python下很帅气的爬虫包 - Beautiful Soup 示例

先发一下官方文档地址。http://www.crummy.com/software/BeautifulSoup/bs4/doc/

建议有时间可以看一下python包的文档。

Beautiful Soup 相比其他的html解析有个非常重要的优势。html会被拆解为对象处理。全篇转化为字典和数组。

相比正则解析的爬虫，省略了学习正则的高成本。

相比xpath爬虫的解析，同样节约学习时间成本。虽然xpath已经简单点了。（爬虫框架Scrapy就是使用xpath）

安装

linux下可以执行

也可以用python的安装包工具来安装

使用简介

下面说一下BeautifulSoup 的使用。

解析html需要提取数据。其实主要有几点

1：获取指定tag的内容。

2：获取指定tag下的属性。

3：如何获取，就需要用到查找方法。

使用示例采用官方

格式化输出。

获取指定tag的内容

上面示例给出了4个方面

1：获取tag

soup.title

2：获取tag名称

soup.title.name

3：获取title tag的内容

soup.title.string

4：获取title的父节点tag的名称

soup.title.parent.name

怎么样，非常对象化的使用吧。

提取tag属性

下面要说一下如何提取href等属性。

获取属性。方法是

soup.tag['属性名称']

常见的应该是如上的提取联接。

代码是

相当easy吧。

查找与判断

接下来进入重要部分。全文搜索查找提取.

soup提供find与find_all用来查找。其中find在内部是调用了find_all来实现的。因此只说下find_all

看参数。

第一个是tag的名称，第二个是属性。第3个选择递归，text是判断内容。limit是提取数量限制。**kwargs 就是字典传递了。。

举例使用。

获取内容和字符串

获取tag的字符串

注意在实际使用中应该使用 unicode(title_tag.string)来转换为纯粹的string对象

使用strings属性会返回soup的构造1个迭代器，迭代tag对象下面的所有文本内容

获取内容

.contents会以列表形式返回tag下的节点。

想想，应该没有什么其他的了。。其他的也可以看文档学习使用。

总结

其实使用起主要是