# 版本和安装问题，工作原理

bs3 版本不再支持，现在，请使用bs4

在linux下：apt-get install python-bs4

通过PyPi：what install beautifulsoup4

bs是安装python2的代码打包的，当使用python3的时候，它会自动转化成python3的代码。

bs使用对一个输入，使用一个解析器来解析，然后生成一棵树。

# 选择HTML解析器

默认 html.parser

lxml

html5lib

bs=BeautifulSoup(markup,’lxml/html5lib/html.parser/lxml-xml’)

每个解析器解析输入的markup输出文档，它本身只能假设输入是给定的类型。

# 编码问题

src源经过soup的处理输出，在此过程中涉及编码。

首先，soup会尽量猜测编码，然后转化为utf-8输出。但是如果猜测错误（文本过少），就应当指定编码如下：

soup=BeautifulSoup(markup,from\_encoding=”what-code”)

如果输出想使用对应编码，使用soup.prettify(‘what-code’)

# 部分解析

使用SoupStrainer类，在BeautifulSoup构造器中指定parse\_only=参数即可

SoupStrainer(tag=’a’,id=’link2’)

# 处理的对象类型

## Tag

对应于html标签，属性{name，[‘class’,font’,…]}

tag.name

tag.attrs

tag.其他-----🡪其他将会查找tag的子标签

可以对tag进行增删改查的操作

导航navigating自上而下的原则：

1. 使用子tag作为属性 （可能会产生冲突）,只返回第一个子标签
2. 使用内置属性如name，attrs
3. .childer生成器和.contents列表，表示tag的直接子tag
4. .parent 父tag对象
5. .next\_sibling 和.previous\_sibling 表示同级的兄弟节点

## NavigableString

处在标签之间的内容就是该对象，它可以当做str的子类

tag.string就是该对象，但是这种类型没有更多的属性了，它不能内嵌标签。

## BeautifulSoup

代表了文档内容本身，某种意义上来说，这就是一个根tag而已，不过这个跟tag没有自己的name属性。

## Comment

# 实例

find\_all(‘tag’) ---> 寻找所有的<tag>…</tag> Signature: find\_all([name](https://www.crummy.com/software/BeautifulSoup/bs4/doc/#id11), [attrs](https://www.crummy.com/software/BeautifulSoup/bs4/doc/#attrs), [recursive](https://www.crummy.com/software/BeautifulSoup/bs4/doc/#recursive), [string](https://www.crummy.com/software/BeautifulSoup/bs4/doc/#id12), [limit](https://www.crummy.com/software/BeautifulSoup/bs4/doc/#limit), [\*\*kwargs](https://www.crummy.com/software/BeautifulSoup/bs4/doc/#kwargs))

find和find\_all是两个最为重要的方法：

可以传递正则表达式:find\_all(re.compile(‘^b’))

传递列表，[‘a’,’h1’]则会查找所有的a，h1标签

传递函数，该函数接受一个tag作为参数，返回布尔值。

class\_=re.compile("itl"))

.contents和.children:区别在于，后者是一个生成器，前者是一个列表

.string 代表该tag的内容部分

## 安装pip：Installing with get-pip.py

To install pip, securely download [get-pip.py](https://bootstrap.pypa.io/get-pip.py). [[2]](https://pip.pypa.io/en/stable/installing/#id8)

Then run the following:

python get-pip.py

**安装bs4**：pip install beautifulsoup4