新闻主题分类

一. 竞赛背景 (Background)

恭喜大家完成了深度学习的基础课程！现在到了检验真理的时刻。

本次大作业采用 **Kaggle 竞赛模式**。我们将处理真实世界中的“**新闻主题分类**”问题。这类技术广泛应用于内容推荐系统、舆情监控和信息检索等领域。

你的任务是构建一个**深度学习模型（如 RNN, LSTM, Transformer 等）**，教会计算机“读懂”一段中文新闻文本的核心主题，并在一个标签完全未知的测试集上达到最高的准确率。

1.1 任务描述 (Task)

你需要区分以下 **10** 不同的新闻主题类别：

序号	中文主题	英文对应
0	财经	Finance
1	房产	Realty
2	股票	Stocks
3	教育	Education
4	科学	Science
5	社会	Society
6	政治	Politics
7	体育	Sports
8	游戏	Game
9	娱乐	Entertainment

二. 数据集说明 (Data Description)

A. 训练集 (train/)

用途： 这是用于 训练 模型参数的主要数据。

规模： 约 100,000 条文本。

B. 验证集 (val/)

用途： 用于在训练过程中对模型进行实时验证（或称 调优），以监控模型是否发生过拟合，并帮助选择最佳的超参数和 Epoch。

规模： 约 10,000 条文本。

C. 测试集 (test/)

用途：提供给模型进行预测的**原始文本输入。**

规模：约 10,000 条新闻文本。

三. 评价标准 (Evaluation)

本次竞赛的核心指标是 **准确率 (Accuracy)**。

$$\text{Accuracy} = \frac{\text{预测正确的样本数}}{\text{总测试样本数}}$$

四. 提交要求 (Submission)

你需要提交两个文件：

由于测试集文件 (test.txt) 中不包含 label 列，您的最终提交文件必须遵循以下严格要求

1. 预测结果 (张三 2025110225.csv) 使用姓名+学号

文件必须包含两列：text 和 label。

- text: 对应测试集中的文本
- label: 预测的类别 ID（整数 0-9）

2. 项目报告 (PDF) —— 评分重点！

仅仅跑通代码是不够的，你需要展示你的思考过程。报告需包含：

数据概览与分布:

统计各类别样本数量，并以**可视化图表**形式展示主题分布的平衡性。

展示每个主题的**典型文本样本。**

文本预处理流程:

详细说明您采取的清洗步骤（如去除 HTML 标签、特殊符号、低信息量字符等）。

分词策略：您选择的中文分词工具（如 jieba 或 THULAC）及词典加载情况。

是否进行了去停用词、低频词过滤？

说明将文本转化为数值向量的方法。

说明词向量维度和最大文本长度（序列截断/填充策略）。

结果分析:

展示**训练集** Loos 与**验证集** Loss 随 Epoch 变化的曲线图
展示**混淆矩阵** (Confusion Matrix): 模型最容易把哪两个类别搞混？
分析其**误判类别原因**