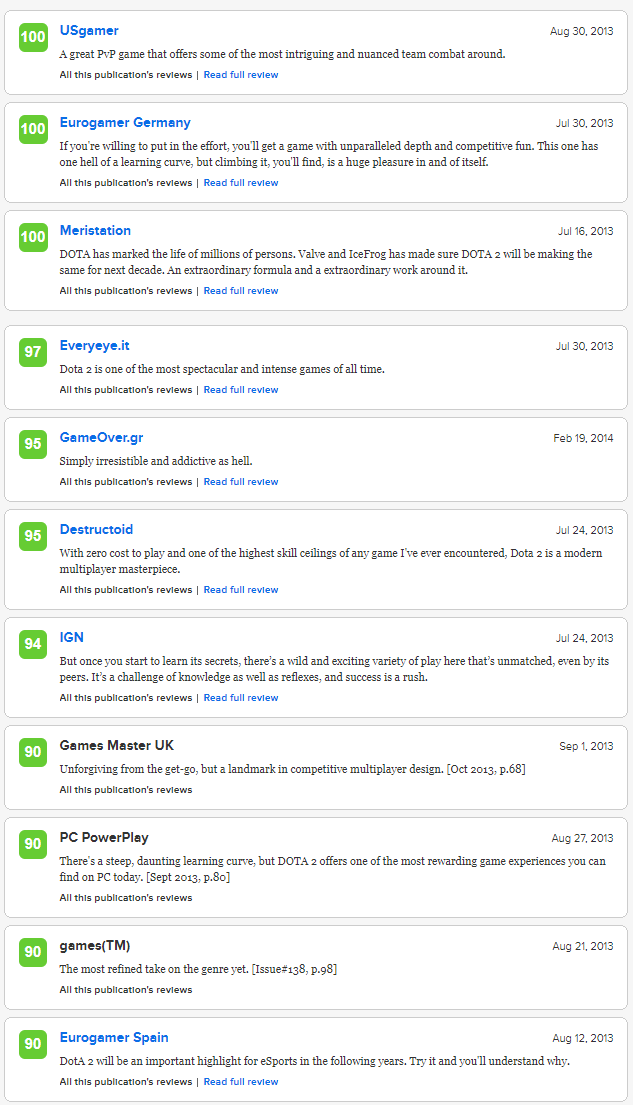
基于自然语言处理的游戏媒体评测可信度分析与预测

摘要：该实验爬取了metacritic网站中的游戏评测文本，探究评测文本的可信度并训练预测模型。本文由经典的二分类问题文本情感分析入手，探究了适合进行文本二分类问题的数据集和机器学习模型。接着利用多项式回归对数据集进行可信度标签。并可以借助更多与游戏本身相关的数据带入模型提高预测准确率。

1. 数据集与数据探索
   1. 数据集介绍

Metacritic。媒体评测文本数据来源。

IMDb（互联网电影资料库）。以电影评论数据集作为参照集。

Steam游戏市场。获取与游戏本身相关的数据。

* 1. 数据探索

问题简化：将可信度分析转化为一个二分类问题——游戏实际表现比媒体评测更优秀/更糟糕

从典型的NLP二分类问题：文本情感分析入手，对比不同模型在不同数据集下对于二分类问题的表现。

1. 数据建模

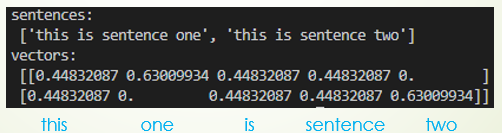
2.1 两类模型尝试：

**Tf-idf+逻辑回归**

字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。

实现方式：sklearn.TfidfVectorizer & sklearn.LogisticRegression

参数设置：max\_features (LogisticRegreesion参数默认)

