工作日记20180102学校图书馆图书数据爬取

Ringo Yungpo Kao

January 02, 2018

# 学校图书馆图书数据爬取

## 选取学校图书馆网站

由于最近要做课程设计，而爬虫的知识并不懂，于是就去图书馆借书。在借书过程中用到了学校的图书馆网站，于是就想着要爬取这个网站的数据。

## 开始尝试

既然有了新的爬取目标，那就开始爬取，原本以为很容易的，但是没想到还是遇到了很多问题。很多书的信息不全，而爬取的过程是在一个文本里面爬取的，如果少了一些数据就会使其他数据爬取错误，于是这个数据就不能用了。不过由于用的是scrapy框架爬取，而且它是将要爬取的网站迭代出来，所以就算是报错，也就是迭代另一个网站，所以这个数据也相当于无效。由于我爬取的信息主要是中间的一段文本，可以通过特征来分割这个文本，通过特征分割可以得到相关数据。

由于数据量大，所以只爬取了搜索热词排行前100的数据，可能将来可以做一个跟借阅量有关的可视化

## 总结

这次总算是有能爬的数据了，其实今天本来想爬大公司的网站，比如外卖网站，但是，由于里面总是有一些完全看不懂的数据，不知道是代表什么意思，可能代表地理位置，于是放弃。经过多次尝试终于找到了能爬取的网址。