

Python 程式設計

林奇賦 daky1983@gmail.com

Outline

- ▶ 網頁抓取與解析
- ▶ urllib
- ▶ HTMLParser

urllib

- ▶ **urllib**這個**module**，提供一般抓取網頁的工作，可以使用**urlopen**函數開啟某個網址，然後將傳回的物件呼叫它的**read**函數，取出所有網頁的內容，最後關閉。原本可能會很複雜的工作全部都已經被包好了

urllib

- ▶ `urlopen()`, 是基於python的`open()`方法
- ▶ `urllib.request.urlopen('網址')`
- ▶ 傳入參數要遵循http、ftp、等網路協議
 - ▶ `urllib.request.urlopen('http://www.yahoo.com.tw')`
 - ▶ 特別注意，協定方式一定要加
- ▶ 也可以是本機端的檔案
 - ▶ `urllib.request.urlopen('file:c:\\python34\\檔名.副檔名')`

讀取網頁內容

- ▶ 使用`read()`方法會將所有內容以`bytes`型態讀取出來
- ▶ `bytes`型態可透過呼叫`decode()`方法來設定編碼，並轉成字串型態回傳
 - ▶ `response = urllib.request.urlopen('http://invoice.etax.nat.gov.tw/')`
 - ▶ `response.read().decode('utf_8')`
- ▶ 其中 `read()` 中可以傳入參數，例如`read(10)`則會回傳長度10的字串
- ▶ 範例 EX09_01.py

HTMLParser

- ▶ 是HTML的解析器，不是嚴謹地去解析網頁，它可以處理像不對稱的HTML語法等等，對於網路上各種千奇百怪出錯的網頁來說，當然是選擇可以容錯的 Parser 比較好
- ▶ 其運作方式是這樣，使用者覆載(override)一系列的 **handle_xxx** 函數，例如handle_data就是負責處理非HTML標籤，也就是不在<>的那些字用的方法，當它分析到這樣的資料就會呼叫handle_data，所以覆載了這個函數就可以處理這些資料，如果你希望可以處理 HTML標籤，你也可以覆載handle_startag等等方法
- ▶ 其中xxx表示html tag的類型

HTMLParser

- ▶ `from html.parser import HTMLParser`
- ▶ 透過繼承的機制繼承 `HTMLParser` 類別
- ▶ 定義我們自己的網頁原始碼的解析器類別
- ▶ 依需求覆載(`override`)一系列的 `handle_xxx` 函數，並實作函式的內容
- ▶ 使用自行定義的類別產生出解析器物件實體
- ▶ 透過呼叫`feed()`方法將傳入的參數進行語法分析

可覆載(override)的函數

- ▶ `HTMLParserhandle_starttag(tag, attrs)`
- ▶ `HTMLParserhandle_endtag(tag)`
- ▶ `HTMLParserhandle_startendtag(tag, attrs)`
- ▶ `HTMLParserhandle_data(data)`
- ▶ `HTMLParserhandle_entityref(name)`
- ▶ `HTMLParserhandle_charref(name)`
- ▶ `HTMLParserhandle_comment(data)`
- ▶ `HTMLParserhandle_decl(decl)`
- ▶ `HTMLParserhandle_pi(data)`
- ▶ `HTMLParserunknown_decl(data)`

