本周主要任务：豆瓣影评数据抓取

豆瓣影评分短影评和长影评

一、现已完成：

短影评的抓取及存储 （短影评是静态html，相对好获取）

长影评的部分获取（是动态js渲染，可以获取前两页，后面出现了编码错误，正在调试）

二、遇到的问题：

1、影评的页数较多，未登录之前只能查看部分页面。后面的部分，需要进行模拟登录。长影评获取代码调好之后，着重解决模拟登录问题（之前尝试用cookie登录，但是失败了）

2、豆瓣反爬虫的问题（昨天爬了两次，就被反爬虫了，心酸）

目前发现反爬虫主要是针对我ip进行了限制，需要用户登录之后才能访问。因此模拟登录是必须要做的。如果后续还有反爬虫的其他机制，可能需要做代理ip及其他反反爬虫的方法。

有相对方便的解决方法，反爬虫限制我ip的时间是几个小时，几个小时之后又可以爬取。可以选择等待几小时之后爬取一次（不安全）

三、后续必做及可选做的

1、模拟登录问题必做，否则即使没有反爬虫，数据获取也不全

2、反反爬虫的可选做，看进度决定

四、关于后续问题

数据存储问题与知识库构建的问题，暂时不清楚是怎么联系的。暂时是将影评的数据存储在txt文件中。解决好爬虫之后，需要阅读相关文献，查找相关方法。