**基于舆情的指数预测思路：**

将舆情分为两大类新闻，一是可以量化的各种指数，另外一种则是普通新闻，进一步如下：

1. 指数新闻，从东方财富网爬取10种主要的指数数据（每个月更新一次），数据库存储格式：时间（Date），种类（Int），值（Float）
   1. CPI
   2. PPI
   3. PMI
   4. 房价指数
   5. 工业增加值增长
   6. 企业商品价格指数
   7. 消费者信心指数
   8. 货币供应量
   9. 外汇储备
   10. 财政收入
2. 普通文本新闻（从雪球网的24小时轮播获取）
   1. 公司相关新闻：该条新闻主要关注某个公司的状况，如业绩情况，财务报表情况，发展前景，股票预期收益水平，包括公司高层相关的新闻都可以包括在内
   2. 行业新闻：行业新闻可以影响该行业的公司股票，也会影响相关行业的前景
   3. 市场新闻：如证券公司相关的新闻
   4. 经济相关新闻：这块涉及的面很广，可以细化为：
      1. 央行相关新闻（汇率，外汇储备）
      2. 货币政策
      3. 楼市房价
   5. 政治新闻
      1. 国内政治新闻
      2. 国际政策新闻

对于指数新闻，我们直接量化为对应的值作为训练输入，对于普通文本新闻，先进行新闻分类，比如分为5大类，对于每一类求出当天该类新闻的平均情感强度作为训练输入，输出为每天的指数，通过递归神经网络学习对舆论和指数的涨跌值进行拟合。

**基于舆情的个股预测：**

相对于指数，个股受宏观因素的影响较小，另外由于股票数目的众多，为每一支股票进行预测不是很好的解决思路，我们可以基于关注度筛选出一些热门股票，基于雪球网对于各个股票的当天总关注人数和评论数目我们可以得到每只股票当天新增关注和评论数，加上股票的成交量等因素，我们可以用加权的办法获取前k只热门股票，用这k只股票的代码从雪球网爬取最近的新闻（跟改只股票紧密相关的）进行情感分析，结合其财务报表等信息预测未来的走向。需要增加两个数据表：

1. 每天各个股票的总关注人数和评论数（时间（Date），股票代码（String），股票名字（String），股票关注数（Int），股票评论数（Int）
2. 每天各只股票新增关注人数和评论数，scheme和1相同