[1])	輸出'netloc'字串和元祖索引值為1的第二
		個元素=>網站域名: netloc = rate.bot.com.tw
15	print('path =', xrt[2])	輸出'path'字串和元祖索引值為2的第三個
		元素=>路徑: path = /xrt
16	print('params =', xrt[3])	輸出'params'字串和元祖索引值為 3 的第四
		個元素=>url 查詢參數: 不存在, 傳回空字串
17	print('query =', xrt[4])	輸出'query'字串和元祖索引值為4的第五個
		元素=>查詢字串: query = Lang=zh-TW
18	print('fragment =', xrt[5])	輸出'fragment'字串和元祖索引值為 5 的第
		六個元素=>框架名稱: 不存在, 傳回空字串
19	print('port =', xrt.port)	輸出'port'字串和元祖通訊埠,因通訊埠在
		本機,沒有索引值,所以用 xrt.port 語法取
		得:不存在,傳回 None
		port = None

1-2 取得網頁內容

1-2-1 requests 套件取得網頁原始碼

requests 套件可以取得網頁原始碼. 流程如下:

- 1. 匯入 requests 套件. 使用 Anaconda 整合環境的好處是常用的數據分析和機器學習套件都已安裝, 直接匯入即可使用.
- 2. 用 requests.get()方法可模擬發出 HTTP GET 方法向伺服器送出請求 (request).
- 3. 當伺服器接受請求後,會回應(response)傳回網頁內容(原始碼).
- 4. 設定好編碼,以 text 屬性取得網頁原始碼.

(ch1-2 requestsXR.py)

我們抓取上節的台灣銀行告牌匯率網頁的原始碼為例,因該頁面原始碼有 2410 列程式敘述,我取得這 2410 列後,僅在螢幕輸出前 20 列.

首先看看網頁原始碼及要輸出的前 20 列:

Cacading-Youth www.941coding.com



- 1. 在網頁上按滑鼠右鍵, 出現上圖選單
- 2. 在選單上選擇「檢視網頁原始碼(v)」按下
- 3. 即可出現該網頁的原始碼, 如下圖:

```
這是網頁的 view-source
\leftarrow
          G
                 ① view-source
                                    https://rate.bot.com.tw/xrt?Lang=zh-TW
   <html lang="zh-TW" class="no-js">
   <head>
        <meta charset="utf-8" />
        <title>臺灣銀行牌告匯率</title>
        <meta http-equiv="X-UA-Compatible" content="IE=edge">
        <meta name="description" content="臺灣銀行匯率利率黃金牌價查詢">
        <meta name="keywords" content="">

        <meta name="format-detection" content="telephone=no">
             <meta name="robots" content="index,follow" />
        <link rel="icon" type="image/x-icon" href="/favicon.ico">
        k rel="stylesheet" href="/Content/css/font-awesome.min.css">
        這是抓取 2410 列全部網頁原始碼後, 要輸出的前 20 列
        <link rel="stylesheet" href="/Content/css/jquery-ui.min.css">
        <!ink rel="stylesheet" href="/Content/main-3.3.0.css";</pre>
22
        <style>
23
             h1.hero__header_small {
                  font-size: 26px;
                  margin: 12px 0;
25
             }
26
                  h1.hero__header_eng {
28
                      font-size: 20px;
29
30
                       margin: -8px 0;
32
33
             .chart-key-value .sp {
                  position: relative;
34
35
                  top: 2px;
             }
36
37
38
             select[name=month], select[name=day] {
39
                  width: 5em;
40
41
             :lang(en-US) {
42
```

Coalcoding/outh www.941coding.com

```
| Simport requests | Simport requests | Surl = 'https://rate.bot.com.tw/xrt?Lang=zh-TW' | Surl = 'https://rate.bot.com.tw/xrt?Lang=zh-TW' | Surl = 'nterest | Surl = 'quises | Surl = 'utf-8' | Surl = 'quises |
```

列	程式	說明
8	import requests	匯入 requests 套件
9	url =	將台灣銀行告牌匯率網頁的網址字串
	'https://rate.bot.com.tw/	'https://rate.bot.com.tw/xrt?Lang=zh-TW'
	xrt?Lang=zh-TW'	指派給變數 url
10	html = requests.get(url)	用 requests.get(url)方法向 url 網址所在伺服
		器發出請求,伺服器接受請求後,傳回 url
		網址網頁的原始碼,並指定給變數 html
11	html.encoding = 'utf-8'	設定 utf-8 編碼
		到 11 列, 使用 html.text 已取得該網頁 2410
		列的所有原始碼, 但原始碼過多, 繼續處理
		以輸出前 20 列觀察結果
12	htmllines =	html.text 取得的原始碼用 splitlines()方法,
	html.text.splitlines()	分割成以列為元素的串列. 並指派給變數
		htmllines
13,	for i in range(20):	用 for 回圈迭代 20 次, 14 列輸出 htmllines
14	print(htmllines[i])	串列索引值 0-19 的前 20 個元素, 即原始碼
		前 20 列

