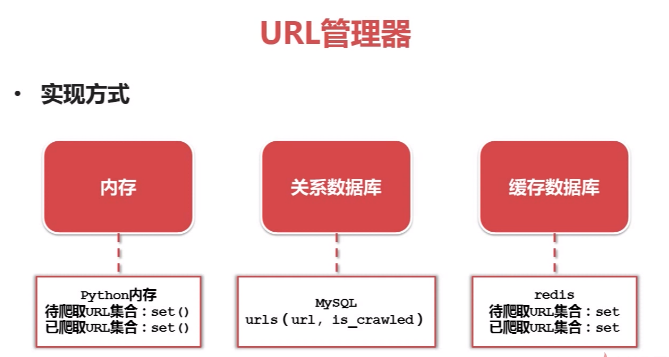




Url管理器的3中存储方式



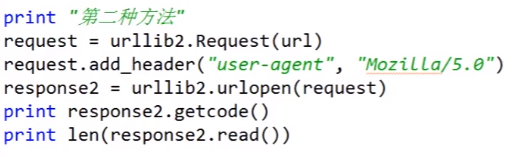
网页下载器原理流程：（通过url下载对应网页html转码成字符串，然后去识别）

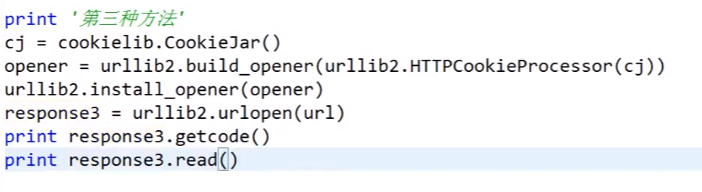


网页下载器urllib2(python自带类)基本用法：

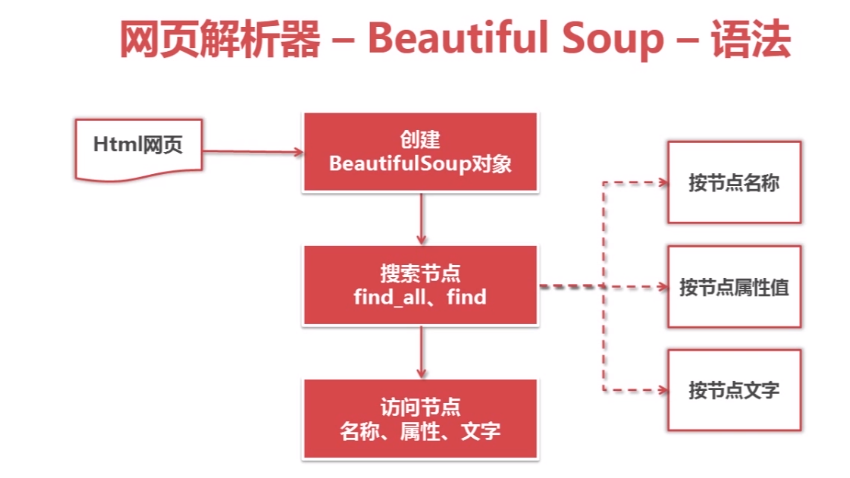




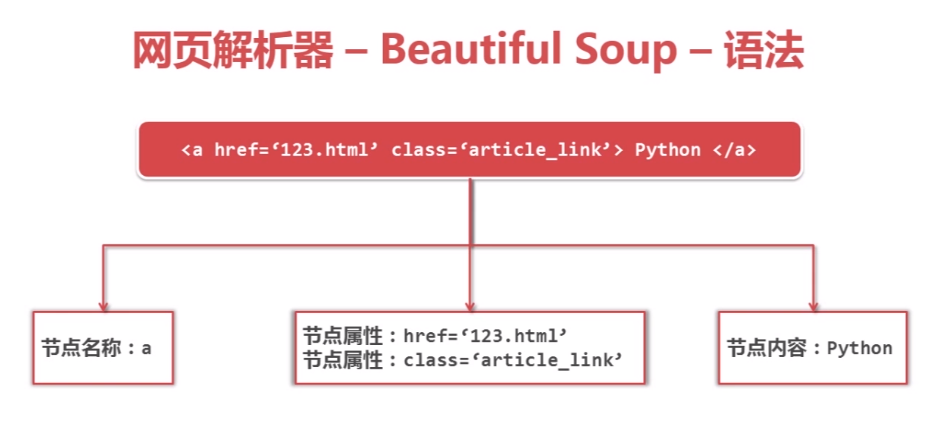




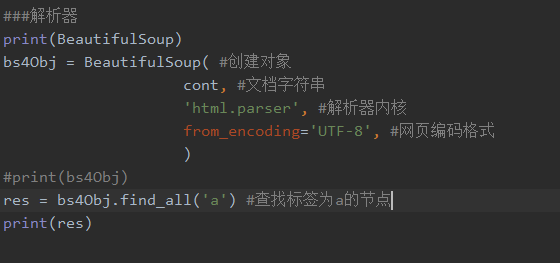
网页解析器基本用法流程：



举个栗子：



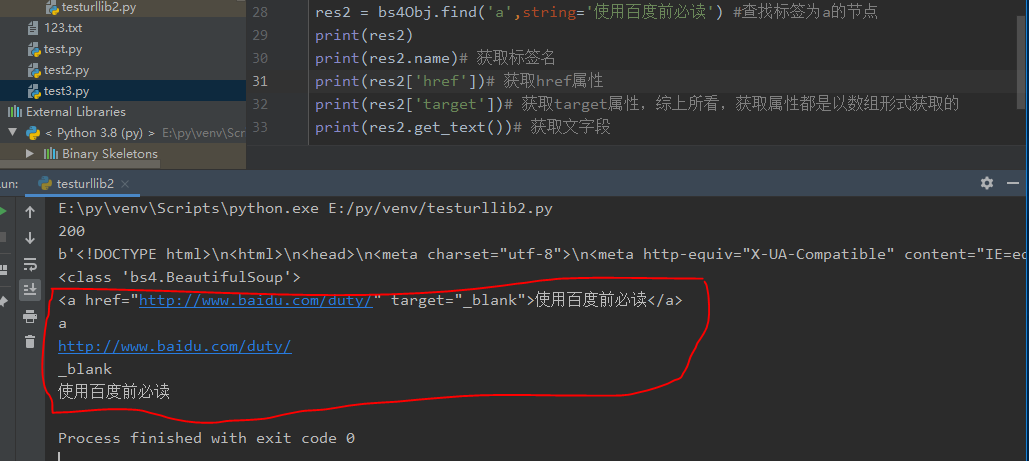
代码栗子：



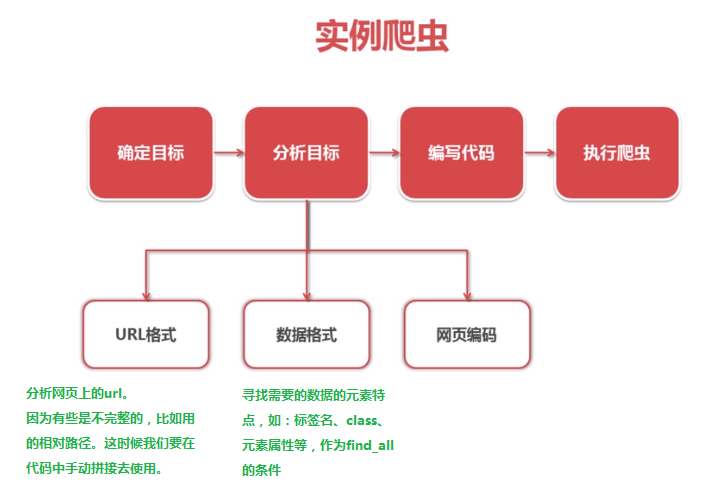
获取节点find\_all，find的用法



获取到节点后的常用使用方法：



## 分析抓取目标



## 代码架构

**主模块：**

**定义入口变量root\_url**

**定义爬虫类Spider**

**使用爬虫类“开始”方法Spider.start(root\_url)**

**Spider类：**

**\_\_init\_\_**

**定义url管理器urlManeger**

**定义下载器downloader**

**定义解析器parser**

**定义输出器outputer**

**start(root\_url)**

**往urlManeger中加入url**

**当urlManeger中的爬取列表还有urL**

**获取首条url**

**下载html**

**获取新的需要的url和需要的数据**

**把获取到的url加入urlManeger**

**记录获取到的新数据**

**urlManeger类：**

**new\_url\_list #待爬取的url列表**

**old\_url\_list #已爬取的url列表**

**\_\_init\_\_**

**Null**

**add\_url(self)**

**has\_url(self)**

**get\_url(self)**