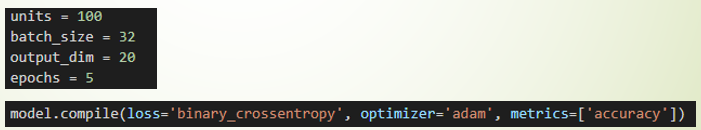


**Word2vec+CNN/LSTM/全连接层**

可以利用更加复杂的神经网络模型解释数据，CNN能有效提取局部特征，LSTM可以体现单词的上下文关系。

实现方式：keras.Sequential & keras.layers

参数设置：损失函数：binary\_crossentropy，优化器：Adam，等等



2.2 数据清洗方式

每条评论的文本和评分作为一个样本

去掉所有的标点符号，仅保留字母和数字

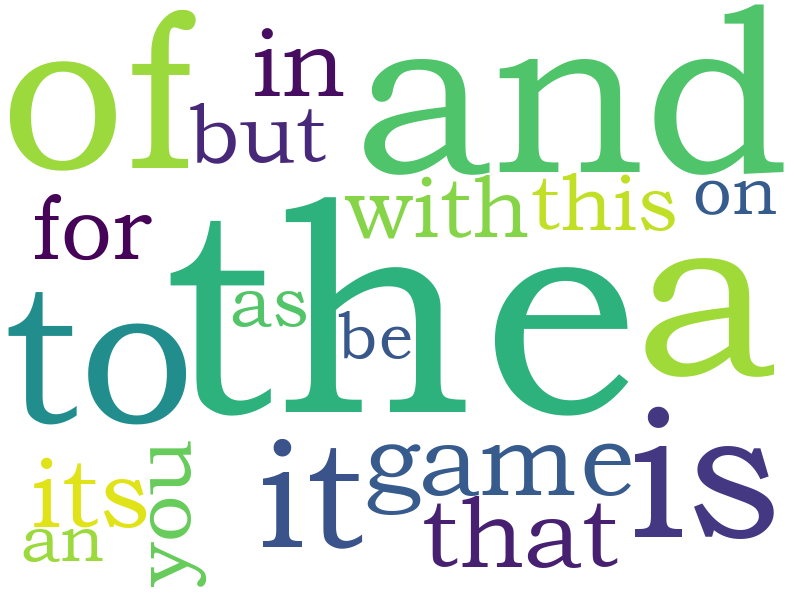
将大写字母转化为小写字母