我們也可以修改第13列程式敘述,輸出後20列原始碼. 程式另存為 (ch1-2_requestsXR2.py)



輸出結果和原始碼比較

```
2385
                //footable 展開所有表格内容
2386
2387
                $(".click ExpandAll").click(function () {
                    $('.footable').trigger('footable_expand_all');
2388
2389
                })
                //footable 關閉所有fsear表格内容
2391
                $(".click_CollapseAll").click(function () {
2392
2393
                    $('.footable').trigger('footable_collapse_all');
2394
2395
                //取得最新xx
2396
                $(".click_reload").on("click", function (e) {
2397
                   location.reload();
2398
2399
                });
2400
                             原始碼後 20 列
            });
2401
2402
            $(window).load(function () {
2403
                //最近一個營業日不可比較幣別,比較幣別為空時不顯示現金&即期(for chrome)
2404
2405
                search_range_check($('input:radio:checked[name="search_range"]'));
                //alert("load");
2406
            })
2407
        </script>
2408
                 . 哇! 果然有 2410 列程式敘述, 超多的吧?!
2409
    </body>
    </html>
```

列	程式	說明
13,	for i in range(len(htmllines)-	用 for 回圈迭代 20 次, 14 列輸出 htmllines
14	20, len(htmllines)):	串列索引值 len(htmllines)-20 到
	print(htmllines[i])	len(htmllines)的最後 20 個元素, 即原始碼最
		後 20 列

如果進一步需了解「匯率」這個詞在網頁出現的次數,一樣將程式另存為 (ch1-3 requestsXRFrequency.py)

列	程式	說明
13	for i in range(len(htmllines)-	用 for 回圈迭代 13 次, 14 列輸出 htmllines
	13, len(htmllines)):	串列索引值 len(htmllines)-13 到
		len(htmllines)的最後 13 個元素,即原始碼最
		後 13 列,輸出較少資料,方便後續程式輸

COCOGINGWOULH WWW.941coding.com

		出觀察結果.
16	print('='*16)	輸出 16 個等號 '=', 用以區隔兩類輸出.
17	count = 0	定義變數 count, 設初始值為 0, 作為匯率出
		現次數的計數器.
18	for line in htmllines:	For 回圈以變數 line 作為迭代串列 htmllines
		的元素.
19,	if '匯率' in line:	If 判斷式, 如果'匯率'有在 line(原始碼的
20	count += 1	2410 列,迭代 2410 次)中,則變數 count 加一
21	print(""匯率"共出現{}次	For 回圈迭代完成,輸出出現總次數
	'.format(count))	==========
		"匯率"共出現 60 次

1-2-2 BeautifulSoup 套件解析網頁

我們爬取網頁內容,一般不是需要整個網站或網頁資料,而是其中特定資料.例如,台灣銀行告牌匯率網頁假設只需台銀各幣別現金買入匯率,如下紅框標示的 19 種幣別和台幣的匯率.





這時就需要功能更強的網頁解析工具 BeautifulSoup 套件. 流程如下:

- 1. 從 bs4 套件匯入 BeautifulSoup 模組. 使用 Anaconda 整合環境的好處是常用的數據分析和機器學習套件都已安裝,直接匯入即可使用.
- 2. 用 html.parser 解析 requests 取得的原使碼. 語發如下:
 - sp = BeautifulSoup(原使碼, 'html.parser')
 - sp 為 BeautifulSoup 的物件
- 3. 接下來就可以始用 BeautifulSoup 物件 sp 的方法和屬性解析網頁. BeautifulSoup 物件的常用方法和屬性:

方法或屬性	說明
title	傳回網頁標題;例: sp.title
text	傳回不含 HTML tag 的網頁文字內容
find()	傳回第一個符合條件的標籤(HTML tag);例:sp.find('tr')
	找到後傳回一個字串,找不到傳回 None
find_all()	傳回所有符合條件的標籤(HTML tag);例:sp.find_all('a')
	找到後傳回一個串列(list), 找不到傳回空串列([])
select()	傳回 CSS 選擇器(id 或 class)或標籤(HTML tag);例:sp.select('#id'),
	sp.select('.class'), sp.select('td')
	找到後傳回一個串列(list), 找不到傳回空串列([])

