题目一

关于视频相关作者基本信息，本来应该是要获取到每个视频的时间的，然后根据这些时间用pandas做为dataframe的索引，然后再用相关的时间函数去列出一个星期一个星期的顺序，然后再查看在这个时间内，相关的作者是否有发两个视频以上，如果有，就说明该作者是fate主题下的活跃作者，但是由于时间关系，我还没来得及做

题目二

关于弹幕，因为b站改版了，变成是seg.so文件，这个是一个二进制文件，就算我获取到了，也是一片乱码，还没遇到过这种情况，需要花费时间去了解才行，目前弹幕只能是获取到文本内容，对应的时间无法获取，可能是需要去配置一些相关环境才能对二进制文件进行转换，接着再去获取对应的文本内容以及时间

关于重点问题

第一点，我的思路是可以把代码挂在服务器里面，然后去检索新的内容，如果有新增的内容，则对数据进行保存，如果没有则每隔5分钟进行爬取一次

第二点，数据库保存，这个在第二题的代码里面有对应的数据库保存方法，如果出现错误，则触发回滚机制，

第三点，对应弹幕上限的问题，对待这种问题，分两步走，首先先了解每一个账号和ip地址，每日能获取弹幕的上限数量为多少，然后根据这个量去制定一个账号池和ip池，用计算器去计数当到达一定量的时候，切换账号的cookie，和ip地址即可，如果出现重复，可以用数据库去重的语句，或者pandas去重的语句即可

第四点，历史数据的爬取，就是通过获取视频的发布时间，然后和现在的时间，从而去构建这一整段时间的，时间列表即可

第三题和第四题

基本都完成了，写在代码里面

第五题

本来是计划去爬取10万+的评论信息，然后再去处理的

但是因为我是个人，只有一个账号，获取评论有上限，每当获取2万左右的数据B站那边的接口就会暂时给我停掉，然后就要等好久，才能去重新获取评论数据，本来已经获取差不多9万的数据了，但是导师你这边要收卷了，所以只能停了

对于日文，英文，中文的文本归类，我是这样的，首先只能对其文本进行打标签行为，

然后用贝叶斯的方法根据这些标签的数据，去进行建模，然后做一个文本分类的模型，从而去识别是日文，还是英文，还是中文，目前我能想到的方法这个是最好用的

然后关于正向，中立，负向，关键词topN的表现，可以用百度的paddle库，去对文本先用jieba库进行分词，然后再加入停用词，去掉无意义的词，接着再对高频词进行情感分析，找出它的情感倾向，从而根据它的高频词的数量和情感分数的表现，去判断正向，中立，负向，关键词top N的表现行为

模糊语境，这个没做过，不过我猜想应该是首先先找出中立的词，然后对应训练作为训练集，去训练出一个新的模型，然后使用snownlp这库，去对其进行训练模型，然后再通过训练好的库对其进行语义分析。