舆情分析项目小组总结

1. 项目数据

本次的项目跨度时间长达一个月，目前项目的完成度为百分之百，基本功能都已实现。

1. 出现的问题
   1. Flask和web前后端的连接问题。
   2. 在爬取数据的过程中的经常会被封ip
   3. ES存储数据的一系列问题
   4. 如何有效的提升模型的正确率
   5. 计算词向量方式的选择
   6. Keras训练出来的lstm模型与flask连接出现了多线程冲突问题
2. 改进和作用
   1. 查阅了大部分的资料，最终决定通过form表单和flask的request进行前后端数据的交互
   2. 加入了sleep的时间，以加长爬取时间为代价减少被封ip的可能性
   3. 对于无法分词的text形式的文段，以创建自定义索引来存储数据，通过加入个人自定义的停止词词典改善分词效果；对于插入数据的速度问题，通过查阅资料，使用bulk语句对数据进行批量插入和清除，快速的提高了插入速度
   4. 在项目的开始阶段，对于网上的数据集直接进行训练，利用knn，决策树，随机森林，svm等将近十种模型进行了训练，并对预测正确率进行了横向比较，发现正确率普遍不高。后来对于数据集进行了处理和重新标注，再次进行训练，并加入深度学习的形式（lstm模型）再付进行横向比较，最终选择的是正确率和f1分值都更胜一筹的lstm模型
   5. 刚开始采用snownlp计算词向量，后来发现模型的预测正确率不甚理想，经查阅资料后采用word2vec来计算词向量，然后正确率有了较大的提升
   6. 通过长时间的查阅和阅读文档，最终发现是flask的名称冲突导致的多线程问题，加入了threaded=False的属性解决了这个问题
3. 成功因素
   1. 组员之间的有效沟通：我们小组为了更好的针对本次项目进行有效的交流，为此专门建立了一个项目交流群。小组的每个成员如果有什么对于项目有什么不了解的地方，并且无法及时面对面交流的话，会在群里提出问题然后开始进一步的交流，最后统一看法，并且将统一的地方通过在线文档的形式记录下来，方便以后的查阅。
   2. 网上资料的帮助：由于是第一次做项目，对于项目没有什么了解， 在参考了多个相关项目之后才确定了做项目的基本流程。在做项目的过程中也时常会遇到无法处理的问题，我们通过参考csdn,博客园,github等多个平台进行了有效的解决。
   3. 老师的帮助：在做项目的过程中，时常会遇到一些思路上的问题，例如利用什么来存储大量的数据，并对他进行快速高效的处理和搜索。老师会在我们在做项目上的错误思路进行及时的纠正，帮助为我们的项目能快速的进行下去 。