**Python学习总结笔记（6）-- HTML解析之HTMLParser**

2016年11月22日 16:38:12

阅读数：292

网页爬虫避免不了要解析html页面，需要获取其中的超链接等信息。HTML本质上也是XML文档，但是其语法没有那么严格，所以不能用XML解析的方法来解析HTML（基于Dom方式解析XML可参见我的博客[Python学习总结笔记（1）–Dom XML操作](http://blog.csdn.net/kikaylee/article/details/53178786)）。

**0x01 HTMLParser**

Python官方提供了一个解析HTML的标准库HTMLParser，分析出HTML里面的标签、数据等。需要指出的是，HTMLParser采用的是事件驱动的方式，即当解析到一个特定的标签时，会触发相应的处理事件，用户通过重写相应事件的处理函数来实现HTML的解析。HTMLParser库中常见的几个处理函数如下：

handle\_startendtag 处理开始标签和结束标签

handle\_starttag 处理开始标签，比如<xx>

handle\_endtag 处理结束标签，比如</xx>

handle\_charref 处理特殊字符串，就是以&#开头的，一般是内码表示的字符

handle\_entityref 处理一些特殊字符，以&开头的，比如 &nbsp;

handle\_data 处理数据，就是<xx>data</xx>中间的那些数据

handle\_comment 处理注释

handle\_decl 处理<!开头的，比如<!DOCTYPE html ...

handle\_pi 处理形如<?instruction>的东西

* 1关于HTMLParser的详细介绍可以查看[官方文档](https://docs.python.org/3/library/html.parser.html)。

**0x02 实例**

下面是一个很简单的获取页面全部静态超链接的例子：

#/usr/bin/env python

#coding:utf-8

\_\_author\_\_ = 'kikay'

from HTMLParser import HTMLParser

import requests

#扩展的HTML解析类(提取全部的超链接)

class MyHtmlParserHelper(HTMLParser):

def \_\_init\_\_(self):

#super(MyHtmlParserHelper,self).\_\_init\_\_()

HTMLParser.\_\_init\_\_(self)

self.\_\_items=[]

self.\_\_href=None

self.\_\_text=None

self.\_\_flag=False

#处理开头

def handle\_decl(self, decl):

pass

#处理开始标签

def handle\_starttag(self, tag, attrs):

if 'a'==tag and attrs:

for key,value in attrs:

if 'href'==key:

self.\_\_href=value

self.\_\_flag=True

#处理结束标签

def handle\_endtag(self, tag):

if 'a'==tag and self.\_\_flag==True:

info=('{txt}({hf})').format(txt=self.\_\_text,hf=self.\_\_href)

self.\_\_items.append(info)

self.\_\_flag=False

#处理标签中间的数据

def handle\_data(self, data):

if self.\_\_flag==True:

self.\_\_text=data

@property

def items(self):

return self.\_\_items

if \_\_name\_\_=='\_\_main\_\_':

url='http://www.baidu.com'

response=requests.get(url=url,verify=False)

#解析

parser=MyHtmlParserHelper()

parser.feed(response.content)

items=parser.items

for i in items:

print i

* 4运行结果：

新闻(http://news.baidu.com)

hao123(http://www.hao123.com)

地图(http://map.baidu.com)

视频(http://v.baidu.com)

贴吧(http://tieba.baidu.com)

登录(http://www.baidu.com/bdorz/login.gif?login&tpl=mn&u=http%3A%2F%2Fwww.baidu.com%2f%3fbdorz\_come%3d1)

更多产品(//www.baidu.com/more/)

关于百度(http://home.baidu.com)

About Baidu(http://ir.baidu.com)

使用百度前必读(http://www.baidu.com/duty/)

意见反馈(http://jianyi.baidu.com/)

* 1**0x03 后记**

HTMLParser是一个很简单的HTML解析库，功能不算强大，并且还是个很老的库，比如其不支持新式类的继承方式（super写法，见上面的例子），解析静态的HTML页面使用BeautifulSoup等库显然更适合，但是作为一个比较经典的库，了解下还是有必要的。