去除出现频率最高的20个单词

去除仅出现一次的单词

将数据集中评分前25%标记为正样本，后25%标记为负样本

训练集：验证集：测试集 = 6：2：2



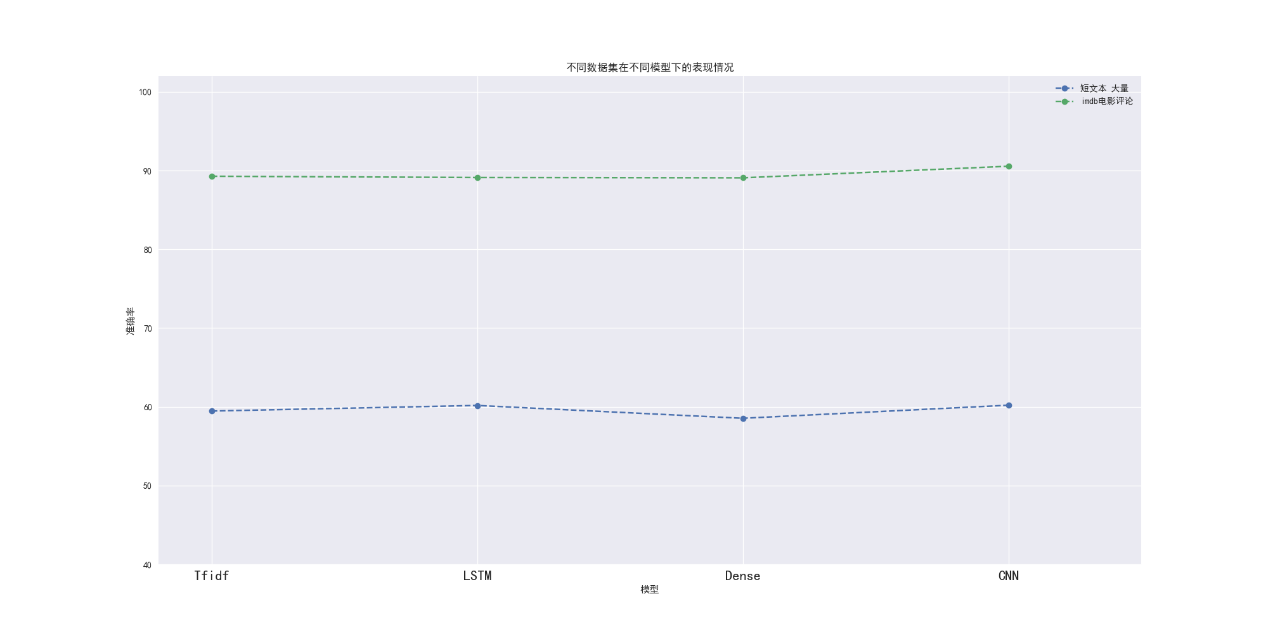
文本中出现频率最高的20个单词

2.3 数据集比较和分析

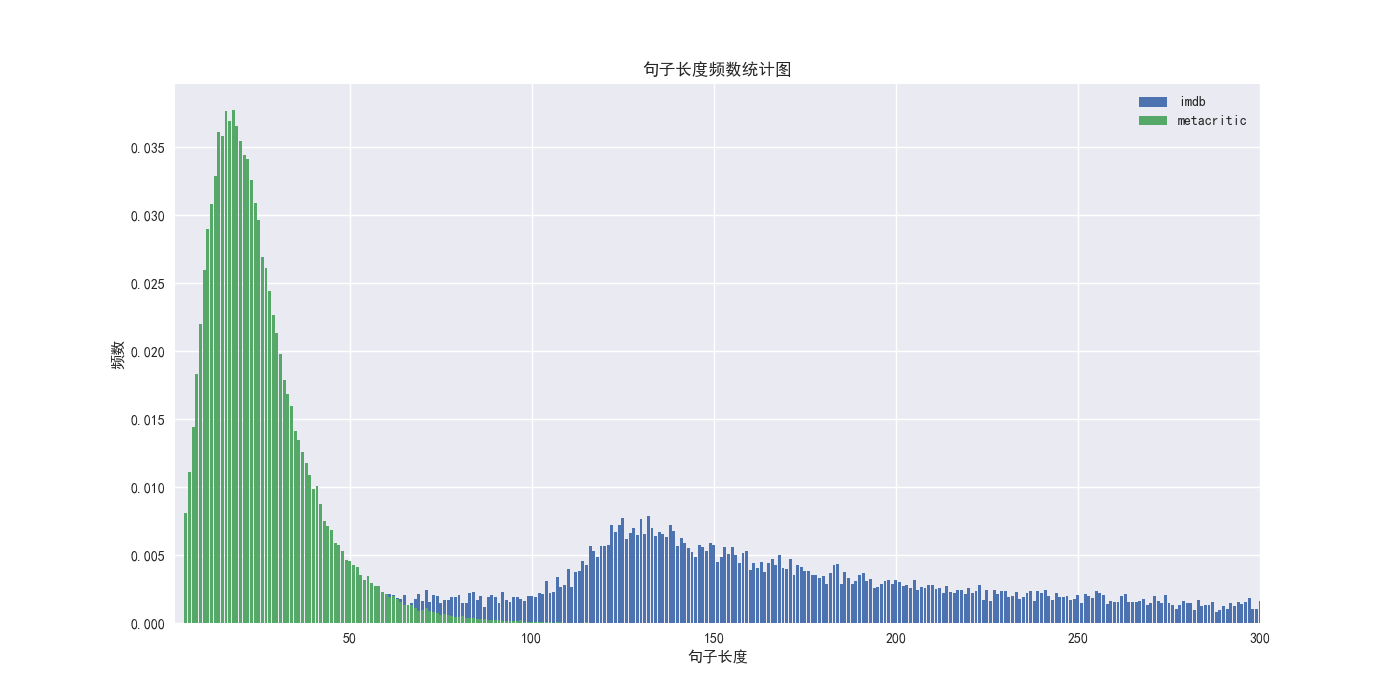
该实验主要尝试了四类数据集：

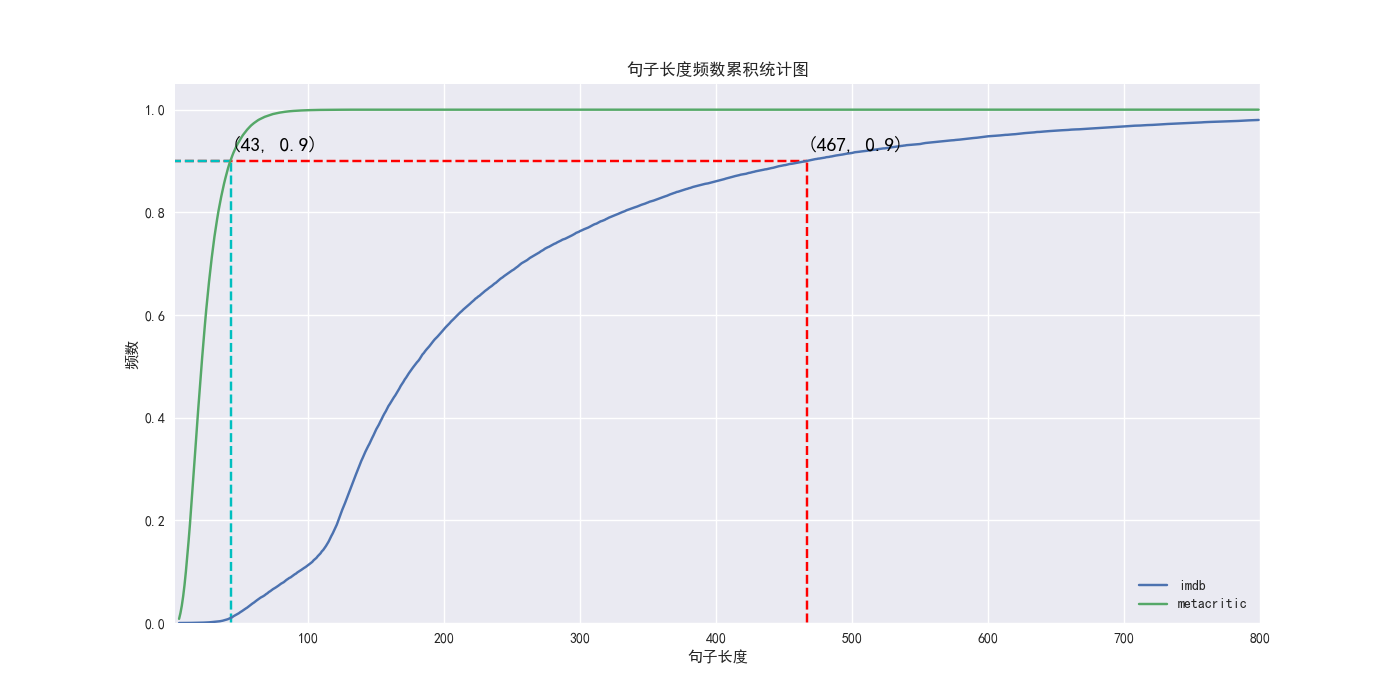
1. 短文本+大量样本
2. IMDb电影评论数据集
3. 长文本+少量样本
4. 长文本+大量样本

