



國立臺灣大學 National Taiwan University

12-擷取網頁資料

陳永樵





12 擷取網頁資料

- 使用Python 所提供模組可以直接擷取網頁中的資料，介紹如何取得網頁



12-1 模組urllib.request、urllib.response 與requests

- 模組urllib.request 用於下載指定網址的網頁，程式碼較為複雜，第三方函式庫requests，也可以下載指定網址的網頁，程式碼較為簡潔。



模組urllib.request的重要函式

重要函式與說明	程式碼與執行結果
<p><code>urlopen(url)</code></p> <p>開啟 url 所指定的網頁，並回傳 urllib.response 物件。</p>	<pre>import urllib.request as ur url='http://www.python.org' resp=ur.urlopen(url) print(resp)</pre>
	<p><http.client.HTTPResponse object at 0x02A92570></p>



模組urllib.response 的重要函式與屬性

重要函式與說明

read()

讀取物件 urllib.response 的所有資料，資料皆為 byte，需要使用函式 decode 轉換成字串。

程式碼與執行結果

```
import urllib.request as ur
url='https://www.python.org'
resp=ur.urlopen(url)
data=resp.read()
print(data)
```

```
b'<!doctype html>\n<!--[if lt IE 7]> <html
class="no-js ie6 lt-ie7 lt-ie8 lt-ie9">...
<![endif]-->\n\n  \n\n  \n \n\n</body>\n
n</html>\n'
```



重要函式與說明

geturl()

讀取物件 urllib.response 的網址。

程式碼與執行結果

```
import urllib.request as ur
url='https://www.python.org'
resp=ur.urlopen(url)
print(resp.geturl())
```

https://www.python.org

getheader()

讀取物件 urllib.response 的網頁表頭。

```
import urllib.request as ur
url='https://www.python.org'
resp=ur.urlopen(url)
print(resp.getheaders())
```

```
[('Server', 'nginx'), ('Content-Type',
'text/html; charset=utf-8'), ..., 'max-
age=63072000; includeSubDomains')]
```



重要函式與說明

status

伺服器回傳的常見狀態碼，如下。

狀態碼	表示
2xx	成功獲得資料，例如：200 表示「OK」。
4xx	用戶端錯誤，例如：404 表示「找不到」
5xx	伺服器錯誤，例如：502 表示「閘道故障」

程式碼與執行結果

```
import urllib.request as ur
url='https://www.python.org'
resp=ur.urlopen(url)
print(resp.status)
```

200



12-1 模組urllib.request、urllib.response與requests

- 模組requests 為第三方函式庫，需要使用pip 進行安裝，若是Windows 作業系統
- 在「命令提示字元」程式下，使用指令「pip install requests」，會自動從網路下載安裝模組requests，模組requests 的重要函式如下。



模組requests 的重要函式

重要函式與說明

`requests.get(url)`

開啟 url 所指定的網頁。

程式碼與執行結果

```
import requests  
url = 'http://www.python.org'  
data = requests.get(url)  
print(data)
```

顯示整個網頁的原始碼，因為過長而省略。



12-1-1 使用模組urllib.request 下載網頁

- (ch12\12-1-1- 模組urllib.request.py)
- 使用模組urllib.request，下載網址為「<https://www.python.org>」的網頁。

行號	範例 ( : ch12\12-1-1- 模組 urllib.request.py)
1	import urllib.request as ur
2	url='https://www.python.org'
3	resp=ur.urlopen(url)
4	print(resp.geturl())
5	print(resp.status)
6	print(resp.getheaders())
7	data=resp.read()
8	print(data)
9	print(data.decode())



程式解說

- 第1 行：匯入模組`urllib.request`，重新命名為`ur`。
- 第2 行：設定`url` 為「`https://www.python.org`」。
- 第3行：顯示模組`ur`的函式`urlopen`，以`url`為輸入，將回傳的`urllib.response`物件，指定物件名稱為`resp`。



- 第4 行：使用函式print 顯示物件resp的函式geturl(網址) 到螢幕上。
- 第5 行：使用函式print 顯示物件resp的屬性status(網頁狀態) 到螢幕上。



- 第6 行：使用函式print 顯示物件resp的函式getheaders(網頁表頭) 到螢幕上。
- 第7 行：顯示物件resp 的函式read，將回傳的網頁資料物件，指定物件名稱為data。



- 第8 行：使用函式print 將物件data的網頁原始碼，以byte方式顯示到螢幕上。
- 第9 行：使用函式decode 將物件data 的網頁原始碼byte 資料轉換成字串，最後使用函式print 將轉換後的字串顯示到螢幕上。



執行結果

```
https://www.python.org
```

```
200
```

```
[('Server', 'nginx'), ('Content-Type', 'text/html; charset=utf-8'), ..., 'max-age=63072000; includeSubDomains']
```

```
b'<!doctype html>\n<!--[if lt IE 7]> <html class="no-js ie6 lt-ie7 lt-ie8 lt-ie9">...資料過長  
省略...<![endif]-->\n\n \n\n \n \n\n</body>\n</html>\n'
```

```
<!doctype html>
```

```
<!--[if lt IE 7]> <html class="no-js ie6 lt-ie7 lt-ie8 lt-ie9"> <![endif]-->
```

```
...資料過長省略...
```


```
</body>
```

```
</html>
```



12-1-2 使用函式庫requests 下載網頁

- (ch12\12-1-2-requests.py)
- 使用第三方函式庫requests，下載網址為「<https://www.python.org>」的網頁。

行號	範例 ( : ch12\12-1-2-requests.py)
1	<code>import requests</code>
2	<code>url = 'http://www.python.org'</code>
3	<code>data = requests.get(url)</code>
4	<code>print(data.encoding)</code>
5	<code>print(data.status_code)</code>
6	<code>print(data.headers)</code>
7	<code>print(data.text)</code>



程式解說

- 第1 行：匯入模組requests。
- 第2 行：設定url 為「<https://www.python.org>」。
- 第3 行：顯示模組requests的函式get，以url 為輸入，將回傳物件命名為data。
- 第4 行：使用函式print 顯示物件data的屬性encoding(網頁編碼) 到螢幕上。



- 第5 行：使用函式print 顯示物件data的屬性status_code(網頁狀態) 到螢幕上。
- 第6 行：使用函式print 顯示物件data的屬性headers(網頁表頭) 到螢幕上。
- 第7 行：使用函式print 顯示物件data的屬性text(網頁內容) 到螢幕上。



12-1-2 使用函式庫requests 下載網頁 (ch12\12-1-2-requests.py)

執行結果

```
utf-8
200
{'Connection': 'keep-alive', 'Public-Key-Pins': 'max-age=600;...', 'Date': 'Tue, 21 Jun 2016
06:01:20 GMT', 'Content-Length': '47462'}
<!doctype html>
<!--[if lt IE 7]> <html class="no-js ie6 lt-ie7 lt-ie8 lt-ie9"> <![endif]-->
…資料過長省略…
</body>
</html>
```