**解析複雜的HTML網頁以擷取你所需的資訊**

愈到複雜的標籤時很容易打算以多行程式強行擷取你要的資訊，這是應該避免的。

假設你有一些目標內容，可能是名稱、統計、或一段文字。它可能埋藏在20曾標籤當中而沒有適合的標籤或HTML屬性可供搜尋。總不可能如下例:

bs.find\_all(‘table’)[4].find\_all(‘tr’)[2].find(‘td’).find\_all(‘div’)[1].find(‘a’)

程式不好看，而且若當網站作者更新網站後(像是新增標籤、清單)，程式可能就沒用了。

遇到複雜的標籤德html有甚麼方法呢?

* 找其他較工整的網站代替(如果都有你想要的資料的話)
* 尋找”列印此頁”連結或格式比較好的行動版HTML(將自己偽裝成行動裝置並接收行動版本網頁)
* 尋找隱藏在JavaScript檔案中的資訊。要記得必須檢查匯入的JavaScript檔案才能這麼做
* 對網頁標頭來說，網站本身的URL經常就有此資訊，可以觀察一下。

**BeautifulSoup的屬性搜尋標籤、操作多個標籤與穿越解析樹。**

幾乎每個網站都有儀式表**(CSS)**

**CSS**以HTML元素間的差異讓他們套用不同的樣式。

<http://www.pythonscraping.com/pages/warandpeace.html>(範例網站)

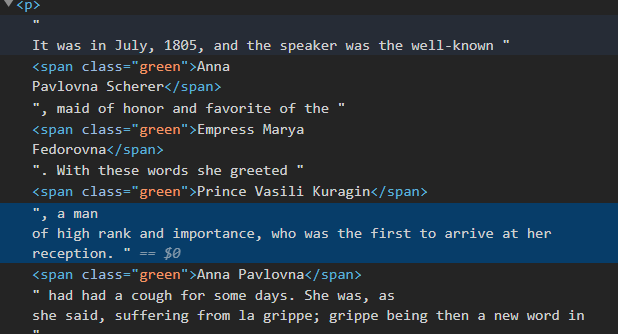
有些標籤可能像這樣:

<span class=”green”></span>

其他像這樣:

<span class=”red”></span>

BeautifulSoup 依靠CSS可抓取有紅色文字而排除綠色文字，或者反之。



**使用get\_text()與保存標籤的時機**

.get\_text()從文件中取出標籤且只回傳文字的Unicode字串。

舉例來說，若操作帶有許多超連結、圖片與其他標籤的一大段文字，這些東西都會被抽離而只剩下無標籤的一段文字。

要記得從BeautifulSoup物件找出東西 比從一段文字中找簡單。

呼叫.get\_text()應該是輸出、儲存或操作最終資料前的最後手段。

一般來說應該要盡可能保存文件的標籤結構。

**BeautifulSoup的find()和find\_all()**

find()和find\_all()是BeautifulSoup最常用的2個函式，可以用他們過濾HTML網頁以根據不同屬性招出所需的全部或單一標籤。

find\_all(tag, attributes, recursive, text, limit, keywords)

find(tag, attributes, recursive, text, keywords)

95%情況只需要前面2個參數。

tag:標籤

attributes:屬性

recursive:布林值

recursive= True 會從子標籤與其子標籤中找尋相符者 (預設)

recursive= False 只找文件的第一層

text:較為罕見，根據標籤的”文字內容”而非標籤屬性搜尋

補充:

allText = bsObj.findAll(id="text") #只放一個參數

allText = bsObj.findAll(' ',{'id':'text'}) #放不只一個參數

#2行是相等的

**其他BeautifulSoup物件**

1. BeautifulSoup物件
2. Tag物件: 呼叫find、find\_all或向下找而取的清單或物件: bs.div.h1
3. NavigableString物件:用於代表標籤中的文字而非標籤本身
4. Comment物件:用於尋找html的註解， <!--例如這個-->

**處裡子節點**

經常在電腦與數學中聽到子節點的操作:移動、儲存”刪除等，但這裡只討論選取。

在BeautifulSoup與其他函式庫中，**子節點(children)**與**後代(descendant)** 不同 : 如同親屬關係，子節點在父節點的下一層，後代在父節點的任一層

舉例來說，tr標籤是table的子節點，而tr、th、td、img、span都是table的後代(參考<http://www.pythonscraping.com/pages/page3.html>來說)。所有子節點都是後代，但並非所有後代都是子節點。

**一般來說，BeautifulSoup函式都處裡目前選取標籤的後代**，舉例來說，bs.body.h1選取body標籤的後代的第1個h1標籤，不會找body外的標籤

同樣的，bs.div.find\_all(‘img’)會找出文件中的第1個div標籤，然後找出該div標籤後帶中的所有img標籤。

**處裡姊妹next\_siblings 、 previous\_siblings()**

BeautifulSoup的next\_siblings()函式方便從表格蒐集資料，特別是有標題列時，next\_siblings()相對的是previous\_siblings()函式，若方便找的標籤在要處裡清單尾部時很好用。

**清單:**



**處裡父節點**

處裡子節點的相反，由後代往先代找

**正規表示式**<https://poychang.github.io/note-regular-expression/>

**正規表示式與BeautifulSoup**

**存取屬性**

已經知道如何存取或過濾標籤並存取其內容。但有時網路擷取不是找標籤的內容；**你要找的是屬性**。這對a等URL放在href屬性的標籤特別有用；或是img標籤，他的圖片放在src屬性中

如下**呼叫可存取標籤物件的屬性清單**:

myTag.attrs

要記得他回傳1個可存取與操作這些屬性的python字典物件。舉例來說，圖形的來源可如下尋:

myImdTag.attrs[‘src’]

**Lambda表示式**

<https://poychang.github.io/note-regular-expression/>