

Python-爬蟲程式設計

蘇柏原(teaching@bo-yuan.net)

網頁的呼叫

網頁的呼叫

- 呼叫方式：
 - HEAD
 - 只取傳回來的標頭資料
 - GET
 - 取得傳回來的全部資料
 - POST
 - 新增資料，同時也取得返回結果
 - PUT
 - 更新資料，同時也取得返回結果
 - DELETE
 - 刪除資料，同時也取得返回結果
 - PATCH
 - 補充資料，同時也取得返回結果
- 不論哪個連線，都是去呼叫網址傳遞資訊，所以實際做甚麼是由程式設計師決定！

網頁的呼叫

- 常見的資料類型：
 - HTML
 - 一般網頁的格式
 - JSON
 - 一種簡易的資料表示格式，是樹狀結構
 - XML
 - 一種資料表示格式，樣子與網頁語法類似
 - CSV
 - 一種資料表示格式，以逗點區隔資料類型，一行存取一筆資料
 - TEXT
 - 沒有特別格式的文字資料
 - BINARY
 - 圖片、影片或檔案等等

Requests函式庫

用來抓取指定的網站資料

安裝Requests

- Requests是一個Python函式庫，他可以用來抓取指定的網址內容，並且也可以模擬瀏覽器傳遞資訊給網址，他需要透過PIP安裝：

```
pip install requests
```

- 在欲使用的程式中載入Requests函式庫：

```
import requests
```

使用Requests

head 、 get 、
post 、 put 、
delete 、 patch

- 讀取指定網址：

操作變數=requests.呼叫方法(
網址,
headers=標頭資訊,
params=傳遞的網址參數資料,
data=傳遞的資料,
json=以JSON格式傳遞的資料
)

如果同時使用則
會以data的資料
為主

如果遇到SSL憑證過期，需要加這個參
數來關閉驗證才能抓取資料：

verify=False

使用Requests

- 讀取的狀態：
操作變數.status_code
- 讀取的標頭資訊(字典類型)：
操作變數.headers
- 讀取和設定編碼：
操作變數.encoding

使用Requests

- 讀取內容(以純文字型態讀取，會自動做編碼的轉換)：
操作變數.text
- 讀取內容(原始型態，可讀取二進位資料)：
操作變數.content
- 讀取內容(JSON資料)：
操作變數.json()

作業

- 作業：
 - 使用requests指令連線至網址：
<http://teaching.bo-yuan.net/test/requests/>
 - 運用各種呼叫方式(head, get, post, put, delete, patch)找尋答案。
 - 完成後把操作過程中的所有指令保留在程式碼中，將檔案繳交上來。

CSV資料轉換

JSON 使用與轉換

- 在欲使用的程式中載入CSV函式庫：
 - `import csv`

- 載入CSV檔案串流：

`結果串列=list(csv.reader(檔案串流))`

傳入`codecs.open`函式開啟的變數



如果想要直接解析字串，可以用
「`io.StringIO(字串)`」函式將字串轉
成檔案串流



JSON資料轉換

JSON 使用與轉換

- 在欲使用的程式中載入JSON函式庫：
 - `import json`
- 載入JSON文字：
`結果變數=json.loads(要載入的文字)`
- 解析物件成JSON字串：
`結果字串=json.dumps(要解析的變數)`

作業

- 作業：
 - 為「PCHome線上購物」建立一個的商品搜尋程式。
 - 網址：<https://shopping.pchome.com.tw/>
 - 繳交程式碼檔案。

BeautifulSoup 函式庫

用來分析HTML或XML

安裝 BeautifulSoup

- BeautifulSoup是一個Python函式庫，它可以用來分析HTML或XML格式的資料，需要透過PIP安裝：
`pip install beautifulsoup4`
- 在欲使用的程式中載入BeautifulSoup函式庫：
`from bs4 import BeautifulSoup`

使用 BeautifulSoup

- 分析指令的資料：
操作變數=BeautifulSoup(要分析的資料, 資料格式)

html.parser => 分析HTML
xml => 分析XML(pip install lxml)



使用 BeautifulSoup

- 標籤變數也可以使用這兩個函式

- 取得指定的**第一個**標籤(返回指向標籤的變數)：
標籤變數 = 操作變數.find(標籤名稱, 屬性名稱, text= 包覆的文字)
- 取得指定的**全部**標籤(返回指向全部標籤的串列變數)：
標籤串列變數 = 操作變數.find_all(標籤名稱, 屬性名稱, text= 包覆的文字)

如果不設可以填
None

字典格式：

```
{  
    屬性名: 屬性值,  
    屬性名: 屬性值  
}
```

使用 BeautifulSoup

- 取得標籤名稱：
標籤變數.name
- 取得標籤包覆的內容：
標籤變數.text
- 取得標籤中全部的屬性：
標籤變數.attrs
- 取得當前標籤的上層標籤(包覆當前標籤的標籤)：
標籤變數.parent

作業

- 作業：
 - 使用「**中央氣象局**」的氣象資料，建立一個可以查詢縣市溫度極值資料的程式。
 - 網址：https://www.cwb.gov.tw/V8/C/W/County_TempTop.html
 - 繳交程式碼檔案。