(ch1-4 BeautifulSoup1.py)

這個案例就來實作抓取 19 種幣別和台幣的匯率:

五山	4	デクロロ
列	程式	說明
8	import requests	匯入 requests 套件
9	from bs4 import	從 bs4 套件匯入 BeautifulSoup 模組
	BeautifulSoup	
10	url =	將台灣銀行告牌匯率網頁的網址字串
	'https://rate.bot.com.tw/	'https://rate.bot.com.tw/xrt?Lang=zh-TW'

Cacading-Youth www.941coding.com

	xrt?Lang=zh-TW'		指派給	變數 url				
11	html = requests.ge	t(url)	用 requ	uests.get(ur	1)方法	向 url 約	網址所在信	
		,	1	請求, 伺服				
			網址網	頁的原始碼	5, 並持	旨定給變	虁數 html	
12	html.encoding = 'utf-8'		設定 utf-8 編碼					
13	sp =		⊞ htm	用 html.parser 解析 requests 取得的原始碼				
	•	Ltout		xt, 傳回 Be				
	BeautifulSoup(htm	i.text,		, , , , ,			****	
	'html.parser')		變數 sp); 就可以使	^፻ 用 sp	Beautif	fulSoup 物作	
			的方法	和屬性解析	requ	ests 取行	得的原始码	
			html.te				133//3.0/12	
	上節已看過網頁	百胎碼	百胎碼	的第 264 列	i <tak< th=""><th>ule> 樗</th><th>築內建構"</th></tak<>	ule> 樗	築內建構"	
					•	// /示	11年1日、	
	包括對網頁按右			的匯率表格				
	「檢視網頁原始码	馬(v)」 <u></u>	原始碼	的第 265 列	∫ <the< th=""><th>ead> 標</th><th>鐵內建構</th></the<>	ead> 標	鐵內建構	
	用 requests 取得的	勺原始碼.	格表頭	,但我們要	的匯	率資料フ	下在此	
	而現在需研究網頁	頁結構(原	原始碼	的第 314 列	[] <tbo< th=""><th>ody> 標</th><th>籤內建構</th></tbo<>	ody> 標	籤內建構	
	始碼內容),才	可用 sp	格匯率	資料, 我們	要抓耳	取此處図	E 室 資料	
					×1/11-		巴丁吳们	
	BeautifulSoup 物化		263	The state of the s				
	和屬性找出資料的 html 標 籤或 css 選擇器, 取出特定		(table title="牌告匯率" class="table table-striped table-bordered table-conder 二直標定本・データ、現金匯率・第三直欄是本行賣出現金匯率・第四直欄是本行買人即期匯率・第五直標					
			265 <t< th=""><th>head class="phone-medi</th><th></th><th>Tam - North Walter</th><th>是本行買人即期匯率,第五</th></t<>	head class="phone-medi		Tam - North Walter	是本行買人即期匯率,第五	
	籤或 css 選擇器,	取出特定	266 267	thead class="phone-medi	um-font"> :_width set-	title-L-min-wid	dth-class noscript" r	
			266 267 268 269 270	<pre>chead :lass="phone-medi</pre>	um-font"> _width set- _width rowS _hide rate- n">	title-L-min-wid P_Ctrl_2_2_2 se content-cash "	dth-class noscript" r et-title-L-min-width- colspan="2">現金匯率	
	籤或 css 選擇器, 資料. 看下圖網頁		266 267 268 269 270 271 272	<pre>chead class="phone-medition</pre>	.um-font"> :_width set- :_width rowS :_hide rate- :n"> :_hide rate- :n">	title-L-min-wid P_Ctrl_2_2_2 se content-cash " content-sight "	dth-class noscript" r et-title-L-min-width-	
			266 267 268 269 270 271	thead class="phone-medical class="print class="print class="print class="print class="hidde ch class="print ch class="print class="prin	.um-font"> :_width set- :_width rowS :_hide rate- :n"> :_hide rate- :n"> :_hide" rows :_hide" rows	title-L-min-wid P_Ctrl_2_2_2 se content-cash " content-sight "	dth-class noscript" r at-title-L-min-width- colspan="2">現金匯率。 " colspan="2">即期匯率	
			266 267 268 269 270 271 272 273	thead class="phone-medical characteristics" (th class="print characteristics") (th class="print characteristics") (th class="print characteristics") (th class="print characteristics") (span class="print characteristics")	.um-font"> :_width set- :_width rowS :_hide rate- :n"> :_hide rate- :n"> :_hide" rows :_hide" rows	title-L-min-wid P_Ctrl_2_2_2 se content-cash " content-sight " pan="2">	dth-class noscript" r at-title-L-min-width- colspan="2">現金匯率。 " colspan="2">即期匯率	
