# 关于新闻爬虫的代码讲解

## 什么是爬虫

网络爬虫是一种按照一定的规则，自动的爬取万维网信息的程序或者脚本。网络爬虫按照系统结构和实现技术，大致分为通用爬虫，聚焦爬虫，增量式爬虫，深层网络爬虫。但是实际应用中一般都是几种爬虫技术相结合实现的。

搜索引擎其实是一种大型的复杂的网络爬虫属于通用爬虫的范畴。专业性质的搜索引擎则是属于聚焦爬虫的范畴。增量式爬虫则是对已下载的网页采取增量式更新和只爬取新产生的或者已经发生变化的网页的爬虫。Web网页按照存在的方式可以分为表层网页和深层网页，表层网页通常是指传统引擎可以索引的页面，以超链接可以到达的静态网页为主构成的Web页面；深层网络是那些大部分内容不能通过静态链接获取的、隐藏在搜索表单后的，只有用户提交一些关键词才能获取的Web页面。爬取深层网络页面的爬虫就属于深层网络爬虫的范畴。

## 爬虫的工作流程



1. 首先选取一部分或者一个精心挑选的中字URL；
2. 将这些URL放入带抓取的URL队列；
3. 从待抓取的URL队列中读取待抓取的URL，解析DNS，并且得到主机的IP，并将URL对应的网页下载下来，存储到已下载的网页数据中。并且将这些URL放进已抓取的URL队列中；
4. 分析已抓取URL队列中URL，从已下载的网页数据中分析出其他的URL，并和已抓取的URL进行去重处理，最后将去重后的URL放入待抓取URL队列，从而进入下一个循环。

## 爬虫的python实现框架Scrapy

爬虫的python实现的库有很多，主要有urllib,httplib/urllib,requests.这里主要讲一下Scrapy框架。

Scrapy使用python写的Crawler Framwork，简单轻巧，并且非常方便。Scrapy使用的是Twisted异步网络库来处理网络通信，架构清晰，并且包含了各种中间件接口，可以灵活的完成各种需求。



1. 引擎打开一个网站，找到处理改网站的spider并向该spider请求第一个要爬取的URL；
2. 引擎从Spider中获取的第一个要爬取的URL并通过调度器（Scheduler）以Requests进行调度；
3. 引擎向调度器（Scheduler）请求下一个要爬取的URL;
4. 调度器（Scheduler）返回下一个要爬取的URL给引擎（Scrapy Engine），引擎将URL通过下载器中间件（Downloader Middleware）转发给下载器；
5. 一旦页面下载完毕，下载器生成一个该页面的Response，并将其通过下载器中间件转发给引擎；
6. 引擎从下载器中间件接收到Response并通过spider中间件发送给spider处理；
7. Spider处理Response并返回爬取到的Item和新的Request给引擎；
8. 引擎将爬取到的item给Item Pipeline,将Request给调度器；
9. 从第2步开始重复直到调度器中没有Request，引擎关闭该网站。