Python爬虫常用模块（二）-Beautiful Soup(bs4)

2017-10-13 [每日一Python](https://mp.weixin.qq.com/s?__biz=MzI2Mjg5NDA4Ng==&mid=2247483699&idx=1&sn=9ed6812b78734ade06f86a588ec95fc6&chksm=ea45665add32ef4c3e3950d600ded3f1f03c6a2f9798e00381272e4046a06d334a0d3ca0a54d&scene=27&key=90b792fe8a39b59d28e6f9170d47abe9203f4d5840ccbc57a2f24cde890b0a4bd6e29f86f221a9803ccd439e04c161c94bd345a714e7d76a5849414f4d4016e59266a1d9653bcb76b37723ce0593b38d&ascene=7&uin=Nzk3NjIyNQ%3D%3D&devicetype=Windows+7&version=6205051a&pass_ticket=pVlB793F3uK4YJsVaPHBeSBPYMcGIImT6eadIav5PlM%3D&winzoom=1##)

爬虫

Python爬虫常用模块（二）-Beautiful Soup(bs4)

爬虫的基本框架是获取HTML页面信息，解析页面信息，保存结果；bs4库主要是解析、遍历、提取页面信息；bs4将复杂的HTML文档转化为一个树形结构，每个节点都是一个Python对象，这些对象可以分为四类：Tag,NavigableString,Comment,BeautfifulSoup

Tag:标签，是最基本的信息组织单元，对应HTML原生文档的tag，Tag的两个基本属性是name和attributes;

BeautfifulSoup:表示的是一个文档的全部内容，大部分时候可以把它当成Tag对象，但并不是真正的HTML的tag，因此没有name和attribute属性，但可以用.name来查看它的名字“[document]”

NavigableString:标签内非属性字符串，<>-<\>内的字符串；

Coment:Comment对象时一个特殊的NavgableString对象，表示HTML文档的注释部分,使用特殊的格式输出(prettify())：

例子：

from bs4 import BeaufifulSoup 

soup = BeautifulSoup(

tag = soup.title

tag.name     # 可以获得tag 的name :'title'

tag.attrs      #获取属性:{},这个例子是空值

tag['class'] = 'for example'  #可以根据字典形式添加属性

tag.attrs     #再获取属性就是{‘class’: 'for example'}

print(tag)     #现在打印tag就可以得到：title class = "for example">网易</title>

遍历文档树：

下行遍历：

1.contents:子节点的列表，把所有<tag>的所有子节点都存入列表；

2.children:子节点的迭代类型，与contents类似，遍历子节点；

3.descendants:子孙节点的迭代类型，包含所有子孙节点，用于循环遍历；

上行遍历：

1.parent：获取某元素的父亲节点；

2.parents：迭代得到元素的所有父辈节点，用于循环遍历先辈节点；

平行遍历：

1.next\_sibling：返回按照HTML文档顺序的下一个平行节点标签

2.nex\_siblings：迭代类型，返回按照HTML文档顺序的后续所有平行节点标签；

3.previous\_sibling: 返回按照HTML文档顺序的上一个平行节点标签

4.previous\_siblings: 迭代类型，返回按照HTML文档顺序的前面所有平行节点标签；