		分析.	266 267 268 269 270 271 272 273 274 275	<pre>chead liass="phone-medi</pre>	.um-font"> :_width set- :_width rowS :_hide rate- :n"> :_hide rate- :n"> :_hide" rows :_hide" rows	title-L-min-wid P_Ctrl_2_2_2 se content-cash " content-sight " pan="2">	dth-class noscript" r at-title-L-min-width- colspan="2">現金匯率。 " colspan="2">即期匯率	
	資料. 看下圖網頁	分析. 15 ^{現金匯8}	266 267 268 269 270 271 272 273 274 275	thead lass="phone-mediation" (th class="print (th class="print (th class="print (th class="hidde (th) class="hidde	.um-font"> :_width set- :_width rowS :_hide rate- in"> :_hide rate- in"> :_hide rowS :_hide rowS :_"> 遠	title-L-min-wid P_Ctrl_2_2_2 se content-cash " content-sight " pan="2">	dth-class noscript" r at-title-L-min-width- colspan="2">現金匯率。 " colspan="2">即期匯率	
	資料. 看下圖網頁 □ 牌價嚴新掛牌時間: 2019/05/24 14:	分析.	266 267 268 269 270 271 272 273 274 275	*************************************	.um-font"> :_width set- :_width rowS :_hide rate- :n"> :_hide rate- :n"> :_hide" rows :_hide" rows	title-L-min-wic P_Ctrl_Z_2_2 secontent-cash " content-sight " content-sight " pan="2"> 明漢率 <td>dth-class moscript" r et-title-L-min-width- colspan="2">現金匯率 " colspan="2">即期匯率 span></td>	dth-class moscript" r et-title-L-min-width- colspan="2">現金匯率 " colspan="2">即期匯率 span>	
	資料. 看下圖網頁 □ 牌價最新掛牌時間: 2019/05/24 14:	分析. 15 現金層: 本行買人	206 207 208 209 209 270 271 272 273 274 275	** cth class="phone-mediation children" print cth class="print cth class="print cth class="print cth class="print cth class="midde cth class="midde cth class="midde cth class="midde cth class="midde cth class="print cth	um-font"> _width setwidth rowS _hide rate- in">	title-L-min-wid P_Ctrl_2_2_ se content-cash " content-sight " pan="2"> 期匯率 <td>dth-class noscript" ret-title-L-min-width-colspan="2")現金匯率 " colspan="2")規度區等 span></td>	dth-class noscript" ret-title-L-min-width-colspan="2")現金匯率 " colspan="2")規度區等 span>	
	資料.看下圖網頁 □ 牌價最新掛牌時間: 2019/05/24 14:	分析. 15 現金種語 本行買人 31.115	286 287 288 289 270 271 271 272 273 274 275	** cth class="phone-mediation characteristics" cth class="print cth class="print cth class="print cth class="print cth class="print cth class="midde cth class="midde cth class="midde cth class="midde cth class="print cth cth cth class="print cth cth cth cth cth cth class="print cth cth cth class="print cth cth cth cth cth cth cth cth cth ct	um-font"> _width setwidth rows _hide rate- n"> _hide rate- n"><	title-L-min-wid P_Ctrl_2_2_2 content-cash " content-sight " pan="2"> 明確率 <td>dth-class noscript" ret-title-L-min-width-colspan="2")現金匯率 " colspan="2")規則匯率 span> 歷史匯率 查詢 查詢</td>	dth-class noscript" ret-title-L-min-width-colspan="2")現金匯率 " colspan="2")規則匯率 span> 歷史匯率 查詢 查詢	
