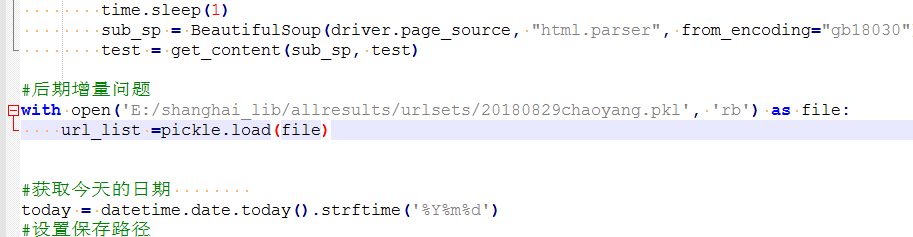
上海图书馆增量更新方案

参考网上增量更新方法后，我们觉得对于每一个excel文件维护一个网址的列表，当要爬的网站url不在此列表中时，代表其是需要爬取的新网站。而当爬的网站url在此列表中时代表该网站已经爬取过了。Python中list的查找代价太过庞大，因此我们选用set这个数据结构，其采用红黑树，查找代价为nlogn。

因此我们首先保存excel中的url至硬盘，然后读取保存的set



之后我们进行判断，当url在url\_list中时就跳过该网站的爬取