其中，IMDb电影评论数据集为参照集，用于比较不同数据集的表现情况。



从图中可以看到自己爬取的数据准确率只有60%，而IMDb数据集准确率接近90%。接着对文本长度进行了比较。



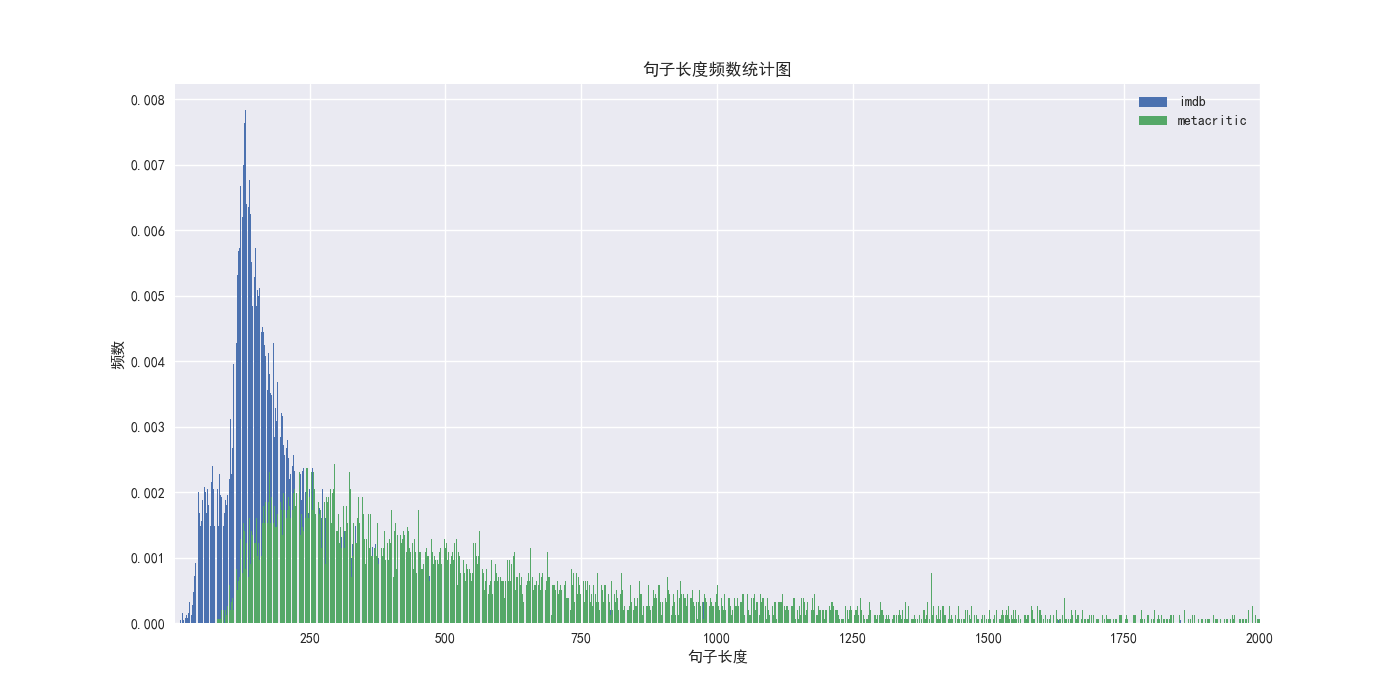


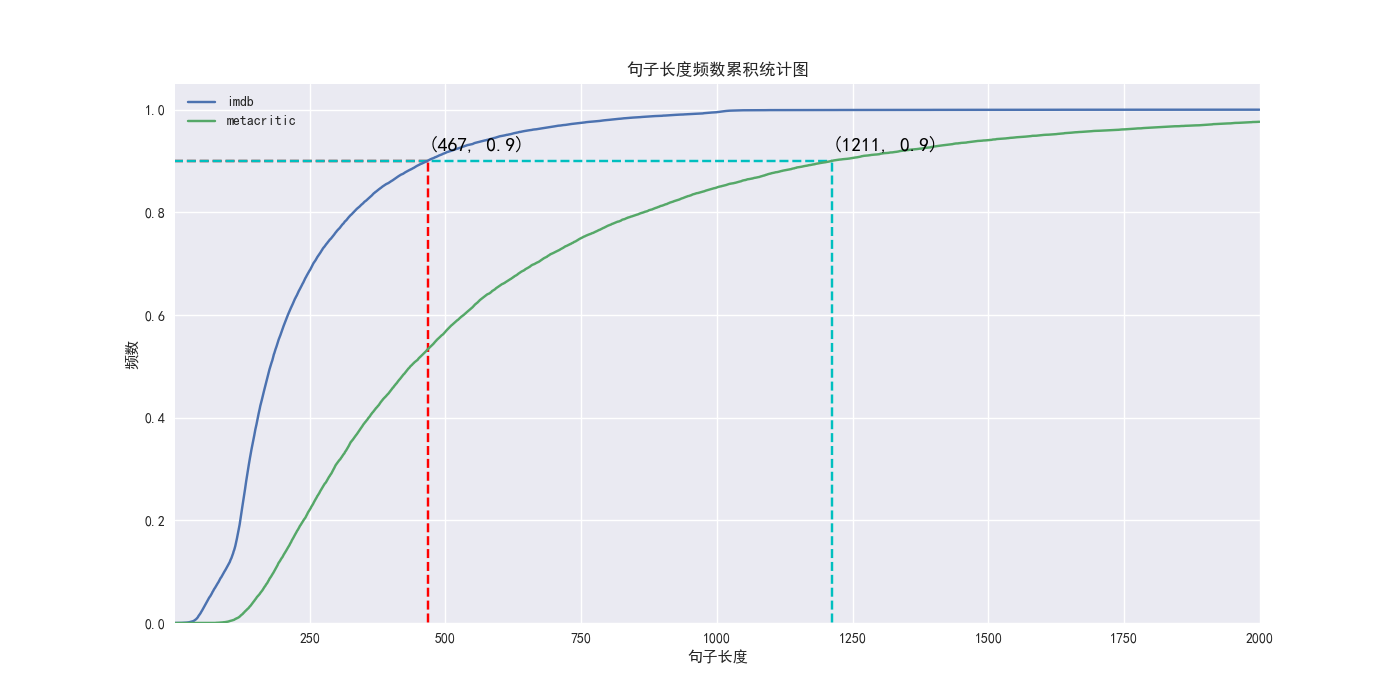
发现自己爬取的数据集句子平均长度只有IMDb电影评论的十分之一不到。

认为效果不佳很可能是因为句子长度太短

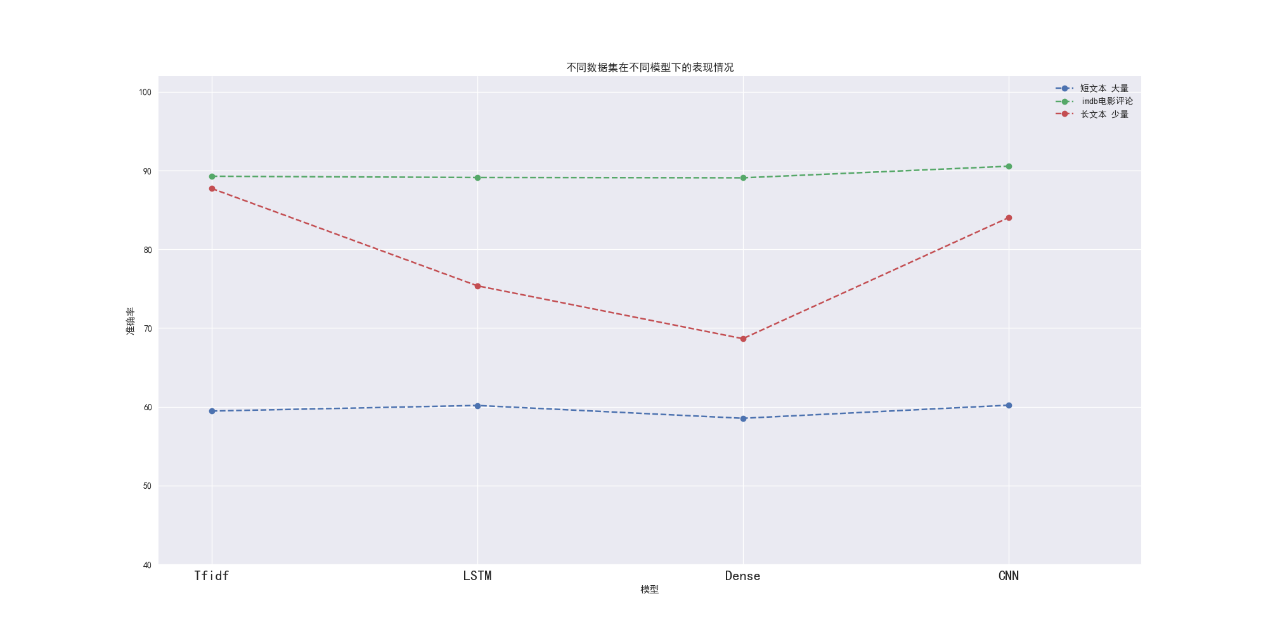
