工作日记20180101动态页面抓取

Ringo Yungpo Kao

January 01, 2018

# 动态页面爬取

## 选取爬取网站

由于题目是校园网的爬取，而且最近有要求在学校的就业网填写信息，然后又看到传输的数据是动态获取的，于是想要爬取这个网站。

## 爬取失败

爬取这个网站关键要知道数据是如何传输的，chrome的控制台可以看到请求的网址，发现数据传输的都是post链接。这样爬取实际上是非常简单的，然而post传送的数据是几行代码，其中有换行符号，而换行符号是/n。在用python传出的时候。一个/会编程两个，就是//，而我没有办法解决。而用selenium爬取有太耗时间，因为需要等待网页加载才能爬取相关数据，而这个网站加载又太慢，爬取一个可观的数量的数据可能要等好久。而我又想在服务器上爬取，所以就放弃了。

## 总结

通常网站前端与后端传输是用的是restful api，而这个网站不是，这么特殊的案例让本来很容易的动态爬取变得极其困难。所以新手爬取可能还是得选好网站。