BeautifulSoup库

文档内容为本人观看北京理工大学嵩天老师公开课的听课笔记与实践总结，图片为从该课程下载资料的截图，感谢嵩天老师。

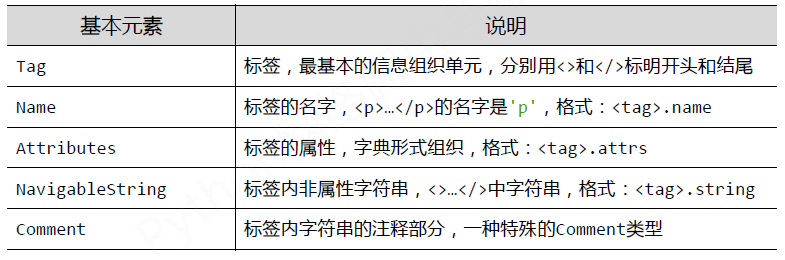
# 一、BeautifulSoup库入门

# 1、理解

该库的作用为：定向网络的数据爬取与网页解析。BS类是对Tag类的继承。提供了html文档到python对象的映射。可以简单地理解为BS将html封装成一个标签集合，我们可以通过标签名字来访问其包含的标签对象，例如.a、.body等，若相同标签有多个则只返回第一个。如果想全部检索，可以使用后面的find\_all方法。

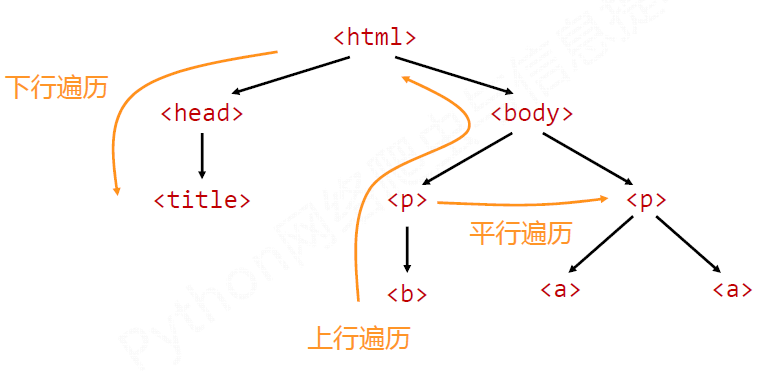
## 2、Tag基本方法

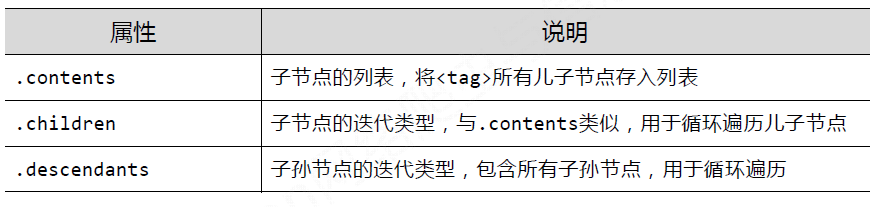
可以分为两类，分别是访问自身的方法和遍历方法。

访问自身的方法包括以下四种：

值得注意的是，本来tag.string是访问标签的字符串部分，但是实际上在有嵌套的标签中，会访问失败，返回值为空，当用于简单的标签（不含嵌套时）可以正常的使用，一种补救的方法是用.get\_text()方法，但返回的是该标签内的所有文本信息，即包含被嵌套标签的string。

遍历其他标签的方法包括：下行遍历、上行遍历、平行遍历。



**下行遍历：**

.contents返回子标签列表，列表元素仍为tag类型。

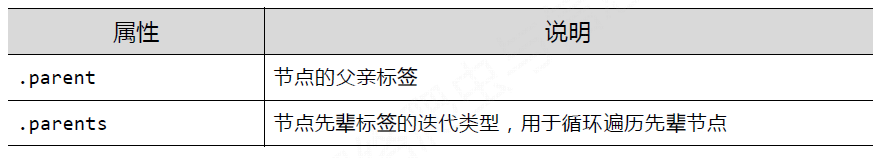
.children返回迭代类型，不能直接访问，需要用for循环来遍历。

for child in soup.body.children:

print(child)

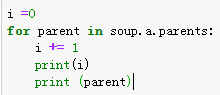
.decendants方法在jupyter测试时候报错，用pycharm正常，不知何故。

for son in soup.body.descendants:  
 print(son)

**上行遍历**

.parent返回父（复合）标签，这个父标签支持遍历，内容是该标签的contents，实际上，实际上只要是标签就有.contents，就支持遍历。

.parents返回有tag和BS类构成的generator类，例如：



1 <p class="course">Python is a wonderful general-purpose programming ...</p>

2 <body><p class="title"><b>......</b></p></body>

3 <html><body>......</body></html>

4 <html><body>......</body></html>

类型分别是：

1 <class 'bs4.element.Tag'>

2 <class 'bs4.element.Tag'>

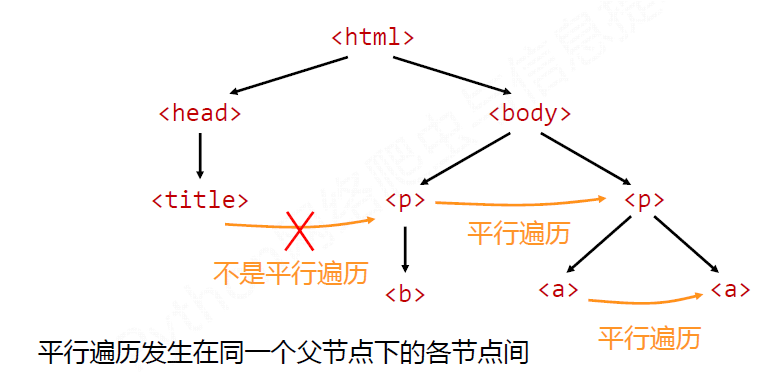
3 <class 'bs4.element.Tag'>

4 <class 'bs4.BeautifulSoup'>

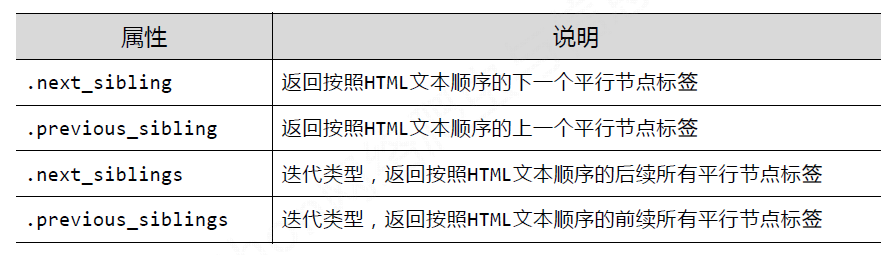
也就是说：上行遍历的终点是beautifulsoup类也就是根。

**平行遍历：**

首先，具有同一个父节点的标签节点才具有平行关系，如下图所示：



平行遍历的方法包括：



平行遍历（无论前后）即是在父节点的.contents列表里面遍历，所以有时会有换行符。

最后，bs4库的prettify()方法可以使得bs类对象更美观的显示，bs4和python3都默认utf-8编码，所以最好用3以上版本开发。

代码链接：