本次赛方给我们的数据集分成两部分，一部分是50万已经标注好的数据用于训练，一部分则是450万的数据为未标注的数据用于最后的模型分类。在对项目50万数据进行初步分析后，我们发现数据集存在严重的数据不平衡的问题，50万数据中共有1258个标签，数据最少的标签仅有11条数据，最多的则有26835条数据。同时在50万数据中可能存在类别缺失的问题，即最后分类的450数据中可能存在50万数据中未有的标签。这将涉及到迁移学习的问题。后来经过赛题方确认后，450万数据包含的标签即为50万数据中包含的标签，这意味着不存在解决迁移学习的问题。最后需要解决的数据问题主要有数据不平衡问题，以及每一类数据量过少的问题。

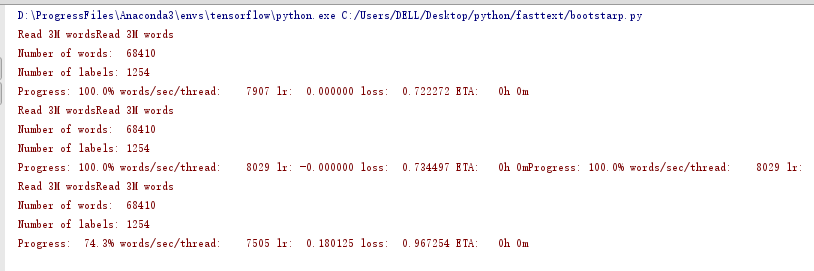
解决数据不平衡的策略大致可以分为两类，一类是从训练集入手，通过改变训练集的分布，从而降低数据不平衡的程度，即采样法，对于增加对稀有类别的数据采样，减少大类别数据的采样；一类则是从学习的算法入手，根据算法来解决相应的数据不平衡的问题，这方面的策略包括分类器集成，代价敏感学习和特征选择等方法。

采样法主要可以分为上采样，下采样和混合采样。上采样中随机上采样是最常使用的，即随机复制少数类的数据加入训练集，这会使得少数类中存在大量相似的数据，会造成训练集过程中过拟合的问题。同时若有少数类存在噪声数据，随即上采样会加大噪声对模型的影响。下采样中随机下采样与随即上采样思想一致，即从多数类的数据中抽出与少数类样本数量一致的规模，再将抽选出的样本与少数类的样本组合成新的平衡样本集。显然下采样会使得原属数据集样本减少，令训练时间更短，但是造成数据信息丢失情况。混合采样则是一种结合上采样，下采样的方法，其分类器的性能是优于单个采样方法，混合采样的主要思想是将随机下采样与SMOTE算法结合，即令多数类随机采样到一定的规模（采样后多数类的规模依旧大于少数类规模），去除多数类样本后和噪声后，使用SMOTE算法人工合成少数类样本，最后使用向量机（SVM）进行分类。但是这会造成多数类的样本被处于边界的少数类或者少数类的噪声”入侵”。

