慕课网：Python开发简单爬虫

1.爬虫简介

爬虫：一段自动抓取互联网信息的程序

价值：互联网数据，为我所用

2.简单爬虫架构

爬虫调度端—>爬虫：（**URL管理器**->**网页下载器**->**网页解析器**：还可以补充URL）->价值数据

3.URL管理器：管理待抓取URL集合和已抓取URL集合——防止重复抓取、防止循环抓取

实现方式：内存(Python内存:两个set)、关系数据库(MySQL)、缓存数据库(redis:两个set)

urllib2下载网页方法1：最简洁方法

```
import urllib2
#直接请求
response=urllib2.urlopen('http://www.baidu.com')
#获取状态码，如果是200，表示获取成功
print response.getcode()
#读取内容
cont=response.read()
```

urllib2下载网页方法2：添加data、http header

```
# url、data、header->urllib2.Request->urllib2.urlopen(request)
import urllib2
#创建Request对象
url='http://www.baidu.com'
request=urllib2.Request(url)
#添加数据
request.add_data('a','1')
#添加http的header
request.add_header('User-Agent','Mozilla/5.0')
#发送请求获取数据
response=urllib2.urlopen(request)
```

urllib2下载网页方法3：添加特殊情景的处理器

```
# HTTPCookieProcessor、ProxyHandler、HTTPSHandler、HTTPRedirectHandler-
>opener=urllib2.build_opener(handler)->urllib2.install_opener(opener)-
>response=urllib2.urlopen(url)  response=urllib2.urlopen(request)
import urllib2,cookielib
#创建cookie容器
cj=cookielib.CookieJar()
#创建一个opener
opener=urllib2.build_opener(urllib2.HTTPCookieProcessor(cj))
#给urllib2安装opener
urllib2.install_opener(opener)
#使用带有cookie的urllib2访问网页
response=urllib2.urlopen("http://www.baidu.com/")
```

网页解析器：补充URL，获取价值数据

种类：模糊匹配：正则表达式

　　结构化解析：html.parser　　BeautifulSoup(第三方插件)　lxml(第三方插件)

结构化解析-DOM(Document Object Model)树

4.网页下载器(urllib2-Python自带)：将互联网上URL对应的网页下载到本地的工具
Python有哪几种网页下载器：urllib2-官方模块　　requests-第三方，更强大
5.网页解析器(BeautifulSoup)：第三方插件
6.完整实例：爬取百度百科Python词条相关的1000个页面数据
实例爬虫：
1.确定目标
2.分析目标：URL格式，数据格式，网页编码
3.编写代码
4.执行爬虫

分析目标：
目标：百度百科Python词条相关词条网页-标题和简介
入口页：https://baike.baidu.com/item/Python/407313
URL格式：
——词条页面URL：/item/%E8%AE%A1%E7%AE%97%E6%9C%BA%E7%A8%8B%E5%BA%8F%E8%AE%BE%E8%AE%A1%E8%AF%AD%E8%A8%80
数据格式：
——标题：
　　<dd class="lemmaWgt-lemmaTitle-title"> <h1>***</h1></dd>
——简介：
　　***
页面编码：UTF-8

管理器：manager　　解析器：parser　　下载器：download　　输出器：outputer