工作日记20180103解决数据格式的问题

Ringo Yungpo Kao

January 03, 2018

# 解决数据格式的问题

## 遇到了数据处理问题

虽然相关数据爬取出来了，但是其实还可以继续解析出关键的数据，于是就用继续切割，并且用正则表达式抓取出关键的数据，比如年份，价格等。这样就能够用于后面的数据可视化。

## 关于如何解析

为了解析出关键数据，比如isbn，我去网上看了相关的正则表达式的写法，发现太过复杂，其实我只要根据我的数据特征去写就行了，于是就写了个只识别数字和横线的识别，发现效果不错，而价格也是这样，只写了个识别数字和点的。至于分类这个数据我不知道怎么存储好，毕竟是树状的数据结构，于是我将其直接提取所以的关键字，去掉重复的，当作标签来使用。其中馆藏数量我用图书馆的链接数获取，借出数量我用借出详情的链接数获取。

## 收获

本次的数据格式处理主要要根据数据的特征进行判断，当然也要做好取舍，毕竟有的数据很难获取，而且格式不一，不是所有数据都获取得来，不过我也就只要获取大多数就行。