爬虫步骤

1. 获取网页的URL地址
2. 分辨请求网页的方法GET POST
3. 查看网页的源代码
4. 撰写爬虫请求头

5.解析页面内容bs4 xpath re

6.存储信息的到本地

网页源代码

直接获取网页源代码是没有什么意义的。我们必须提取出我们需要的信息来，才能为数据分析和后续处理做准备。网页都是有规则的，我们只需要找出规则，写好一条URL的爬去代码，就可以不断循环迭代带其他网页代码上。常见的解析库有BeautifulSoup，Lxml，pyquery

信息存储

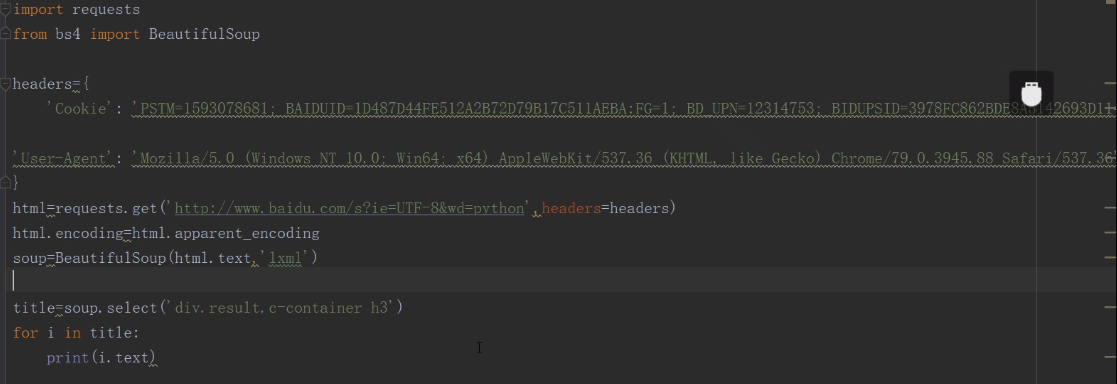
下载了源代码，解析了我们需要的信息，之后就是存储信息了。我们既可以直接存储在我们熟悉的文本格式里，比如csv，txt，word，excel等等文件里。也可以存储在各类数据库里，他们不占空间，易于管理，方便统一化，规范化编写，是实际开发中的不二选择。比如MYSQL,Mongodb，Redis数据库等。开发者工具的使用

开发人员工具-网络-状态码是200是正常返回

JSON是一种轻量级的数据交换格式应用于如今前后端分离的情景当中主要用途是数据交互也就是我们口中的接口（发起请求，返回数据）通常与Ajax请求搭配使用







Requests请求库

1.get请求相当于我们在浏览器中按下一次回车



(1)通过params加参数 (2)直接写在url里



1. 如果说get是获取页面资源那POST就是提交资源给页面

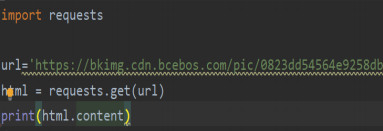


1. 设置编码格式：\*\*utf-8 utf8 UTF8 UTF-8 都是可以的，一样的



获取二进制文件

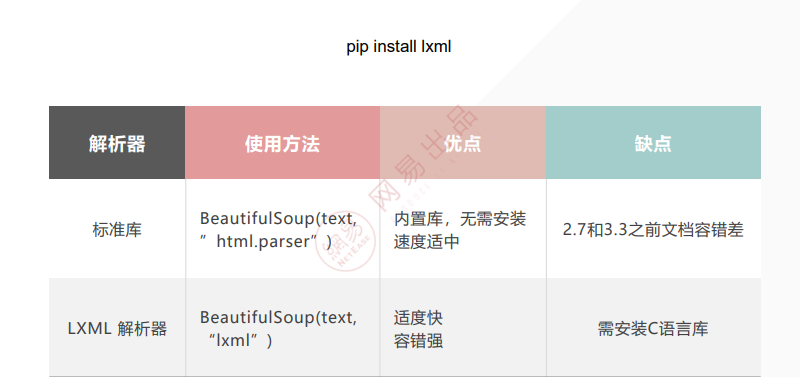
何为二进制文本？图片，音乐，视频等在爬虫是获取到的就是二进制文本我们不能使用text来获取，要使用content我们需要根据爬取的数据使用不同的方法



BeautifulSoup语法

为什么要用：（1）Beautiful Soup 是一个可以从HTML文件中提取数据的Python库

1. Beautiful Soup将复杂HTML文档转换成一个树形结构,每个节点都是Python对象

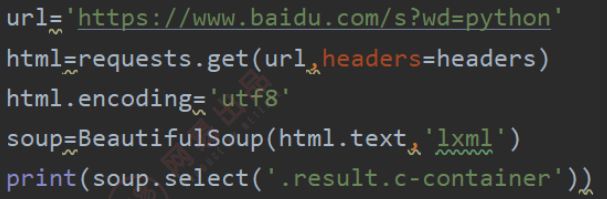


CSS选择器

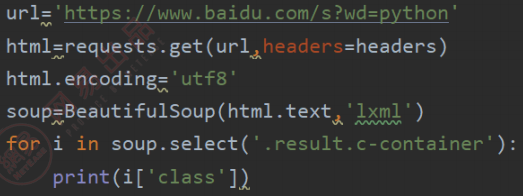
节点选择须知class属性用.id属性用#

获取了一个页面中所有的含有result和c-container class的标签

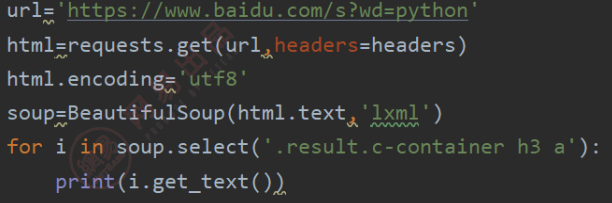
我们发现标签名是可省略的（在有class或者id的情况下）



获取属性的时候可以像取得字典一样去直接拿



获取文本

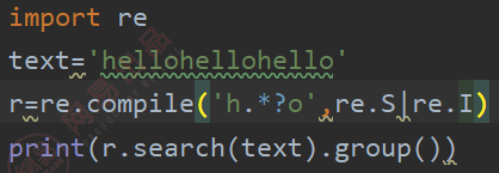




Python中的正则表达式

match 从头匹配 search 全局匹配 group 获取内容 findall 匹配全部

re.S 匹配换行符 re.I 不区分大小写



m.group() == m.group(0) == 所有匹配的字符，与括号无关

m.group(N) 返回第N组括号匹配的字符

m.groups() 返回所有括号匹配的字符

m.groups() == (m.group(0),m.group(1), ...)

compile里边编写正则表达式 findall用于查找全部 .\*?为惰性匹配

