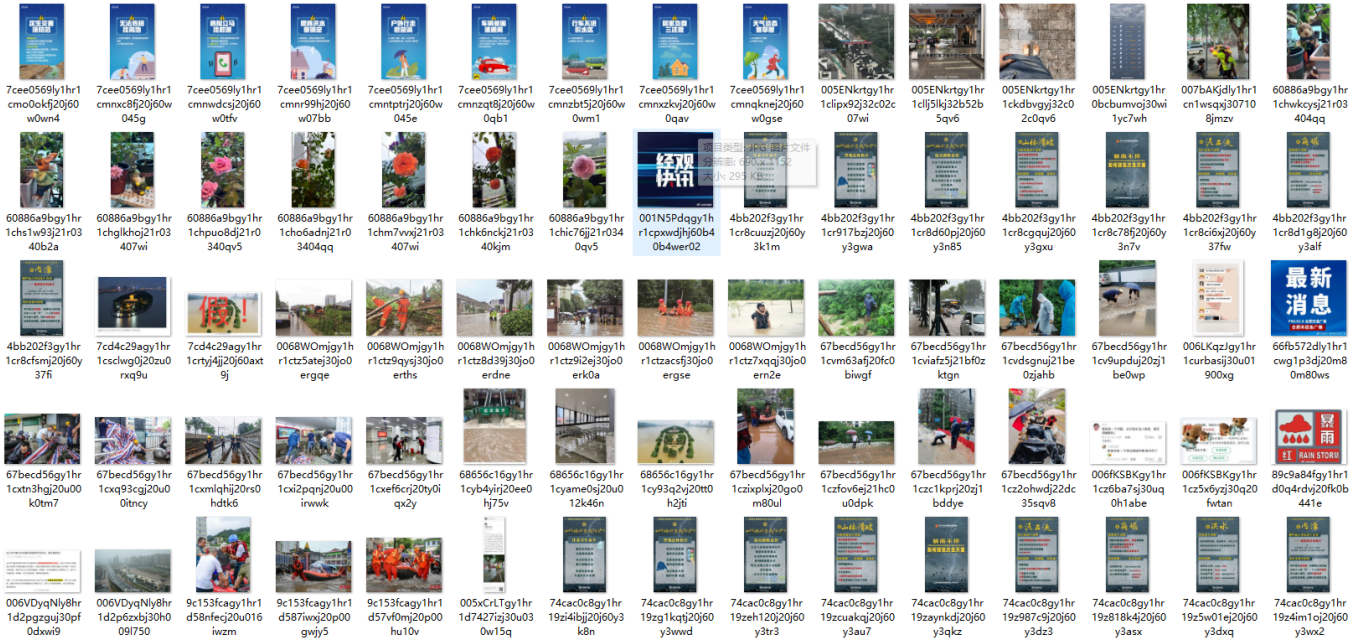


一、本周成果：

利用大模型识别路牌等信息对图片清晰度有较高要求，之前爬取的图片为缩略图，**分辨率较低**，尝试利用selenium获取高分辨率图片时被微博反爬，多次尝试无果后。改用pyautogui库结合selenium进行爬取，pyautogui能够模拟电脑鼠标键盘的操作，从而在一定程度上避免微博的检测，**最终实现了高清晰度图片的下载**，缺点是速度较慢



二、遇到问题：

1.高德地理编码/逆编码API虽然能够返回地理坐标，但需要的参数是结构化地理位置，如**省-市-区-街道**类似格式，而从爬取到的信息中地理位置往往是只有地名，如**北京师范大学**，此时返回会返回多个结果，可能会导致坐标的误差。

