1. 问题的分析与研究

首先训练数据集只是给了用户ID、博文标记（这一栏数据我还没准确理解它是什么意思）、转赞评、微博内容，这七类数据。预测数据集是除了转赞评其余都有。我们要做的事情就是预测用户发微博之后，一周内的转赞评。

首先的思路可以很自然地想到利用相关NLP算法，构造一些特征，然后运用机器学习建模。还有就是暴力的使用纯统计预测结果。

现在的思路整理是：

1. 使用用户的历史数据中的转发数的规律作为它未来发微博一周后的转发数能够到达的水平。
2. 对微博内容进行分析，相应的提一些特征。
3. 数据分析与预处理
4. 首先了解数据集的情况，用户的总数，微博总数，数据集时间跨度。找到需要预测的用户，从训练集中抽取出他们的历史数据（也应该会存在没有历史数据的用户）。
5. 互动数的分布情况，用户的分布情况。数据集中肯定存在大量的“无用”用户，他们可能微博数比较多，但是互动为0。这一类用户可以设置一个集合过滤，预测为最低档。不用参与提取特征的模型预测。
6. 没有历史数据的用户，可以直接处理为最低档，也可以根据用户发表的微博内容中的特征进行适当预测。
7. 特征工程

这个问题的特征从三个维度考虑：用户特征，微博特征。

1、用户特征

用户的特征主要包括活跃度，自身属性，基础属性。

活跃度：连续发微博的天数；连续不发表微博的天数；各个时间段内发微博的频率。

自身属性：所有微博中3无微博的条数和比例；

基础属性： 收到的转发，点赞，评论的总数，平均数，中位数，最大数；不同档位的微博数，占总微博数的比例；档位的众数

2、微博特征

要提取微博特征，首先要对微博内容进行预处理。这里预处理主要包括：分词处理(过滤停用词)，提取邮箱，hashtag，@的数目，URL，日期等。

微博特征主要包括基础属性，时间属性，文本属性

基础属性： 微博的长度，hashtag是否含有及数量，@的数量，url的数量

时间属性： 周几特征，微博所在时间段，今天是否是节假日，休息日，调休日

文本属性：这个需要进行 LDA 主题分析，统计该微博的 topic分布，词袋模型，微博发表后前后 1 分钟，15分钟，1小时，3小时内本人微博的相似度。

1. 模型构建
2. 模型调整
3. 评测模型与提交结果