	資料. 看下圖網頁 □ 牌價嚴新掛牌時間: 2019/05/24 14:	分析. 15 現金種語	286 287 288 289 270 271 271 272 273 274 275 31.805 4.067 40.9	*** *** *** *** *** *** *** *** *** **	um-font") _width setwidth rowShide rate- n"> _hide rate- n"> _hide rate- n"> _hide rate- n"> _ide rate- n"> _ide rate- n"> _ide rate- 31.585 4.047	title-L-min-wid P_Ctrl_2_2_2 secontent-cash " content-sight " pan="2"> 明確率 <td>dth-class noscript" ret-title-L-min-width-colspan="2")现金匯率 " colspan="2">即期匯率 span> E.p.匯率 查詢 查詢 查詢</td>	dth-class noscript" ret-title-L-min-width-colspan="2")现金匯率 " colspan="2">即期匯率 span> E.p.匯率 查詢 查詢 查詢	
	資料. 看下圖網頁 □ 牌價嚴新掛牌時間: 2019/05/24 14:	分析. 15 現金種	286 287 288 289 270 271 272 273 274 275 274 275 31.805 4.067 40.9	whead lass="phone-mediation of the class="print of the class="pri	um-font") _width setwidth rowShide rate- n"> _hide rate- n"> _hide rate- n"> _hide rate- n"> _hide rate- 1"> _hide rate- 1"> _hide rows- 1">	title-L-min-win P_Ctrl_2_2_2 se content-cash " content-sight " pan="2"> 期匯率	dth-class noscript" ret-title-L-min-width-colspan="2") 現金匯率 " colspan="2") 即期匯率 span> Exper	
	資料. 看下圖網頁 □ 牌價嚴新掛牌時間: 2019/05/24 14:	分析・ 現金庫 本行買入 31.115 3.851 38.78 313 314 415 316 317	286 287 288 289 270 271 272 273 274 275 274 275 31.805 4.067 40.9	### class="phone media chickens chicke	um-font") _width setwidth setwidth rows _hide rate_ n"> _hide rate_ n"> _hide rate_ n"> _inder now ""> /a 31.585 4.047 40.2	title-L-min-win-win-win-win-win-y-Ctrl_2_2_2 secontent-cash "content-sight "pan="2"> 期匯率(span> 透期匯率 查詢 查詢 查詢 phone-small-finance	dth-class moscript" ret-title-L-min-width-colspan="2") 想应证率 " colspan="2") 期期匯率 span> Exp在率 查詢 查詢 查詢 查詢 查詢 查詢	
	資料. 看下圖網頁 □ 牌價嚴新掛牌時間: 2019/05/24 14:	分析. 現金匯 本行買入 31.115 3.851 38.78 313 314 314 316	286 287 288 289 299 270 271 272 273 274 275 274 275 31.805 4.067 40.9 head2 ody> 4td dat	*************************************	um-font")width setwidth rowst _hide ratehide ratehide rateh'> _hide rateh'> _h'> _hide rateh'> _h'> <td>title-L-min-win P_Ctrl_2_2_2 secontent-cash " content-sight " pan="2"> 明蓮率(/span><!--/i--> 臺詢 臺詢 臺詢 whith phone-small-fi-div"></td> <td>dth-class noscript" ret-title-L-min-width-colspan="2">期期匯率 colspan="2">期期匯率 colspan="2">即期匯率 colspan="2">更重的 查询 查询</td>	title-L-min-win P_Ctrl_2_2_2 secontent-cash " content-sight " pan="2"> 明蓮率(/span> /i 臺詢 臺詢 臺詢 whith phone-small-fi-div">	dth-class noscript" ret-title-L-min-width-colspan="2">期期匯率 colspan="2">期期匯率 colspan="2">即期匯率 colspan="2">更重的 查询	
	資料. 看下圖網頁 □ 牌價嚴新掛牌時間: 2019/05/24 14:	分析. 現金種 本行資人	286 287 288 289 299 270 271 272 273 274 275 274 275 31.805 4.067 40.9 head2 ody> 4td dat	### class="phone-media characteristics" ### characteristics ### characteristics ### characteristics ### characteristics ### characteristics #### characteristics #### characteristics #### characteristics ##################################	um-font") _width setwidth ovetswidth ovetshide rate_ n"> _hide rate_ n"> _hide rate_ n"> _inder note_ n"> _inder note_ all	title-L-min-win P_Ctrl_2_2_2 secontent-cash " content-sight " pan="2"> 期匯率(/span> /i 透期匯率 查詢 查詢 查詢 本語 phone-small-f-ddiv"> "常別國旗" spc:"	dth-class moscript" ret-title-L-min-width-colspan="2">即期匯率 colspan="2">即期匯率 colspan="2">即期匯率 colspan="2">即期匯率 colspan="2">即期匯率 colspan="2">原史僅率 查詢 查詢 查詢 查詢 查詢 = min-cont"> 查詢 查詢 = min-cont"> 查詢 查詢 = min-cont"> 查詢 查詢	
