bs4解析-HTML语法

bs4解析比较简单, 但是呢, 首先你需要了解一丢丢的html知识. 然后再去使用bs4去提取, 逻辑和编写难度就会非常简单和清晰

HTML(Hyper Text Markup Language)超文本标记语言, 是我们编写网页的最基本也是最核心的一种语言. 其语法规则就是用不同的标签对网页上的内容进行标记, 从而使网页显示出不同的展示效果.

```
<h1>
    我爱你
</h1>
```

上述代码的含义是在页面中显示"我爱你"三个字, 但是我爱你三个字被"<h1>"和"</h1>"标记了. 白话就是被括起来了. 被H1这个标签括起来了. 这个时候. 浏览器在展示的时候就会让我爱你变粗变大. 俗称标题, 所以HTML的语法就是用类似这样的标签对页面内容进行标记. 不同的标签表现出来的效果也是不一样的.

```
h1: 一级标题
h2: 二级标题
p: 段落
font: 字体(被废弃了, 但能用)
body: 主体
```

这里只是给小白们简单科普一下, 其实HTML标签还有很多很多的. 我们不需要一一列举(这是爬虫课, 不是前端课).

OK~ 标签我们明白了, 接下来就是属性了.

```
<h1>
    我爱你
</h1>
<h1 align='right'>
    我爱你妹
</h1>
```

有意思了. 我们发现在标签中还可以给出xxx=xxx这样的东西. 那么它又是什么呢? 又该如何解读呢?

首先, 这两个标签都是h1标签, 都是一级标题, 但是下面这个会显示在右边. 也就是说, 通过xxx=xxx这种形式对h1标签进一步的说明了. 那么这种语法在html中被称为标签的属性. 并且属性可以有很多个. 例如:

```
<body text="green" bgcolor="#eee">
    你看我的颜色. 贼健康
</body>
```

总结, html语法:

```
<标签 属性="值" 属性="值">
    被标记的内容
</标签>
```

有了这些知识, 我们再去看bs4就会得心应手了. 因为bs4就是通过标签和属性去定位页面上的内容的.