极客学院
jikexueyuan.com

定向爬虫
动态加载网页的爬取

定向爬虫：动态加载网页的爬取 — **课程概要**

- AJAX介绍与网页展示
- 从js文件读取内容
- 构造目标地址
- 实战——腾讯视频评论爬虫

AJAX介绍与网页展示

AJAX介绍与网页展示

- AJAX定义与介绍
- AJAX网页特点
- AJAX网页举例

AJAX介绍与网页展示 — **AJAX定义与介绍**

AJAX即“Asynchronous Javascript And XML”（异步JavaScript和XML），是指一种创建交互式网页应用的网页开发技术。

通过在后台与服务器进行少量数据交换，AJAX 可以使网页实现异步更新。这意味着可以在不重新加载整个网页的情况下，对网页的某部分进行更新。

AJAX介绍与网页展示 — **AJAX网页特点**

- 页面加载快速
- 不刷新网页就能更新信息
- **源代码内容与网页内容不同**

从js文件读取内容

从js文件读取内容

- 审查元素列出js文件
- 寻找可疑文件
- 解析js文件内容

从js文件读取内容— **解析js文件内容**

- 解析json文本
- 获取文本内容

构造目标地址

构造目标地址

- 根据规律构造
- 来自文件
- 手动生成

构造目标地址— **根据规律构造**

- 页数
- 每页个数
- 其实数
- 其他

构造目标地址— **来自文件**

- id
- cid
- vid
- xxx

构造目标地址— **手动生成**

- 时间戳

实战——腾讯视频评论爬虫

实战——腾讯视频评论爬虫

目标网站：腾讯视频

目标网址：http://video.qq.com

目标内容：

某一个视频下面的评论，具体包括：

- 评论内容
- 评论人名称
- 评论时间

输出结果保存到MongoDB中。

定向爬虫：动态加载网页的爬取

本套课程中我们学习了动态加载网页的爬取，你应当掌握以下知识：

- 获取动态网页的加载内容
- 手动构造特殊js文件路径
- 爬取腾讯视频评论

这是《定向爬虫入门》的最后一次课程，希望各位同学能继续关注我的其他课程。

极客学院

jikexueyuan.com

中国最大的IT职业在线教育平台