python 魔鬼训练营 第11周

DATAGURU专业数据分析社区

法律声明

【声明】本视频和幻灯片为炼数成金网络课程的教学资料，所有资料只能在课程内使用，不得在课程以外范围散播，违者将可能被追究法律和经济责任。

课程详情访问炼数成金培训网站

http://edu.dataguru.cn

第11课：一只来自网页的爬虫

课程内容：

- 什么是爬虫
- 爬虫的原理
- 爬虫的实现

本次课内容重点讲解在python中经常应用的爬虫技术；了解爬虫的原理及如何实现一个简单的爬虫程序。

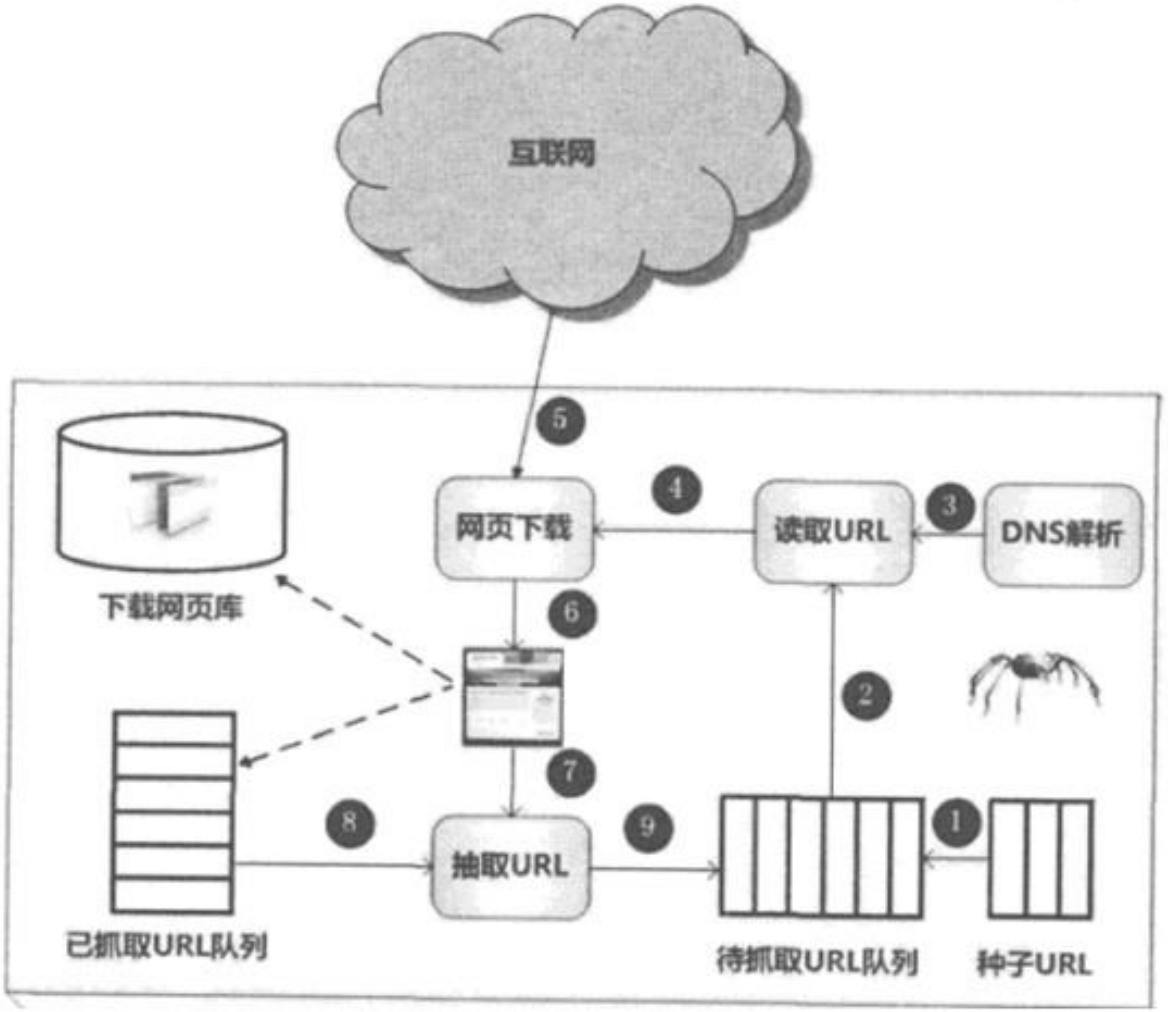
第1节：什么是爬虫

- 网络爬虫（又被称为网页蜘蛛，网络机器人，在FOAF社区中间，更经常的称为网页追逐者），是一种按照一定的规则，自动地抓取万维网信息的程序或者脚本。

- 爬虫的作用，主要用于在网络上抓取网页信息并存储在本地；便于其它程序对内容进行扫描和检索。

第2节：爬虫的原理

- 爬虫工作流程



第2节：爬虫的原理

- 爬虫抓取策略

1. 深度优先遍历策略：递归实现
2. 宽度优先遍历策略：追加
3. 反向链接数策略
4. Partial PageRank策略
5. OPIC策略策略
6. 大站优先策略

第2节：爬虫的原理

- 爬虫的分类

1. 批量型爬虫
2. 增量型爬虫
3. 垂直型爬虫

- 爬虫更新策略

1. 历史参考策略
2. 用户体验策略
3. 聚类抽样策略

第2节：爬虫的原理

- 爬虫实现机制

1. 多线程
2. 分布式

第3节：爬虫的实现

- 组成部分

1. 待抓取URL：url列表
2. 抓取程序：urllib2.urlopen
3. 分析程序：HTMLParser，SGMLParser，pyquery ，BeautifulSoup，re
4. 存储程序：file

第11课：一只来自网页的爬虫

要点回顾：

- 爬虫的原理
- 爬虫的组成部分

炼数成金逆向收费式网络课程

- Dataguru（炼数成金）是专业数据分析网站，提供教育，媒体，内容，社区，出版，数据分析业务等服务。我们的课程采用新兴的互联网教育形式，独创地发展了逆向收费式网络培训课程模式。既继承传统教育重学习氛围，重竞争压力的特点，同时又发挥互联网的威力打破时空限制，把天南地北志同道合的朋友组织在一起交流学习，使到原先孤立的学习个体组合成有组织的探索力量。并且把原先动辄成千上万的学习成本，直线下降至百元范围，造福大众。我们的目标是：低成本传播高价值知识，构架中国第一的网上知识流转阵地。
- 关于逆向收费式网络的详情，请看我们的培训网站 http://edu.dataguru.cn

Thanks

FAQ时间