项目设计书：唐诗生成器

设计人：罗乐恒

1. 项目背景

唐诗作为中华文化的瑰宝，拥有丰富的文学价值和历史意义。本项目旨在利用机器学习技术，开发一个唐诗生成器，通过学习唐诗的语言模式，生成具有韵律和意境的诗歌。数据源来自于 GitHub 的唐诗宋词数据库（https://github.com/chinese-poetry/chinese-poetry）。

2. 项目目标

- 使用唐诗数据库训练一个 LSTM 模型，生成符合唐诗风格的诗句。
- 探索文本生成模型在古典诗词创作中的应用。

3. 数据集构建

3.1 数据来源

- 数据源：https://github.com/chinese-poetry/chinese-poetry
- 数据格式：JSON 文件，包含作者、正文、标题、ID 四个部分。本项目仅使用诗词正文进行训练。

3.2 数据预处理

1. 读取所有 JSON 文件，提取诗词正文，合并成一个字符串用于训练。
2. 使用 Keras 的 Tokenizer 对字符进行编码，设定最大词汇量为 10000 个字符。
3. 将诗词分割成训练样本，每个样本包含一个长度为 20 的子串和下一个字符作为标签。

4. 模型设计

4.1 模型框架

- 使用 Keras 构建一个基于 LSTM 的序列生成模型。
- 模型结构：
    - Embedding 层：将字符编码映射到 128 维度的向量空间。
    - LSTM 层：两层 128 单元的 LSTM 网络，添加 Dropout 防止过拟合。
    - Dense 层：输出层使用 softmax 激活函数，预测下一个字符的概率分布。

4.2 模型参数

- 最大词汇量：10000
- 输入序列长度：20
- 嵌入维度：128
- LSTM 单元数：128
- Dropout 率：0.2
- 优化器：Adam
- 学习率：0.001
- 损失函数：categorical_crossentropy

5. 训练与验证

5.1 数据集划分

- 使用 sklearn 将数据集划分为训练集和测试集，比例为 7:3。

5.2 训练过程

- 使用批量大小 32，训练 40 个 epoch。
- 保存训练过程中最佳模型及其 Tokenizer。

6. 结果展示与分析

- 模型能够生成具有唐诗韵律的诗句，但在内容逻辑和意境上还有提升空间。
- 未来可以尝试更复杂的模型结构，增加训练数据，提升生成诗句的质量。

7. 主要贡献点与解决难点

- 贡献点：
    - 提供了一个基于 LSTM 的唐诗生成模型，为古典诗词的数字化创作提供了新思路。
    - 展示了从数据预处理、模型训练到结果生成的完整流程。
- 解决难点：
    - 如何处理唐诗的特殊格式，进行有效的特征提取和编码。
    - 在生成过程中保持诗句的韵律和基本语法正确性。

实现功能与缺陷：

实现根据前两句生成诗篇（当输入的内容不足两句时，仍能生成，但可能影响语篇连贯

性）

实现在生成结束后调用 opencc 库将繁体转化为简体

缺陷：

1.无法指定诗人的风格和诗体

2.会出现括号等标点符号（注：我们尝试利用正则表达式在生成后删除，但这也导致 了有时生成字数与要求字数不符等问题）

2024.8.10 罗乐恒