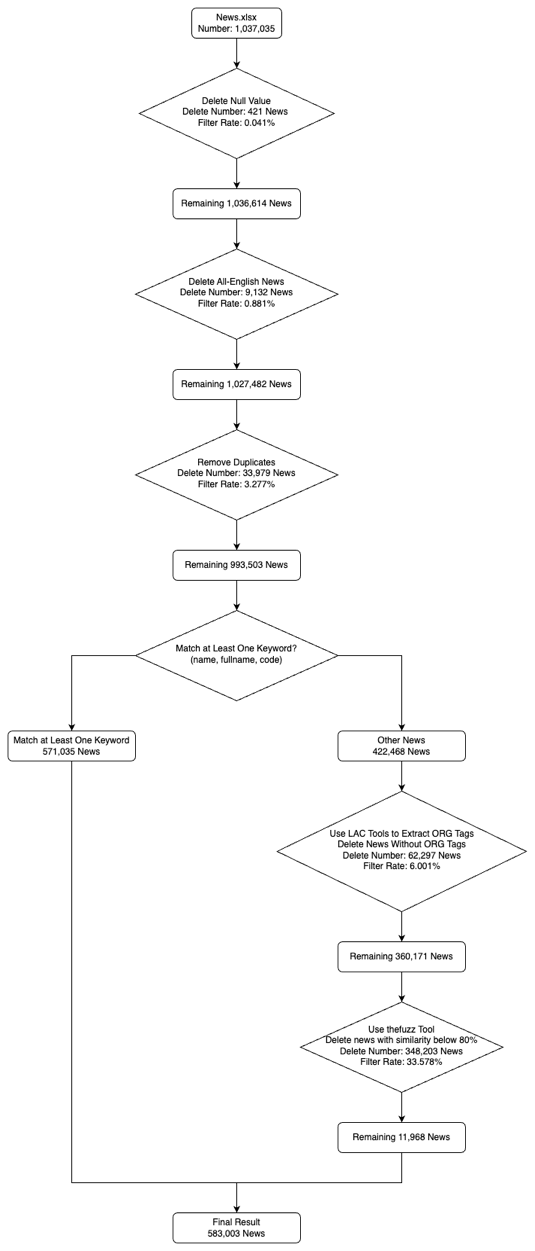
以上报告内容简短的描述了我的过程。如果想看到详细的过程，可以查看Code。里面使用markdown清晰的介绍了每一步代码。

Q1



以下是我数据处理的具体描述：

原始新闻数据：1037035

1. 删除NewsContent列
2. 删除空值

处理过程删除的数据条数：421

处理后剩余的数据条数：1036614

过滤率：0.041%

1. 删除非中文新闻

处理过程删除的数据条数：9132

处理后剩余的数据条数：1027482

过滤率：0.881%

1. 删除新闻中的特殊字符和数字
2. 对新闻进行去重

处理过程删除的数据条数：33979

处理后剩余的数据条数：993503

过滤率：3.277%

1. 通过正则表达式，删除A股列表中name的一些特殊含义字符，并将所有新闻按照是否出现了name，fullname和code中的至少一种，将数据集分为完全匹配和其他新闻

r'\\*?ST', r'^PT', r'^S', r'B股$', r'B$', r'A$'

完全匹配新闻数：571035

其他新闻数：422468

1. 使用百度的LAC库对其他新闻进一步处理，提取出ORG标签，删除没有提取出ORG标签的新闻

处理过程删除的数据条数：62297

处理后剩余的数据条数：360171

过滤率：6.001%

1. 使用thefuzz库对提取出的ORG标签和name分别计算文本相似度，删除相似度低于80%的新闻

处理过程删除的数据条数：348203

处理后剩余的数据条数：11968

过滤率：33.578%

1. 将得到的新闻与第七步中完全匹配的新闻进行合并，得到最终结果。

得到的最终数据数量：583003

Q2

在这一问中，我使用了IDEA-CCNL/Erlangshen-Roberta-110M-Sentiment模型，对得到的新闻进行情感分析。由于计算资源的限制，我将数据集切割为12份，并行计算情感后，合并为最终的结果。

Q3

Reference

百度LAC

thefuzz

IDEA-CCNL/Erlangshen-Roberta-110M-Sentiment