	資料. 看下圖網頁 □ 牌價嚴新掛牌時間: 2019/05/24 14:	分析・ 現金種 本行買入 31.115 3.851 38.78 31.115 3.851 31.115 3.851 31.115 3.851 31.115 3.851 31.115 3.851 31.115 3.851 31.115 3.851 31.115 3.851 31.115 3.851	286 287 288 289 299 270 271 272 273 274 275 274 275 31.805 4.067 40.9 head2 ody> 4td dat	### class="phone-media characteristics" ### characteristics #### characteristics #### characteristics #### characteristics #### characteristics ###################################	um-font") _width setwidth rows _hide rate- ""> _hide rate- ""> _hide rate- ""> _hide rate- "> _hide rate* ""> _hide rate* ""> _hide rate* "">> _hide rate* ""> _hide rate* "" ""> _hide rate* "" _hide rate* " _hide ra	title-L-min-win P_Ctrl_2_2_2 secontent-cash " content-sight " pan="2"> 明確率 "常別國旗" spc-t_hide" />	dth-class noscript" ret-title-L-min-width-colspan="2">期期匯率 colspan="2">期期匯率 colspan="2">即期匯率 colspan="2">更重的 查询	
	資料. 看下圖網頁 □ 牌價嚴新掛牌時間: 2019/05/24 14:	分析・ 現金匯 本行買入 31.115 3.851 38.78 313 314 315 316 317 319 320 321 322 323 324	286 287 288 289 299 270 271 272 273 274 275 274 275 31.805 4.067 40.9 head2 ody> 4td dat	### class="phone-medical color of the class="print of the class="hidde of the class="specific of the class="specific of the class="specific of the class="visible o	um-font")width setwidth volve _hide rate- n"> _hide rate- n"> _hide rate- n"> _hide rate- n"> _inder now- ""> 31.585 4.047 40.2 21.05 "currency sp-america spinging alt- phone print- phone print- phone print-	title-L-min-win-win-win-win-win-win-win-win-win-w	dth-class noscript" net-title-L-min-vidth-colspan="2">期期匯率 colspan="2">期期匯率 colspan="2">期期匯率 colspan="2">更更重要	
	資料. 看下圖網頁 □ 牌價嚴新掛牌時間: 2019/05/24 14:	分析・ 現金匯 本行買人 31.115 38.78 313 344 315 317 318 319 320 321 322 323 324 325	286 287 288 289 299 270 271 272 273 274 275 274 275 31.805 4.067 40.9 head2 ody> 4td dat	### ### #############################	um-font")width setwidth volve _hide rate- n"> _hide rate- n"> _hide rate- n"> _hide rate- n"> _inder now- ""> 31.585 4.047 40.2 21.05 "currency sp-america spinging alt- phone print- phone print- phone print-	title-L-min-win-win-win-win-win-win-win-win-win-w	dth-class noscript" net-title-L-min-vidth-colspan="2">期期匯率 colspan="2">期期匯率 colspan="2">期期匯率 colspan="2">更更重要	
	資料. 看下圖紙頁 □ 牌價嚴新掛牌時間: 2019/05/24 14:	分析. 現金種 本行資人	286 287 288 289 299 270 271 272 273 274 275 274 275 31.805 4.067 40.9 head2 ody> 4td dat	### Class="phone-medical control cont	um-font")width setwidth volve _hide rate- n"> _hide rate- n"> _hide rate- n"> _hide rate- n"> _inder now- ""> 31.585 4.047 40.2 21.05 "currency sp-america spinging alt- phone print- phone print- phone print-	title-L-min-win-win-win-win-win-win-win-win-win-w	### class moscript" ret-title-L-min-width-colspan="2"> ### colspan="2"> ### colspan="2" ### colspan=	
