BeautifulSoup 库的理解：

HTML 都是由一对对标签组成的，标签是<html></html>, 而一堆标签组成的叫做标签树

标签树：

<html>

<body>

<p class”title”> ... </p>

</body>

<html>

所以我们可以说BeaurifulSoup是解析、遍历、维护“标签树”的功能库

标签的格式：

<p class=”title”> ... </p>

- <p> ... 中间的叫标签对 ...</p>

- <p>和</p>是叫标签的名Name，一般上都成对出现

- class=”title”则代表属性Attributes（可以理解为变量），可以有很多个

Beautiful Soup 的引用：

- from bs4 import BeautifulSoup (单单引入BeautifulSoup库)

- import bs4 （引入bs4 里包过 BeautifulSoup 和其他的库）

Beautiful Soup类：

HTMl == 标签树 == BeautifulSoup类

这里我们给出几个代码(这里要注意，as关键字在很复杂的代码里尽量不要使用，会导致代码变得混淆，因为要去猜库的名是什么)：

- from bs4 import BeautifulSoup as bs

- soup = bs(“<html>data</html>”, “html.parser”)

- soup2 = bs(open(“homt/simon/demo.html”), “”html.parser)

BeautifulSoup是对应一整个XML/HTML文档的全部内容

BeaurifulSoup解析器：

解析器 使用方法 条件

- bs4 的 HTML 解析器 BeautifulSoup(mk,”html.parser”) 安装bs4库

- lxml的 HTML 解析器 BeautifulSoup(mk,”lxml”) pip3 install lxml

- lxml 的XML 解析器 BeautifulSoup(mk,”xml”) pip3 install lxml

- html5lib 的 解析器 BeautifulSoup(mk,”html5lib”) pip3 install html5lib

Beautiful Soup 类的基本元素（也可以说是HTML中的基本元素）

- tag : 标签，最基本的信息组织单元，分别为<>和</>标明开头和结尾

- Name ： 标签的名字，<p> ... </p>的名字是’p’, 格式： <tag>.name

- Attributes ： 标签的属性，字典形式组织，格式：<tag>.attrs

- NavigableString : 标签内非属性字符， <> ... </>中的字符串，格式：<tag>.string

- Comment： 标签内字符串的注解部分，一种特殊情况的Comment类型

代码：

from bs4 import BeautifulSoup

soup = BeautifulSoup(demo, “html.parser”)

soup.title # 获得demo的标签信息