想法：目前尝试利用高德API，通过关键字返回精确的结构化地址，目前仍在尝试

2.目前选用的大模型为智谱清言，利用API能够识别图像，但需要的图片参数为url，不能对本地图片识别

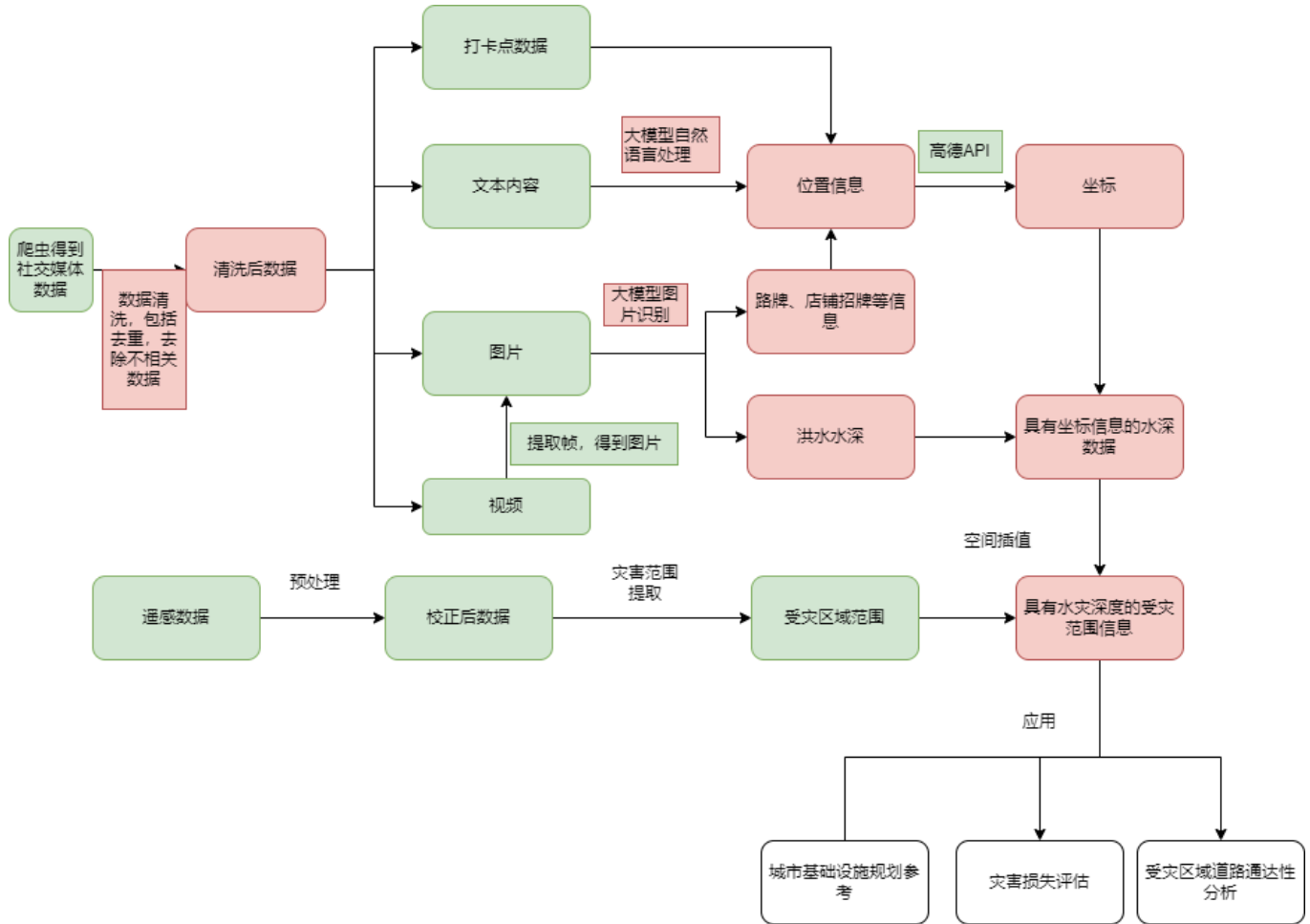
想法：尝试利用图床或搭建服务器解决。

3.目前经过测试已经爬取清晰图片7000多张，利用文件名称去重后还有5000多张，在剩余图片中仍有大量无关图片

想法：目前设想通过大模型对图片内容识别并清洗无关图片

4.爬虫过程中，虽然能够爬取图片，但是不能很好的将同一推文中的图片与文本内容想对应（疑似存在反爬措施，为避免被反爬，需要将内容分开派爬取）

实验流程（暂定）进度



绿色：基本实现或已有成熟方法

红色：仍待解决或效果不佳

对当前实验流程的总结：

1、目前主要问题就是大模型的识别能力，如果采用的大模型效果好，则就能够得到好的实验效果。

但是从测试过程中发现

- **图像招牌文字识别**对数据要求较高
 - 需要图像清晰
 - 招牌占图像比重高
- **水深识别精度较低**：虽然能够返回结果，但是精度太低

👤 用户_SjSqWR

图中洪水有多深



🌀 ChatGLM AI

图中洪水深度无法确定，但根据图片中的车辆部分被水淹没，可以推测洪水深度至少在1米以上。

•

用户_SjSqWR

图中洪水多深



ChatGLM AI

图中洪水已经淹没了路牌，水深无法确定。

用户_SjSqWR

参考图中行人推测水深

ChatGLM AI

图中洪水深度约为1米左右。



- **文本内容总结**不够精确且返回结果啰嗦，不确定性较大，利用代码流程化处理时较为麻烦，需要对模型进行调教

用户_SjSqWR

近日，湖南岳阳，多地因暴雨出现不同程度的险情。#洞庭湖决堤村民讲述紧急转移经过#@长沙消防 救援力量赶赴增援，立即集结五个排涝编队及战保分队分赴三塘镇、静河镇两地进行排涝工作，已分别在两个受灾点持续排涝16小时，累计排水约30万立方米。上面这段话中描述的是在哪里发生的事？

ChatGLM AI

复制入框

上面这段话中描述的是在中国湖南省岳阳市发生的暴雨导致的险情。具体提到的地点是洞庭湖以及三塘镇和静河镇。长沙消防的救援力量被派遣到这些地区进行排涝工作。

用户_SjSqWR

具体地址是哪里，越简洁越好

ChatGLM AI

湖南岳阳。

用户_SjSqWR

具体地址是哪里，回答简洁，位置越详细越好

ChatGLM AI

湖南岳阳

用户_SjSqWR

再详细一点

ChatGLM AI

湖南省岳阳市三塘镇、静河镇

总结：大模型效果还有待改善

2、数据爬取及清洗

- 当前以长沙6.24暴雨为例，以“长沙暴雨”、“长沙洪水”为关键字搜索结果，预计数据数量约为一万左右，其中主要数据集中于6.24日前后，理论设想大模型能够比较好的区分无关图片从而辅助数据清洗（或可尝试获取其他社交媒体数据如抖音等，增加数据量）
- 数据爬取过程仍存在平台反爬问题

下周任务

1. 尽量完成社交媒体数据的爬取及清洗工作
2. 尝试搭建图床，利用大模型识别
3. 尝试利用大模型总结文本内容