改进方式：将同一个游戏下的多条评测汇总成一个样本，评分取每个评测的均值

改进后的数据集情况：





训练情况：



汇总后的数据集虽然每句话长度显著增加，但是样本量也相应变小（2000条左右）

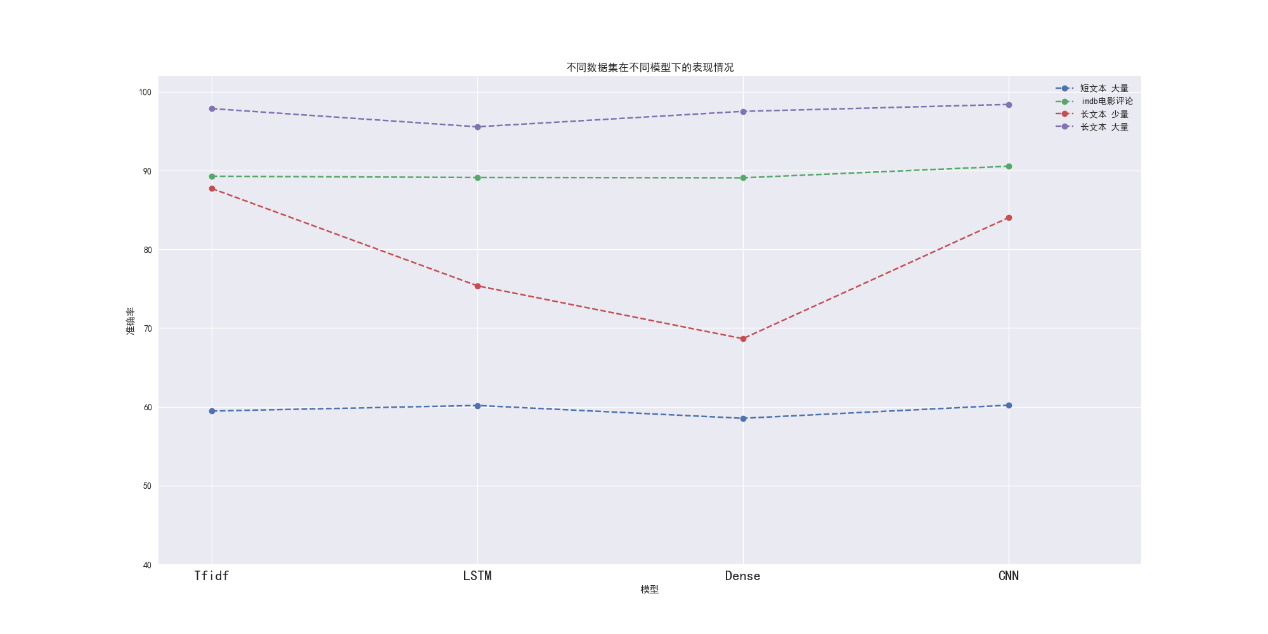
导致机器学习效果不稳定

首先体现在不同模型的效果差异上，Tfidf和CNN模型在样本量较小的情况下表现相对优秀，而LSTM和全连接层则明显较差。（或许说明越复杂的模型对数据量的要求也越大）