	資料. 看下圖紙頁 □ 牌價嚴新掛牌時間: 2019/05/24 14:	分析・ 現金種 本行買人 31.115 3.851 38.78 313 314 《大 文 文 大 大 の 313 315 316 317 318 319 320 321 322 323 324 325 327	286 287 288 289 270 271 272 273 274 275 31.805 4.067 40.9 head	### Class="phone-medical control cont	um-font")width setwidth volve _hide rate- n"> _hide rate- n"> _hide rate- n"> _hide rate- n"> _inder now- ""> 31.585 4.047 40.2 21.05 "currency sp-america spinging alt- phone print- phone print- phone print-	title-L-min-win-win-win-win-win-win-win-win-win-w	### class moscript" ret-title-L-min-width-colspan="2"> ### colspan="2"> ### colspan="2" ### colspan=	
	資料. 看下圖紙頁 □ 牌價最新掛牌時間: 2019/05/24 14:	分析・ 現金厘3 本行買入 31.115 3.851 38.78 313 314 316 317 319 320 321 322 323 324 325 326 327 328	286 287 288 299 270 271 772 273 274 275 31.805 4.067 40.9 head〉 〈tr〉	### Class="phone-medical control cont	um-font")width setwidth volve _hide rate- n"> _hide rate- n"> _hide rate- n"> _hide rate- n"> _inder now- ""> 31.585 4.047 40.2 21.05 "currency sp-america spinging alt- phone print- phone print- phone print-	title-L-min-win-win-win-win-win-win-win-win-win-w	### class moscript" ret-title-L-min-width-colspan="2">現金匯率 ### colspan="2">即期匯率 ### colspan="2">即期匯率 ### colspan="2">即期匯率 ### colspan="2"> ### colspan	
	資料. 看下圖紙頁 □ 牌價最新掛牌時間: 2019/05/24 14:	分析・ 現金種 本行資人 31.115 3.851 38.78 (大b) 313 314 (大b) 319 320 321 322 323 324 325 326 327 328 329	286 287 288 289 270 271 271 272 273 274 275 274 275 31.805 4.067 40.9 40.9 40.9 40.9 40.9 40.9 40.9 40.9	### Company of the c	wm-font") _width setwidth resthide ratenide raten'> _hide raten''> _hide raten'''''''''''''''''''''''''''''''''''	title-L-min-win P_Ctrl_2_2_2 secontent-cash " content-sight " pan="2"> 開議率(/span> 清別匯率 重詢 重詢 重詢 本語 phone-small-fi -div"> "帝別國族" snc: t_hide" /> t_hide"> t_show" style=	### class moscript" ret-title-L-min-width-colspan="2">現金匯率 ### colspan="2">即期匯率 ### colspan="2">即期匯率 ### colspan="2"> ### colspan="2"	
		分析・ 現金種 本行買人 31.115 3.851 38.78 313 344 《tb 319 320 321 322 323 324 325 326 327 328 329 33.449	286 287 288 288 270 270 271 271 272 273 274 275 274 275 31.805 4.067 40.9 40.9 40.9 40.9 40.9 40.9 40.9 40.9	### Company of the c	wm-font") _width restwidth resthide ratem'> _hide ratem'> _hide ratem'> _hide ratem'> _hide ratewide rate	title-L-min-win P_Ctrl_2_2_2 secontent-cash " content-sight " pan="2"> 附張率(/span>	### class moscript	
	資料. 看下圖紙頁 跨別	大子 : 現金種 : 本行資人	286 287 288 288 289 270 271 271 272 273 274 275 274 275 274 275 4.067 40.9 40.9 40.9 40.9 40.9 40.9 40.9 40.9	### Company of the c	wm-font") _width restwidth resthide ratem'> _hide ratem'> _hide ratem'> _hide ratem'> _hide ratewide rate	title-L-min-win P_Ctrl_2_2_2 secontent-cash " content-sight " pan="2"> 明確率(/span> 清明確率 查詢 查詢	### class moscript" ret-title-L-min-width-colspan="2"> ### colspan="2"> ### colspan="2" ### colspan=	

