哈尔滨工业大学（深圳）

Harbin Institute of Technology (Shenzhen)

实验名称：基于朴素贝叶斯（或其它方法）的文本分类算法

姓名：罗千原

学号：2023311206

日期：2025.10.31

1.实验内容

1. 理解文本分类的基本概念和任务；

2. 学习如何使用朴素贝叶斯算法或其他方法进行文本分类；

3. 实现一个文本分类模型，能够对给定的新闻标题文本进行分类；

4. 分析实验结果并评估模型的性能。

2.实验方法

1. 数据准备与数据预处理

a. 加载和预处理数据集：通过 load_data 函数加载 train.txt、dev.txt、test.txt 数据集，自动适配空格、制表符等分隔格式，过滤空行与无效格式数据；使用 jieba 对新闻标题进行分词，过滤长度≤1 的词、纯数字及标点符号，完成文本预处理。

b. 特征表示形式转换：将分词后的文本转换为词列表，后续通过朴素贝叶斯算法的词频统计与概率计算，实现文本到分类概率的特征映射。

2. 构建模型与训练

a. 算法选择：选择朴素贝叶斯算法，基于 “特征条件独立假设”，通过贝叶斯公式计算文本属于各类别的后验概率，实现文本分类。

b. 模型训练：遍历训练集，统计每个类别下的样本数、词频、总词数等核心参数，结合拉普拉斯平滑处理零概率问题，完成模型参数学习。

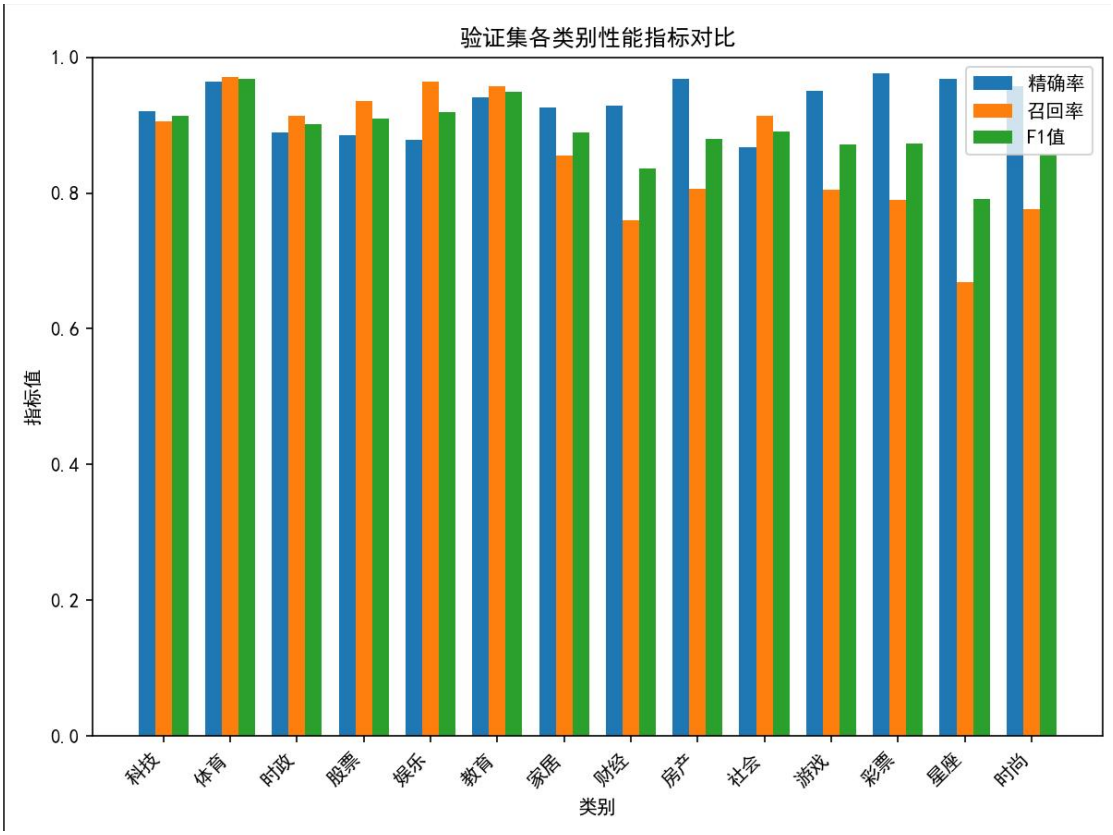
c. 训练过程控制：无显式训练轮数和批量大小（朴素贝叶斯为生成式模型，训练为统计参数过程），通过逐行遍历训练集完成参数统计。

