

Python程式設計#9

林奇賦 daky1983@gmail.com

Outline

- urllib
- Simple HTML and XHTML parser

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
- 如果想送表單、改變header怎麼辦？
 - 可以使用 urllib2

- `urlopen()`,是基於python的`open()`方法的
- `urllib.request.urlopen('網址')`
 - 傳入參數要遵循http、ftp、等網路協議
 - 例如:`urllib.request.urlopen('http://www.yahoo.com.tw')`
 - 也可以是本機端的檔案
 - `urllib.request.urlopen('file:c:\python33\檔名.副檔名')`

讀取網頁內容

- 使用`read()`方法會將所有內容讀取出來
- 可透過呼叫`decode`方法來設定編碼
 - `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

抓取網頁的方法

```
# -*- coding: utf-8 -*-  
import urllib.request  
response = urllib.request.urlopen('http://invoice.etax.nat.gov.tw/')  
html = response.read().decode('utf_8')  
print(html)
```

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

HTMLParser

- `from html.parser import HTMLParser`
- 覆載(override)相對應的函數，格式如下

```
class myparser(HTMLParser):  
    def handle_starttag(self, tag, attrs):  
        print("Encountered a start tag:", tag)  
    def handle_endtag(self, tag, attrs):  
        print("Encountered an end tag :", tag)  
    def handle_data(self, data):  
        print("Encountered some data :", data)
```


HTMLParser

- `Parser = myparser()`
- 透過呼叫`feed()`方法將傳入的參數進行語法分析
- ex.
`Parser.feed('<html><head><title>Test</title></head>'
'<body><h1>Parse me!</h1></body></html>')`

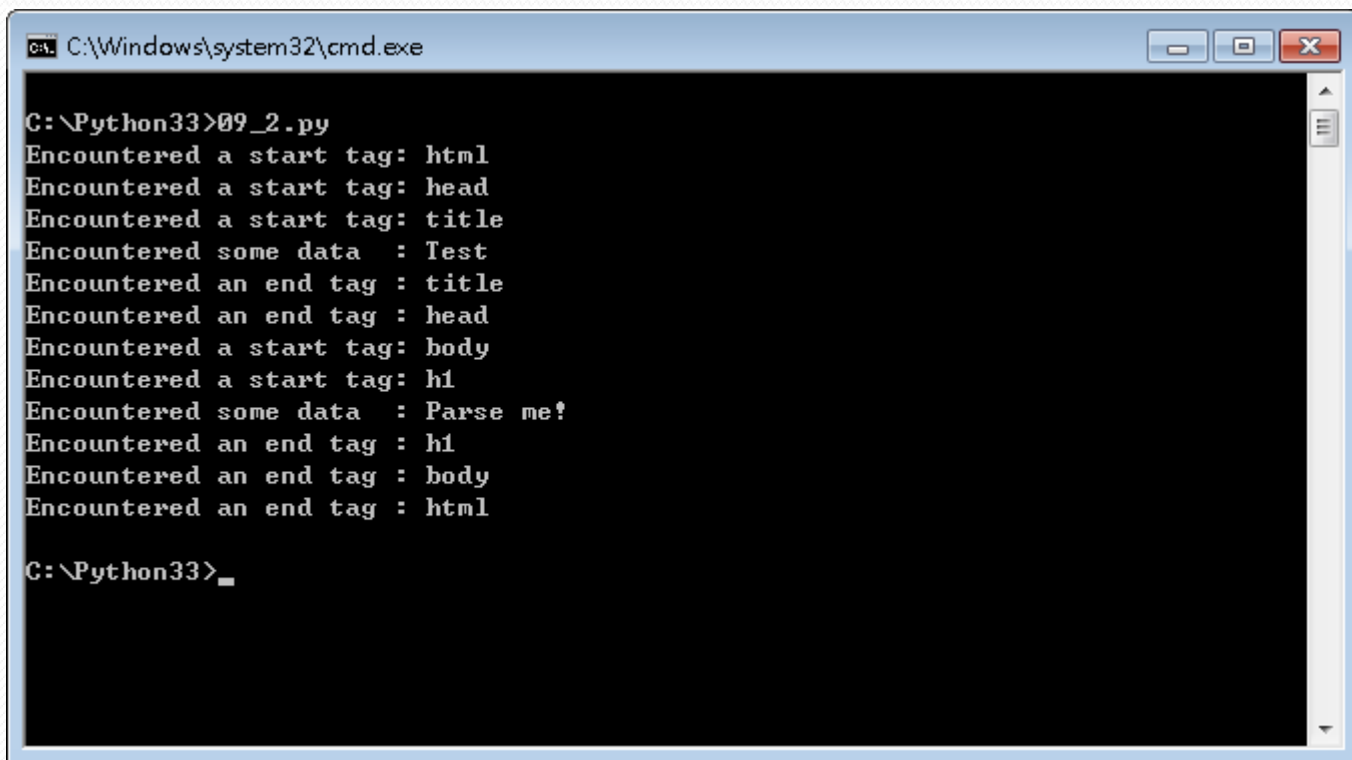
範例程式

```
from html.parser import HTMLParser

class MyHTMLParser(HTMLParser):
    def handle_starttag(self, tag, attrs):
        print("Encountered a start tag:", tag)
    def handle_endtag(self, tag):
        print("Encountered an end tag :", tag)
    def handle_data(self, data):
        print("Encountered some data :", data)

parser = MyHTMLParser()
parser.feed('<html><head><title>Test</title></head>'
          '<body><h1>Parse me!</h1></body></html>')
```