Cocading/auth www.941coding.com

列	程式	說明
14	data1 = sp.find('tbody').find_all('tr')	網頁原始碼再深入分析,可觀察到每一筆幣別匯率都在一 和匯率都在一 標籤內,因此 1. 使用 BeautifulSoup 的物件 sp 的 find()方法; sp.find('tbody')找到表格內文標籤 2. 用 find_all()方法, sp.find('tbody').find_all('tr')再找到 內的所有 標籤,每一 等的有一筆幣別匯率資料 3. 將上一步驟找到由 规的串列,指定給變數 data1
15	print(len(data1))	輸出串列的長度 19; 即網頁有 19 種幣別匯率
16	print(data1)	輸出串列 data1,繼續分析內容 html 結構
	看下圖是由第 16 列輸出 data1 的結果, 紅色箭頭所 指為第一筆資料, 即美金匯率資料; 而每一筆内有 11 筆標籤	而我們要抓取的資料「美金 (USD)」是在第一筆, class="visible-phone print_hide" 的 <div>標籤內 資料「31.085」是在第二筆標籤內</div>

(ch1-4 BeautifulSoup2.py)



將 ch1-4_BeautifulSoup1.py 另存為 ch1-4_BeautifulSoup2.py 繼續用BeautifulSoup的物件 sp解析網頁, 取得目標資料

```
8 import requests
                                                                          美金 (USD) :31.085
9 from bs4 import BeautifulSoup
                                                                          港幣 (HKD) :3.848
10 url = 'https://rate.bot.com.tw/xrt?Lang=zh-TW'
                                                                          英鎊 (GBP) :38.74
11 html = requests.get(url)
                                                                         澳幣 (AUD):21.36
12 html.encoding = 'utf-8'
                                                                         加拿大幣 (CAD):22.93
13 sp = BeautifulSoup(html.text, 'html.parser')
                                                                          新加坡幣 (SGD):22.31
14 data1 = sp.find('tbody').find_all('tr')
15 #print(len(data1))
                                                                         瑞士法郎 (CHF) :30.62
16 #print(data1)
                                                                          日圓 (JPY):0.2782
17 for data2 in data1:
                                                                         南非幣 (ZAR) :-
      data3 = data2.find_all('td')
18
      usd = data3[0].find('div', {'class':'visible-phone print_hide'})
                                                                         瑞典幣 (SEK) :2.9
     usd_text = usd.text.strip()
print(usd_text, ':', end='')
print(data3[1].text)
                                                                         紐元 (NZD) :20.1
21
                                                                          泰幣 (THB):0.8652
                                                                         菲國比索 (PHP):0.5299
                                                                         印尼幣 (IDR):0.00188
                                                                         歐元 (EUR) :34.44
                                                                         韓元 (KRW) :0.0248
                                                                          越南盾 (VND):0.00098
                                                                         馬來幣 (MYR) :6.429
                                                                          人民幣 (CNY):4.457
```

列	程式	說明
15,	#print(len(data1))	將 15, 16 列改為註解
16	#print(data1)	原輸出只是要分析已抓取的資料,並非目標 資料
17	for data2 in data1:	以 for 回圈的變數 data2 迭代 data1 串列中的元素, data2 即為 data1 串列中的每一筆
18	data3 = data2.find_all('td')	以之前的分析,每一筆標籤中有 11 筆 標籤,以 data2.find_all('td')找出所有 標籤,傳回的串列指派給變數 data3
19	<pre>usd = data3[0].find('div', {'class':'visible-phone print_hide'})</pre>	之前已分析過要抓取的幣別資料「美金(USD)」是在第一筆(索引值為 0, 即data3[0]), class="visible-phone print_hide"的 <div>標籤內;以(data3[0].find('div', {'class':'visible-phone print_hide'})的語法表示,找到 class為 visible-phone print_hide 的div. 其中注意{'class':'visible-phone print_hide'}語法說明如下: BeautifulSoup物件方法(find or find_all)(tag, { 屬性名稱:屬性內容 }),可取得標籤 tag中符合的屬性. 傳回的字串指定給變數 usd</div>

Cocading-Youth www.941coding.com

20	<pre>usd_text = usd.text.strip()</pre>	再次觀察原始碼如下圖第 323 列要抓取的 資料「美金 (USD)」,有換行或空白等字符, 因此需去除這些多餘的字符,才不會影響輸 出格式. usd.text 取得字串,用 strip()方法去 除字串左右多餘的字符,並將結果指定給變 數 usd_text 代表各幣別
21	print(usd_text, ':', end='')	輸出變數 usd_text 幣別,加上冒號「:」且不換行(end=")
22	print(data3[1].text)	之前已分析過要抓取的匯率資料「31.085」 是在第二筆標籤內(索引值為 1, 即 data3[1]), data3[1].text 取得資料輸出

```
314
                                                                                                                                       315
316
                                                                                                                                                                    ·/

〈div class="sp-div sp-america-div"〉

〈img title="幣別國旗" alt="幣別國旗" src="<u>/Content/images/spr</u>
  318
319
    320
                                                                                                                                                                                                                               </div>
                                                                                                                                                                                                                            <br/>
class="visible-phone print_hide" /><br/>
<br/>
<b
    321
                                                                                                                                                                                                                                                        業金 (USD) 從呈現內容可知有換行和空白
 325
326
                                                                                                                                                                                                                            <div class="hidden-phone print_show" style="text-indent:30px;">
                                                                                                                                                                                                                                                      美金 (USD)
    327
                                                                                                                                                                                                                            </div>
                                                                                                                                                                                                </div>
```

哇! 我們輕鬆就從台灣銀行網頁抓取需要的 19 種幣別的匯率,是不是很酷? 接下來我們要玩更酷的實作喔!

註:每個網頁的 HTML 結構都不同,因此要有 HTML 網頁標籤語言基礎知識,在爬取網頁資料時才能應付不同的網頁結構.而學員孩子在學習 Python程式語言之前,也已學習網站設計.對於爬取網頁資料就游刃有餘了!因此對 HTML 網頁標籤語言和 CSS 階層樣式表不熟悉的同學,建議先上 w3c 網站熟悉喔!