工作日记20180104分页的处理

Ringo Yungpo Kao

January 04, 2018

# 分页的处理

## 为何需要分页处理

由于我的数据来源是热搜词前百名的搜索结果的图书信息，而刚开始没有分页，直接获取信息。每页只有20个数据，所以最后去掉信息不全的，总共才获取了1600条数据。

## 分页处理开始

我大概先爬取了每个热词搜索首页的总页数信息，然后统计了一下，发现总共有20万左右的书籍信息。而我之前爬取1600条信息就需要一定的时间了。为了能够解析出所有的页面，我本来打算找下一页的链接的，结果发现没有办法分别上一页和下一页的链接，可能会无限循环解析。于是观察获取下一页的链接，发现只要更改一个页数数据就行。我就在获取了页数的情况下去解析每一个页面。

## 机器的强大

之前我抓取的数据也就是能够获取1600条，没想到现在就获得了20万条数据获取的能力。仅仅是加了几行代码。不过也是运行了好久就是，毕竟机器的运算速度和网络带宽是有限的。