目录

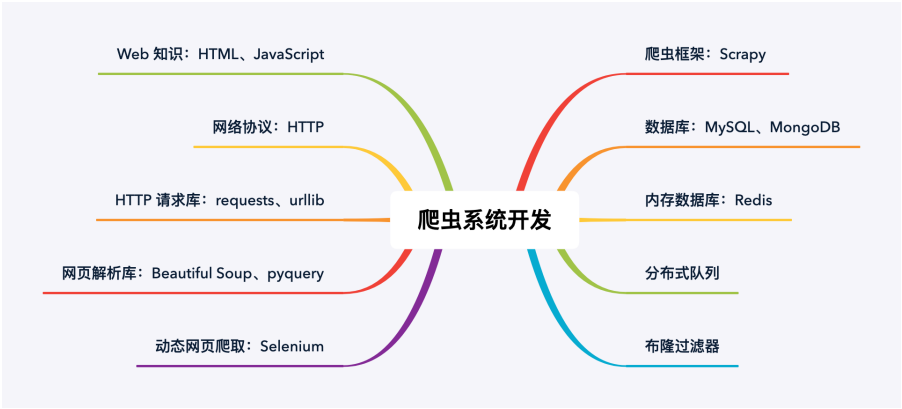
30 爬虫系统

更新时间：2019-11-06 09:33:26

人不可有傲气，但不可无傲骨。

——徐悲鸿

技术栈梳理



爬虫，即自动化的网页抓取程序，它能从网络中的大量网页里提取出所需的信息。网络是张大网，每个网页就是其中的一个交叉点，爬虫就顺着这张大网去爬取网页内容。

因为是从网页中提取内容，所以我们首先需要掌握基本的网页知识，如 HTML 和简单的 JavaScript。浏览器与 Web 服务器之间的通信采用 HTTP 协议，所以基本的 HTTP 知识也不能少。

爬虫需要模拟浏览器来向 Web 服务器发起请求，以获取网页内容。可以用 Python 的标准库 urllib，或者更好用的第三方库 requests 来达到这个目的。

拿到网页内容后，需要对网页进行解析，提取出其中的所需要的信息，或该网页上的其它网页链接。这时需要用到 Python 第三方库 Beautiful Soup 或 pyquery。

以上的抓取、解析过程也可以直接用专业的爬虫框架来完成，如 Scrapy，这是一种更工程化的方式。

当待抓取的网页数量很大时，这时需要注意网页判重的效率，也就是说需要有一种高效的方式来检查一个网页是否之前被抓取过，这就要用到布隆过滤器了。Redis 支持布隆过滤器扩展，能方便解决这个问题。

当待抓取的网页数量进一步扩大时，单机的爬虫程序效率就十分低下了，需要考虑构建分布式的爬虫程序。也就是说在多台机器上同时来跑爬虫任务。我们需要一个分布式队列来统一管理、调度所有的抓取任务，Scrapy-Redis 可以做这件事。

入门资料

了解和认识爬虫的最基本架构和运行流程，可以学习慕课网免费课程《Python开发简单爬虫》，或者阅读这篇知乎回答。

慕课网免费课程《Python最火爬虫框架Scrapy入门与实践》，可以带你了解和上手 Scrapy 的基本使用。Scrapy 的更详细内容可以阅读翻译文档《Scrapy入门教程》。

MySQL 数据库入门，参看慕课网免费课程《与MySQL的零距离接触》。MongoDB 数据库入门，参看 W3Cschool 在线教程《MongoDB教程》。

至于前面所涉及技术的全面讲解，可以阅读图书《Python3网络爬虫开发实战》的免费在线版，基本上所有的爬虫相关知识和工具都有覆盖，值得一读。

好了，资料不多，但要也要加油！

← 29 数据科学 31 机器学习 →

精选留言 0

欢迎在这里发表留言，作者筛选后可公开显示

目前暂无任何讨论