小象学院

爬虫基础课程介绍

主　讲:　杨真

Part I 基础

- 环境搭建：Python 环境
- HTML 基础：网页知识，了解如何解析一个网页
- 第一个10行代码的爬虫
- 内容抽取及解析：XPATH、正则表达式，网页解析利器
- HTTP 协议：Header 定义、返回值 200、300、400、500
- 辅助工具：用于辅助分析网站通信协议、数据格式的工具

Part II 基础爬虫

小象学院

- 网站结构及常见爬虫策略：给出一个网站，分析它的结构并制定策略
- BBS网站结构分析及方案：对于一个BBS网站，来分析如何抓取
- 控制节奏：不能爬太快，IP 地址或账号会被屏蔽的

Part III 进阶

- MySQL 数据库：如何存储网页提取出来的数据
- 多线程：如何执行异步任务，比如抓网页和下载图片
- 并行抓取
- 网站服务架构：通过 Sitemap 来了解网站结构
- 表单、登录及Cookie处理：如果网站需要登录，怎么办？

Part IV 全面实战

通过对微博的接口分析，抓取某个微博的文本、用户信息、评论等数据，结合数据库、异步文件加载、多线程的使用，完成一个完整的微博的爬虫

就业方向

爬虫工程师：网络爬虫、数据抓取方向，包括搜索引擎公司、舆情监控、大数据处理与分析、人工智能的训练数据等

数据工程师：对系统数据做清洗、异构、分析、统计和简单的批量处理

谁适合？

系统编程经验少、Java 等语言并不熟悉，没有丰富的 Spring、Hadoop 等方面经验的初级工程师，可以从爬虫、数据工程师开始做起，熟悉计算机服务体系，例如 HTTP 协议、数据库、正则、服务器架构、防火墙等，一步步成为后台工程师，然后成为系统架构师

小象学院

只为遇见明天更优秀的你！