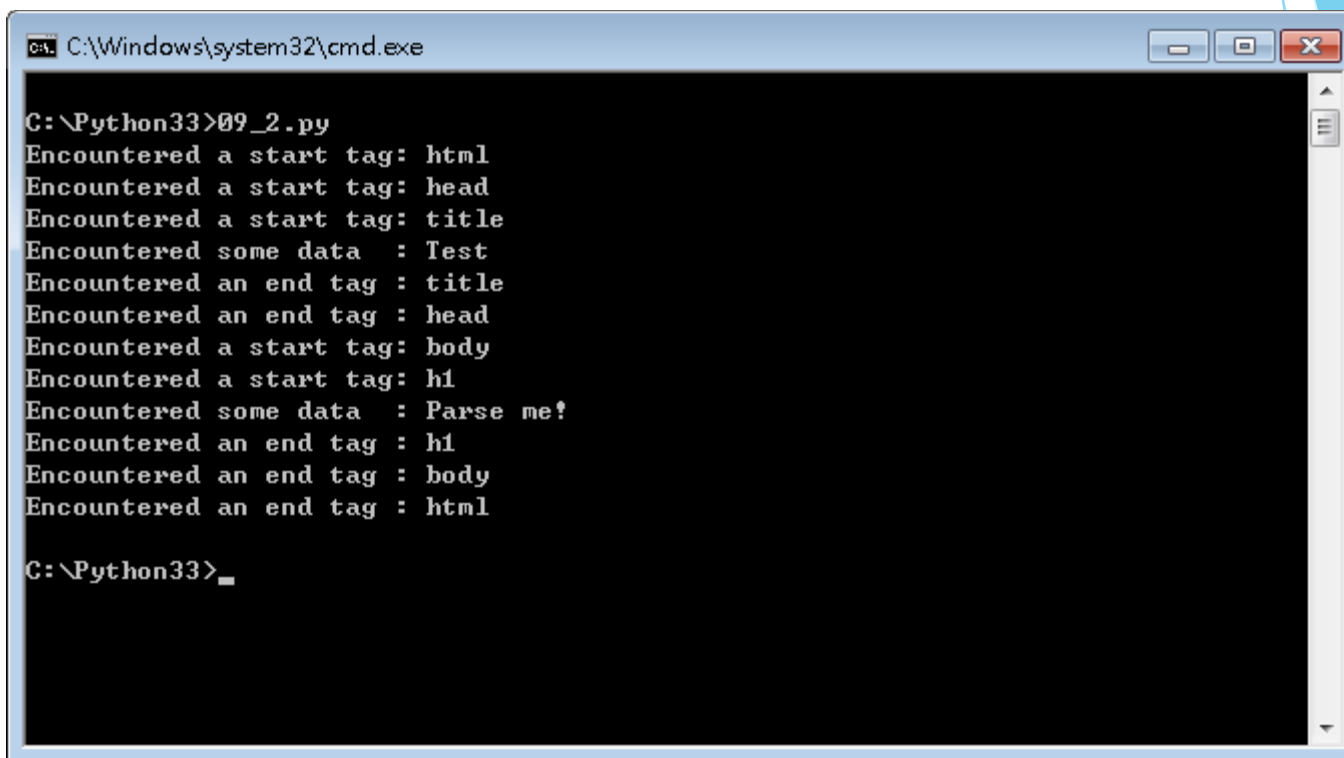
Html Tag 類型

- ▶ 以下針對常見的 tag 類型做說明:
- ▶ starttag
 - ▶ 無屬性(attrs)的如: `<head>`
 - ▶ 有包含屬性的: ``
 - ▶ 其中屬性會以 `[("class", "t18Red")]` 形式存放內容
- ▶ endtag
 - ▶ `</head>`
- ▶ startendtag
 - ▶ `<meta http-equiv="Content-Type" content="text/html; charset=utf-8" />`
- ▶ data
 - ▶ 被tag夾住的內容，非任何tag形式稱之為data

範例程式

- ▶ EX09_02.py
- ▶ 務必要理解解析時的流程

輸出結果



```
C:\Windows\system32\cmd.exe

C:\Python33>09_2.py
Encountered a start tag: html
Encountered a start tag: head
Encountered a start tag: title
Encountered some data : Test
Encountered an end tag : title
Encountered an end tag : head
Encountered a start tag: body
Encountered a start tag: h1
Encountered some data : Parse me!
Encountered an end tag : h1
Encountered an end tag : body
Encountered an end tag : html

C:\Python33>_
```

HTMLParser Methods

- HTMLParser 包含以下的方法
 - `HTMLParser.feed(data)`
 - `HTMLParser.close()`
 - `HTMLParser.reset()`
 - `HTMLParser.getpos()`
 - `HTMLParser.get_starttag_text()`

參考資料

- <http://docs.python.org/3.3/library/html.parser.html?highlight=htmlparser#html.parser.HTMLParser.close>

抓取統一發票範例

- ▶ EX09_03.py
- ▶ 問題1, (下面這種、分隔的字串要怎麼分割出來)
 - ▶ '82267055、72762106、06820335'
- ▶ 問題2, 擷取到的內容似乎包含兩期對獎的號碼，要怎麼限制抓取的是某一組 (ex. 最新一期的)

中文URL的編碼/解碼

- ▶ `import urllib.parse`
- ▶ `urllib.parse.quote(str)`
 - ▶ 此方法可將`str`中的字串轉為url編碼
- ▶ `urllib.parse.unquote(str)`
 - ▶ 將url碼解碼

Homework

- ▶ 抓取yahoo!電影的某部電影, 例如:
 - ▶ https://tw.movies.yahoo.com/movieinfo_main.html/id=5644
- ▶ 需要抓取的資訊如下:
 - ▶ 電影名稱 (中英)
 - ▶ 上映日期
 - ▶ 類 型
 - ▶ 片 長
 - ▶ 導 演
 - ▶ 演 員
 - ▶ 發行公司
 - ▶ 官方網站
 - ▶ 劇情介紹
- ▶ 將擷取出來的資料存檔, 檔名: **編號.txt**, 以這部電影為例存檔為**5644.txt**