Python网络爬虫简介与表达式基础

阿里云 韦玮

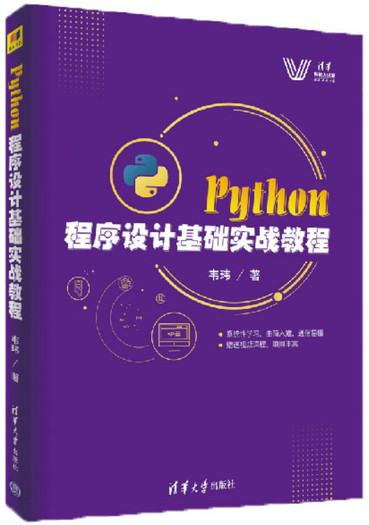
目录

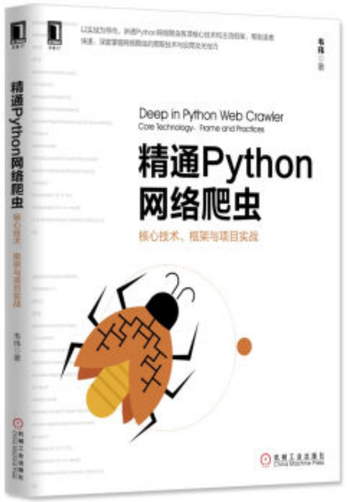
作者简介

韦玮

阿里云云栖社区专家

企业家，资深IT领域专家/讲师/作家，畅销书《精通Python网络爬虫》作者。





Python网络爬虫课程体系简介

- Python网络爬虫简介与表达式基础（本节）
- Urllib爬虫项目编写实战
- 抓包分析技术精讲
- Requests爬虫项目编写实战
- Scrapy爬虫项目编写实战
- 前程无忧招聘信息爬虫项目开发实战
- 淘宝网商品信息爬虫项目开发实战
- 知乎网信息爬虫项目开发实战（含登录）
- 爬虫常见的反爬策略与反爬攻克手段
- 分布式爬虫编写实战

Python基础课程：https://edu.aliyun.com/course/154

Python网络爬虫是什么？

网络爬虫是一种互联网信息的自动化采集程序，主要作用是代替人工对互联网中的数据进行自动采集与整理，以快速地、批量地获取目标数据。

如下所示，是网络爬虫可以做的一些事情：

- 批量采集某个领域的招聘数据，对某个行业的招聘情况进行分析
- 批量采集某个行业的电商数据，以分析出具体热销商品，进行商业决策
- 采集目标客户数据，以进行后续营销
- 批量爬取腾讯动漫的漫画，以实现脱网本地集中浏览
- 开发一款火车票抢票程序，以实现自动抢票

……

正则表达式基础实战

网络爬虫程序在将网页爬下来之后，其中还有一个关键的步骤就是需要对我们关注的目标信息进行提取，而表达式一般就是用于信息筛选提取的工具。

正则表达式是一种功能强大的表达式，并且非常好用，所以建议大家掌握。

本知识点将为大家介绍正则表达式的基础，接下来将进入实战讲解。

XPath表达式基础实战

除了正则表达式之外，还有一些非常好用的信息筛选的工具，比如XPath表达式、Beautiful Soup等等，当然，我们不可能也不需要都进行掌握，在此，我们讲解一下XPath表达式。

- / 逐层提取
- text() 提取标签下面的文本
- //标签名** 提取所有名为**的标签
- //标签名[@属性='属性值'] 提取属性为XX的标签
- @属性名 代表取某个属性值

接下来进入实战介绍。

为了无法计算的价值 | 阿里云