其次，即使模型和参数不变，多次运行同一模型最终得到的效果也不稳定，尤其是LSTM，准确率相差可达20%

改进方式：通过爬取更多数据的方式增加数据量

最终训练情况：



最终得出结论，自然语言处理解决二分类问题的数据集要求：文本长度>=400 样本数量>=10000。

决定用长文本+大量数据作为最终可信度预测的数据集。

2.4 数据标签

多项式回归拟合曲线pred\_score。

越接近曲线说明媒体评分与实际越相符。

引入参数 偏差值 x = score - pred\_score。

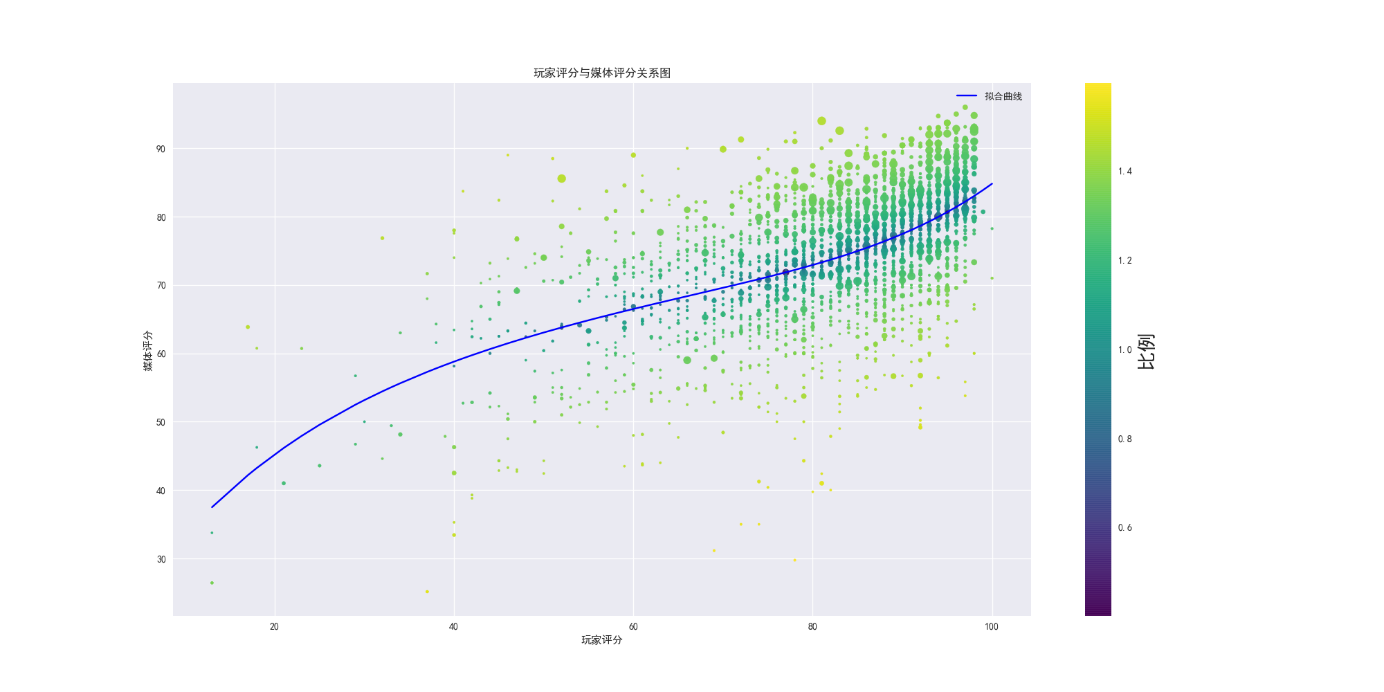
x越大说明游戏越被高估，x越小说明游戏越被低估。

移除值绝对值较小的x。

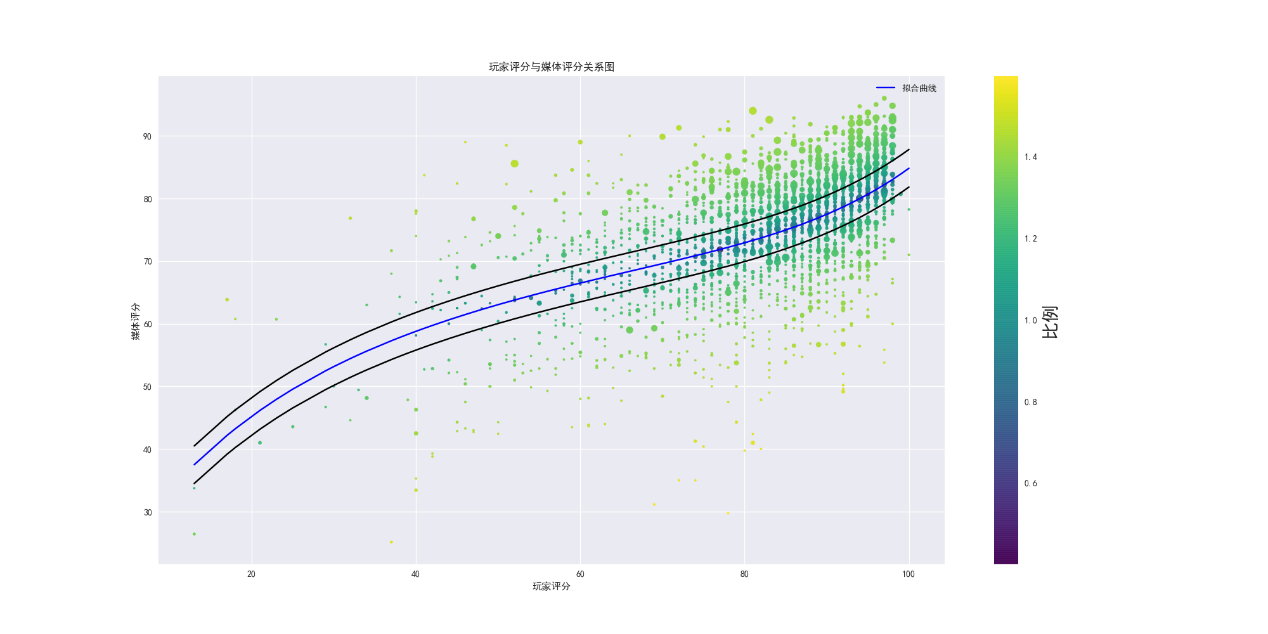