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()`

可覆載(override)的函數

- `HTMLParser.handle_starttag(tag, attrs)`
`HTMLParser.handle_endtag(tag)`
`HTMLParser.handle_startendtag(tag, attrs)`
`HTMLParser.handle_data(data)`
`HTMLParser.handle_entityref(name)`
`HTMLParser.handle_charref(name)`
`HTMLParser.handle_comment(data)`
`HTMLParser.handle_decl(decl)`
`HTMLParser.handle_pi(data)`
`HTMLParser.unknown_decl(data)`

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

```
import urllib.request
from html.parser import HTMLParser
data = urllib.request.urlopen('http://invoice.etax.nat.gov.tw')
content = data.read().decode('utf_8')
data.close()
```

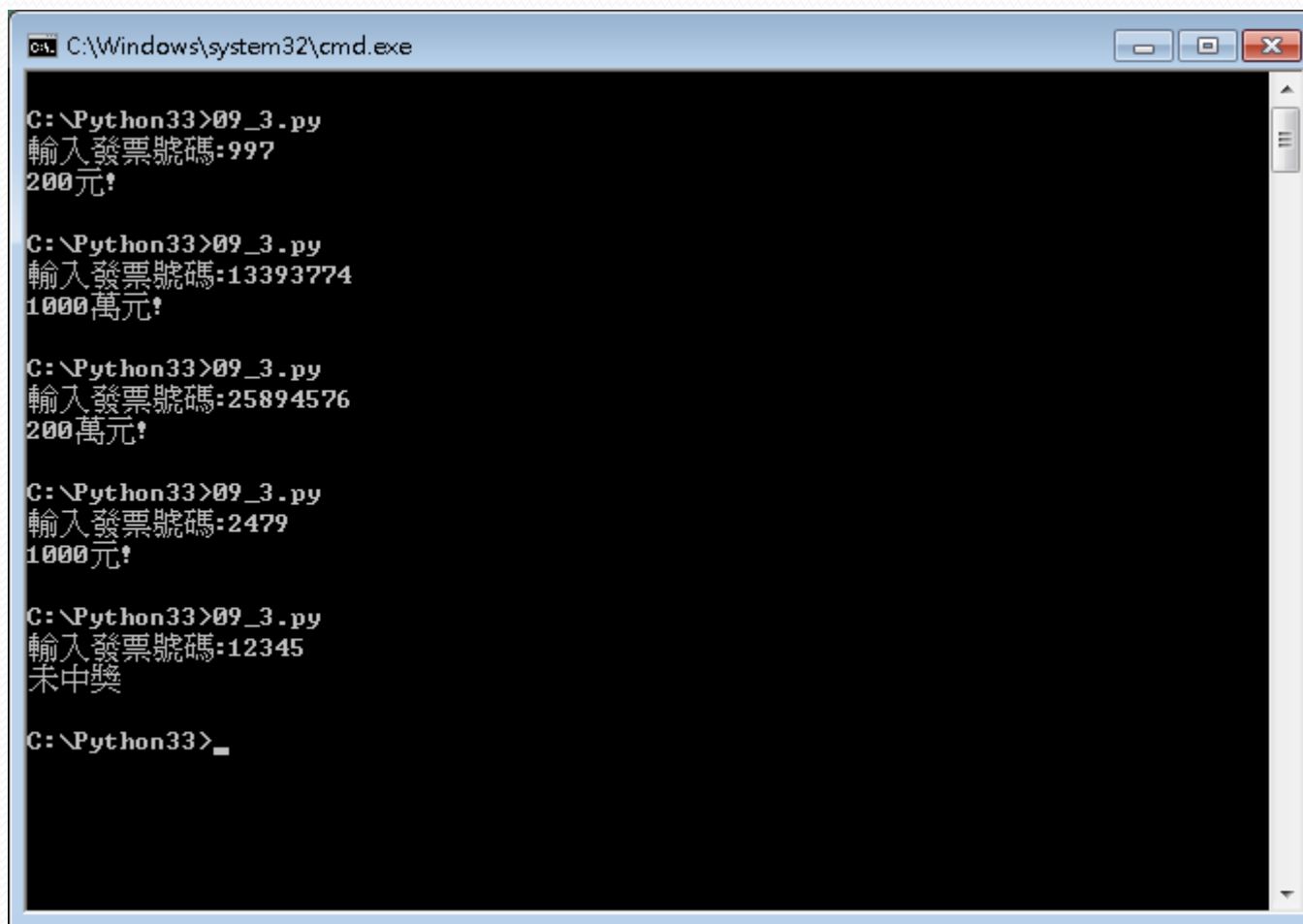
```
class myparser(HTMLParser):
    def __init__(self):
        HTMLParser.__init__(self)
        self.isNumber = 0
        self.numbers = []
    def handle_data(self, data):
