Requests,BeautifulSoup小抄

Requests

Lower-Level Classes中有class Request, Response

encoding：得到/設定r.text的編碼模式(text屬性：unicode編碼的response內容)

Request.get(url)：Send a GET request，回傳一個Response Object

Response.json()：把取得的json格式Response Object解析為字典

BeautifulSoup

BS處理四種物件：Tag、NavigableString(NS，沒有children)、BeautifulSoup(BS)、Comment

soup.title ➔ 回傳一個Tag物件

soup.prettify() ➔ 排好看，如檢查看到的Html排版方式輸出(跟elementtree的indent很像)

tag.text ➔ 取得tag的文字內容

tag.name ➔ title(標籤名)

tag.[屬性] ➔ 屬性內容

tag.get('src') ➔ 取得屬性內容

tag.attrs ➔ 看有哪些屬性及內容，會得到一本字典({'src' : ....., 'alt' : .....})

tag.string ➔ 標籤中要顯示於網頁的文字(NavigableString物件)，str(tag.string)可讓其變為純文字

Navigating the tree(Going Down，使用tag導航)

soup.body.b ➔ 到body中的第一個b標籤

soup.a ➔ 到第一個a標籤

soup.find_all('a') ➔ 串到.內所有的a標籤

取得網頁元素

CSS Selectors	find
1.所有標籤為p 'p'	1.所有標籤為p 'p'
2.body內所有a 'body a'	2. 所有b開頭標籤 're.compile("^b")'
3.符合路徑 'body>div>div>p'	3.標籤a及標籤b ["a", "b"]
4.Id為tag1 '#tag1'	4.標籤p且屬性title "p","title"
5.class為center '.center'	5.ID為link2 id = "link2"
6.存在href屬性的a 'a[href]'	6.class=body的p標籤 "p",class_="body"
	7.class=sister的a標籤 "a", attrs={"class": "sister"}