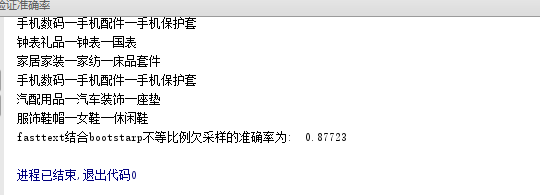
针对数据不平衡的算法方面，大致有基于聚类的挖掘算法，基于Boosting的挖掘算法，代价敏感的挖掘算法，基于核的算法等。基于聚类的挖掘算法的主要思想是按照数据内部的特征将数据集分成多个簇，令簇内的数据相似度尽可能高，簇间的数据相似度尽可能低。但是目前的基于聚类的算法虽然能够支持稀疏矩阵，但是通过算法产生的新样本数据集并不是特别的稀疏，即新样本无法很好的反映原样本的特征，故这方面的算法处理问题时很低效。基于Boosting的挖掘算法主要是将“弱分类器”提升为“强分类器”, 通过反复学习得到的一系列弱分类器组合得到一个强分类器。这类算法虽然具有较高的准确度，选择恰当的迭代次数即能获得好的性能，但是正是由于该类算法由许多的弱分类器组成，而且不支持并行化学习，会是的分类器学习速度缓慢，数据集不充分，选择的弱分类器不一致，均会导致分类器的性能下降。同时该类算法会增加错误样本的权重，较易令分类器受到噪声的影响。代价敏感的挖掘算法的主要思想是利用代价矩阵使不同的分类错误导致不同的惩罚力度。虽然相比采样方法具有较低的算法复杂度，但是代价矩阵通常是未知的，导致难以判断代价矩阵的值。基于核的挖掘算法，基于训练集在样本空间中找到一个具有最大间隔划分超平面将类别不同的样本划分开, 但是在线性不可分的情况中, 没有办法找到一个超平面能够将样本划分, 因此SVM利用核方法对样本空间进行高维映射来达到划分样本的目的。对于核方法的研究分为两种，一种是将核方法和采样方法结合起来，主要过程是将不平衡数据集分为少数类数据集和多数类数据集, 对多数类数据集下采样N次, 再将N次下采样后的多数类数据集分别同少数类数据集组成新的N组平衡数据集, 最后利用集成的方法将使用这N组新数据集分别训练的N个SVM分类器集成起来。一种则是核修正方法，即通过调整内部机制进而优化分类器性能的方法。但是基于核的挖掘方法依旧存在局限性：(1)当存在噪声数据时, 这些基于噪声而生成的数据并不可靠;(2)算法中使用的弱分类器性能各有差异，无法准确的确定每个分类器的权重。（3）该方法仅仅适用于二分类的问题，对于多分类的问题并没有解决方案。（4）该方法是基于向量机（SVM）的，在大规模样本中难以实现。

面对本项目中已标注的数据集数量过少，需要考虑到半监督学习的方法。传统的监督学习，通常需要用到大量良好的标记数据对模型进行训练，在处理高维度的问题时(例如：图像分类，语音分类)，监督学习中所需要用到的数据便会呈指数级暴增。由于实际生活中获取标记样本的成本往往较高，无标记的数据获取成本较低。半监督学习便有助于改善这一状况，其思想是在标记的样本数量很少的情况下，通过模型训练引入无标记样本来避免监督学习中遇到的训练样本不足而出现性能退化的问题。目前虽然通过半监督学习将有标记和未标记的数据进行结合，在某些模型和方法上取得效果，但是倘若未标记的数据本身存在偏差和错误，训练过程中可能由于相应的错误累加导致分类器退化。同时鉴于半监督学习过于复杂，时间需求较长，因此在本次比赛中我们放弃了半监督学习的方法。

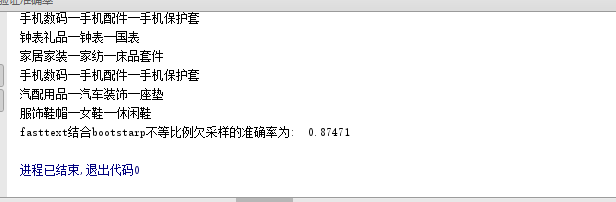
最终我们选择了词向量匹配的方法，即将相应的词向量转换为句子向量进行最后的分类。目前网上已经拥有许多公司公布出来已训练好的词向量，鉴于这些词向量均由海量的数据训练而成，因此并不会存在样本不均衡的问题，同时数据量足够丰富，由此而来的半监督学习的问题也一并解决了。

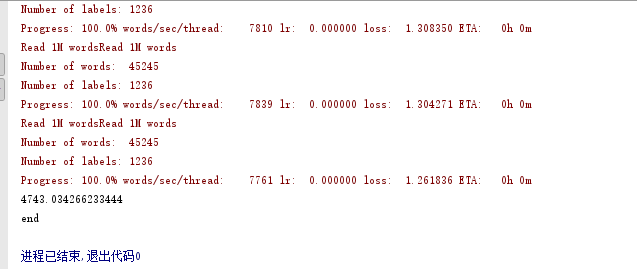


10份重采样

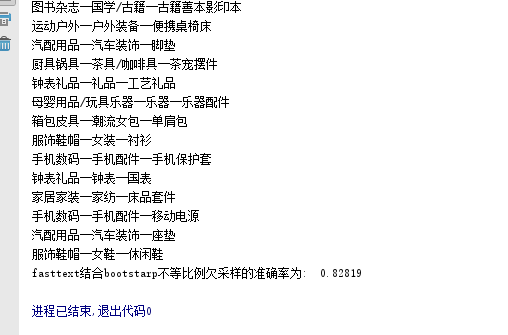


四十折份重采样





四十折



80折