根据x的正负进行标签。

可视化过程如下：

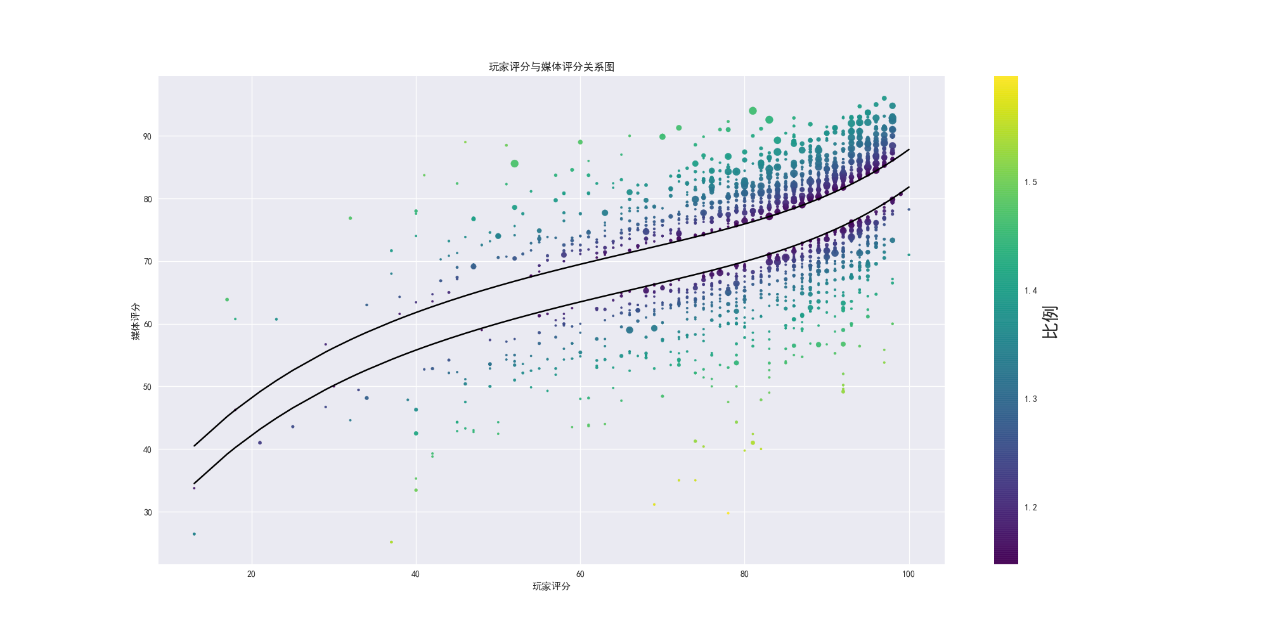
1. 多项式回归拟合均值曲线



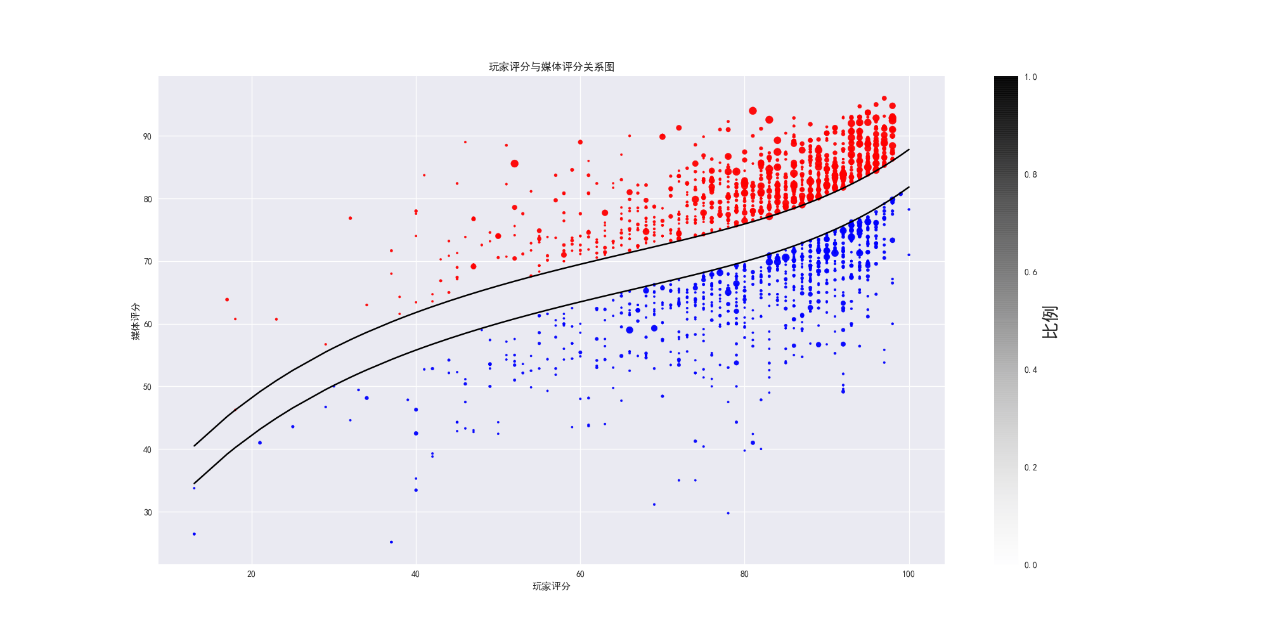
1. 将均值曲线上下平移一定单位，画出两条黑线



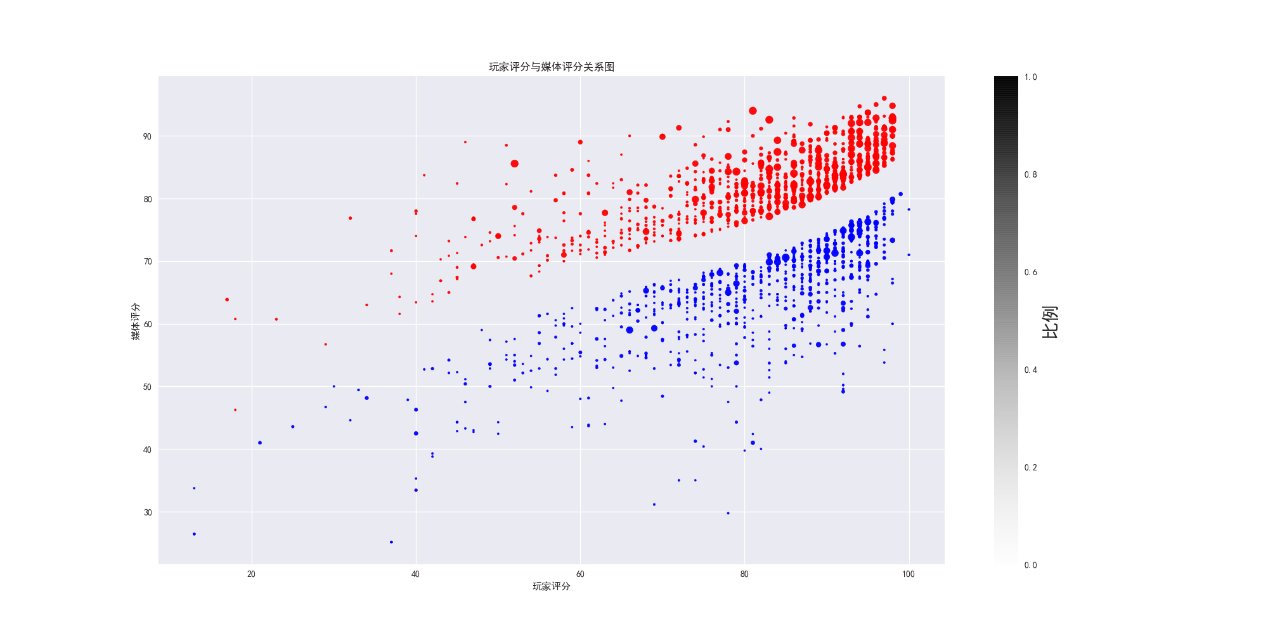
1. 移除两条黑线之间的数据点



1. 对上方和下方剩余的数据点分别进行染色

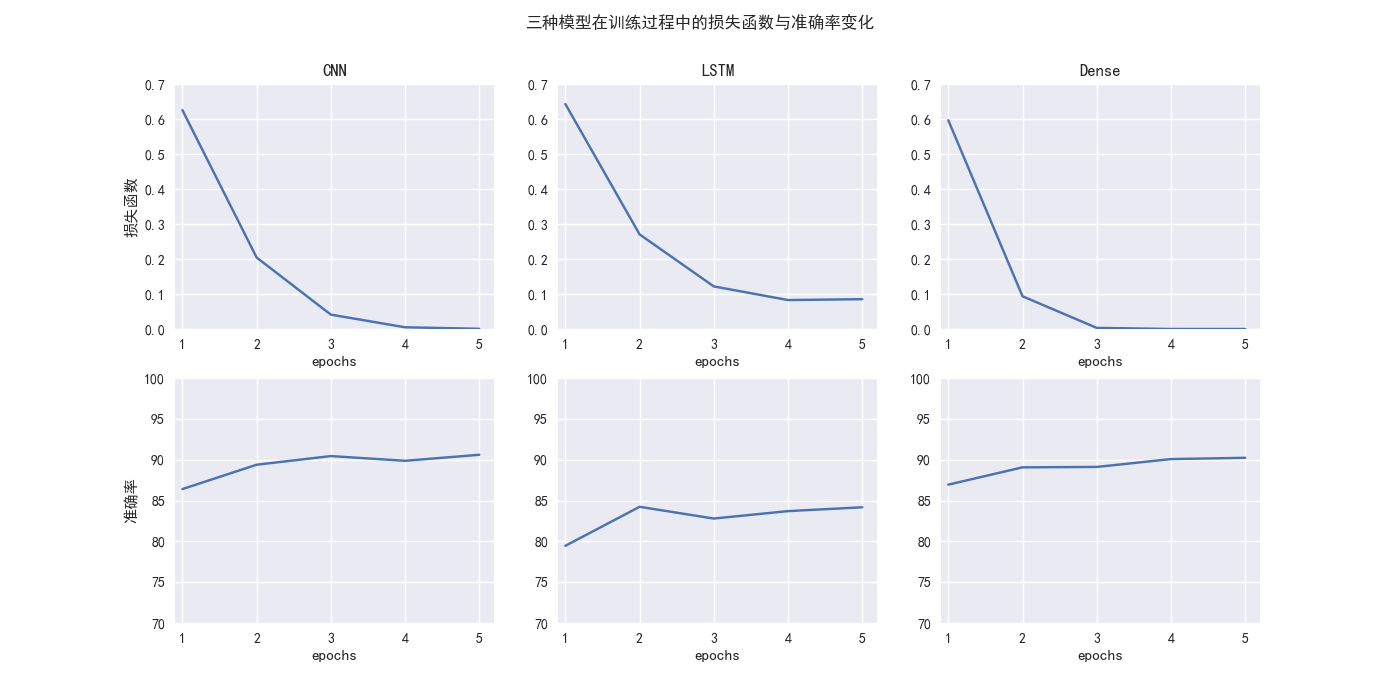


1. 得到最终标签后的数据集



2.5 训练结果

有了标签和文本后，我们把可信度预测问题也转化成了一个二分类问题，将数据带入不同机器学习模型后得到的结果如下图所示。



2.6 结合更多相关数据

媒体评测可信度不仅在于评测文本本身。

评分的高低以及游戏的定价也是相关变量。

把根据文本生成的可信度结果，与评分和定价一起带入逻辑回归模型进行预测。

由于可供训练的数据集较小（约1000条），训练表现不够稳定。

但可以看到整体准确率得到进一步提升。









三、 结果和分析

得到了一个根据媒体评测文本预测可信度的模型

如果有游戏本身的评分和定价等信息，可以带入模型进一步提高准确率

该项目重点在于探究怎样的数据集更适合进行机器学习，而不是机器学习模型和参数选择上。通过一系列实验可以发现，在机器学习中，数据集本身的质量对结果的影响远远大于模型和参数的选择。

四、 思考和拓展

该实验还可以在许多方面进一步优化和拓展。例如可信度标签方式，或者将二分类问题变成一个量化评估的问题。

同时，也可以将该可信度分析方式扩展到更多商品类评测或评论上。