3. 模型评估

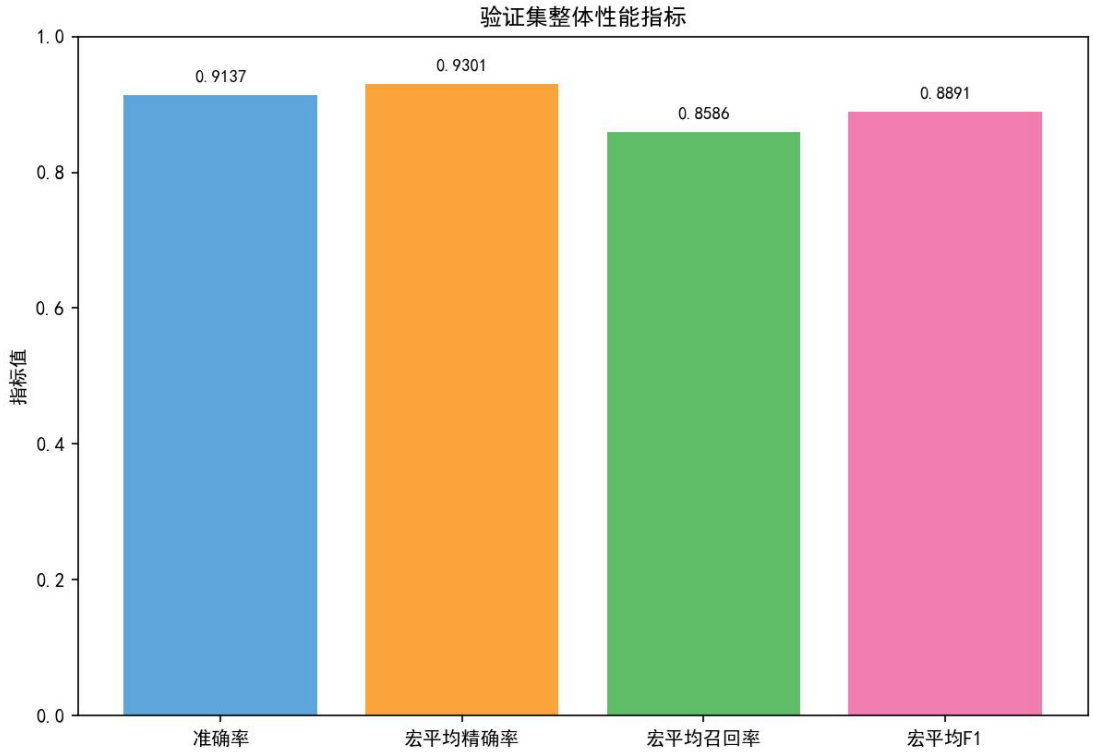
a. 评估指标选择：使用准确率、精确率、召回率、F1 值（宏平均）作为评估指标，从整体和类别维度全面评估模型性能。

b. 评估与分析：使用验证集和测试集进行预测，计算各类评估指标，结合可视化图表分析模型在不同类别上的性能差异，如体育、科技类表现优异，星座、游戏类存在改进空间，进而探究特征优化、类别平衡等潜在改进方向。

3.实验结果和分析



```
测试集指标：
准确率：0.8989 | 宏平均F1：0.8673
类别详情：
类别          精确率    召回率    F1
科技          0.9058    0.8970    0.9014
体育          0.9569    0.9655    0.9612
时政          0.8716    0.8962    0.8837
股票          0.8763    0.9293    0.9020
娱乐          0.8598    0.9466    0.9011
教育          0.9269    0.9430    0.9349
家居          0.9026    0.8290    0.8642
财经          0.9162    0.7457    0.8222
房产          0.9504    0.7735    0.8529
社会          0.8364    0.8896    0.8622
游戏          0.9253    0.7554    0.8318
彩票          0.9783    0.7500    0.8491
星座          0.9685    0.6324    0.7651
时尚          0.9235    0.7209    0.8097
```



整体性能：测试集准确率达 0.8989，宏平均 F1 值为 0.8673，说明朴素贝叶斯模型在新闻标题分类任务中表现较好。

类别表现：

- 体育、科技、教育等类别表现突出，如体育类精确率 0.9569、

F1 值 0.9612，科技类 F1 值 0.9014，教育类 F1 值 0.9349，这些类别标题特征词明确，模型易区分。

- 星座、房产、游戏等类别表现稍弱，如星座类 F1 值 0.7651，房产类 F1 值 0.8229，游戏类 F1 值 0.8318，可能因这些类别标题特征词模糊或样本量较少导致模型区分难度大。

- 图表辅助分析：验证集各类别性能对比图直观呈现了不同类别在精确率、召回率、F1 值上的差异；整体性能指标图则清晰展示了模型在准确率、宏平均精确率 / 召回率 / F1 上的表现，进一步验证了模型的有效性与类别间的性能差异。

4.总结

本实验基于朴素贝叶斯算法完成了新闻标题文本分类任务，整体流程涵盖数据预处理、模型训练与评估三个核心环节。

在数据处理上，通过自动适配分隔格式、`jieba`分词与无效词过滤，将原始文本转换为模型可处理的词列表形式，确保了数据的有效性。

模型训练阶段，利用朴素贝叶斯的概率统计特性，统计类别样本数、词频等核心参数，结合拉普拉斯平滑解决零概率问题，完成模型构建。

评估结果显示，测试集准确率达 0.8989，宏平均 F1 值 0.8673，整体分类效果良好；但不同类别表现存在差异，体育、科技等类别因特征词明确性能突出，星座、游戏等类别因特征模糊或样本量少表现稍弱。

综上，朴素贝叶斯算法在该文本分类任务中具备较好实用性，后续可通过特征优化、类别平衡等方式进一步提升模型性能。