        if self.isNumber == 1:
            print(data)
            self.isNumber = 0
    def handle_starttag(self, tag, attrs):
        if tag == 'span' and attrs == [('class', 't18Red')]:
            self.isNumber = 1
```

```
Parser = myparser()
Parser.feed(content)
```


課堂練習

- 試著讓使用者輸入一串數字，並且判斷輸入的數字是否中獎
 - 如果中獎，則顯示 恭喜中了什麼獎，得到xxx元
 - 反之則顯示未對中的訊息

結果示意圖



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

C:\Python33>09_3.py
輸入發票號碼:997
200元!

C:\Python33>09_3.py
輸入發票號碼:13393774
1000萬元!

C:\Python33>09_3.py
輸入發票號碼:25894576
200萬元!

C:\Python33>09_3.py
輸入發票號碼:2479
1000元!

C:\Python33>09_3.py
輸入發票號碼:12345
未中獎

C:\Python33>_
```

中文URL的編碼/解碼

- `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

課堂/作業練習

- 傳入yahoo!電影的某部電影連結
 - http://tw.movies.yahoo.com/movieinfo_main.html/id=4569
- 輸出結果至csv檔案
 - 欄位分別為
 - 電影名稱
 - 上映日期
 - 類型
 - 片長
 - 導演
 - 劇情介紹