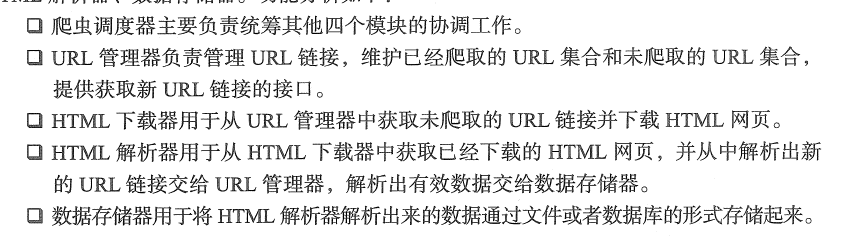
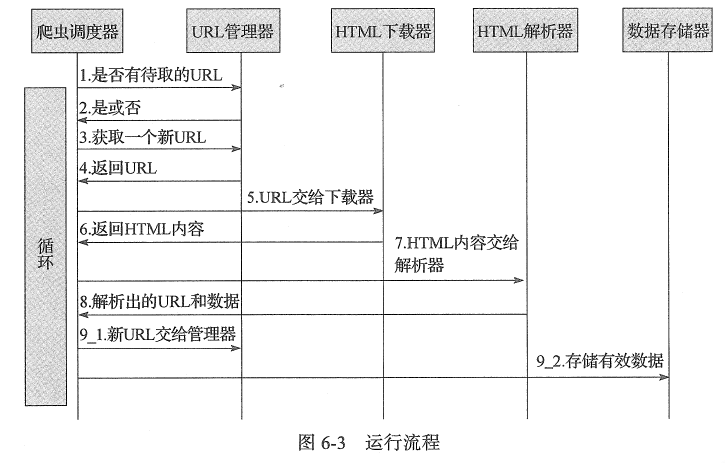
基础爬虫架构

爬虫调度器，URL管理器，HTML下载器，HTML解析器，数据存储器



运行流程



**URL管理器**

1. 判断是否有待取的URL，方法定义为has\_new\_url
2. 添加新的URL到未爬取集合中，方法定义为add\_new\_url（url），add\_new\_urls(urls)
3. 获取一个未爬取URL，方法定义为get\_new\_url（）
4. 获取未爬取URL集合的大小，方法定义为new\_url\_size（）
5. 获取已经爬取的URL集合的大小，方法定义为old\_url\_size（）

**HTML下载器**

下载器需要用到Requests模块，实现一个download（url）接口

**HTML解析器**

解析器使用BeautifulSoup4进行HTML解析。需要解析部分主要分为提取相关词条页面的URL和提取当前词条的标题和摘要信息，最后需要分析一下需要提取的URL的格式，使用urllib.requests.urljoin函数将当前网址和相对网址拼接成完整的URL路径

解析器主要提供一个parser对外接口，输入参数为当前页面的URL和HTML下载器返回的网页内容

**数据存储器**

数据存储器主要包括两个方法：store\_data(data)用于将解析出来的数据存储到内存中，output\_html()用于将存储的数据输出位指定的文件格式

**爬虫调度器**

初始化各个模块，然后通过crawl（root\_url）方法传入入口URL