小象学院

BBS 站点分析及爬虫实战

主　讲:　杨真

小象学院

Part 1 基础

- 环境搭建
- HTML 基础
- 第一个10行代码的爬虫
- 内容抽取及解析
- HTTP 协议
- 辅助工具

Part 2 爬虫

- 网站结构及常见爬虫策略
- BBS网站结构分析及方案
- 控制节奏

Part 3 进阶

- MySQL 数据库
- 多线程
- 并行抓取
- 网站服务架构
- 表单、登录及Cookie处理

Part 4 实战

- 网站结构分析
- 网页抓取方案
- 数据提取
- 存储方案

BBS 结构

- 首页：TOP10 热门文章
- 板块：获取所有的板块入口
- 文章：获取每个板块，每一页的文章列表
- 回帖：获取每一个帖子的主文，及所有跟帖

Board List

Format:

http://[domain]/nForum/section/[index]?ajax

注意：

存在子版块：汽车由 4 个子版块组成

Article List

Format:

http://www.newsmth.net/nForum/board/[BoardName]?ajax&p=[page]

注意：

http://www.newsmth.net/nForum/#!board/[BoardName]?p=[page] 这个请求只返回框架，没有数据

Post Detail

Format:

http://www.newsmth.net/nForum/article/[board_name]/[article_id]?ajax&p=[page]

小象学院

Part 1 基础

- 环境搭建
- HTML 基础
- 第一个10行代码的爬虫
- 内容抽取及解析
- HTTP 协议
- 辅助工具

Part 2 爬虫

- 网站结构及常见爬虫策略
- BBS网站结构分析及方案
- 控制节奏

Part 3 进阶

- MySQL 数据库
- 多线程
- 并行抓取
- 网站服务架构
- 表单、登录及Cookie处理

Part 4 实战

- 网站结构分析
- 网页抓取方案
- 数据提取
- 存储方案

控制节奏

- 网站对爬虫的限制，最主要依赖于每个IP（或每个用户）的访问频次，过高频率的访问会被网站限制访问
- 控制节奏主要针对每个目标地址的访问频率

小象学院

只为遇见明天更优秀的你！