隐马尔可夫模型实验

概述

- 利用隐马尔可夫模型进行中文语句的分词。

数据说明

- 数据集是人民日报 1998 年 1 月份的语料库，对 600 多万字节的中文文章加入了词性标注以及分词处理，由北京大学开发，是中文统计的常用资料，可以在语料库基础上构建词典、进行统计、机器学习等。
- 数据集被划分为训练集和测试集，分别存储在 data 文件夹中的 train.txt 和 test.txt 中。其中，训练集中的语句已完成分词。

实验内容

- 中文信息处理是自然语言处理的分支。和大部分西方语言不同，书面汉语的词语之间没有明显的空格标记，句子是以字串的形式出现。因此对中文信息进行处理的第一步就是进行分词，将字串（character string）转变成词串（word string）。
- 依据字在词语中位置，为每个字赋予不同的状态（如句子开始、句子中间、句子结尾、单字成词等），将输入的中文句子转化为状态序列。
- 在训练集上统计语料信息，训练隐马尔可夫模型，对测试集中的中文句子进行分词测试，并选取部分实验结果进行分析。
- 基于 MindSpore 平台提供的官方模型库，对相同的数据集进行训练，并与自己独立实现的算法对比结果（包括但不限于准确率、算法迭代收敛次数等指标），并分析结果中出现差异的可能原因。
- （加分项）使用 MindSpore 平台提供的相似任务数据集（例如，其他的分类任务数据集）测试自己独立实现的算法，并与 MindSpore 平台上的官方实现算法进行对比，进一步分析差异及其成因。

实验要求

- 推荐使用 Python（在独立实现算法时，可采用 Numpy，Pandas，Matplotlib 等基础代码集成库；在使用 MindSpore 平台时，可使用平台提供的代码集成库）。
- 在独立实现算法时，不得使用集成度较高、函数调用式的代码库（如 sklearn，PyTorch，Tensorflow 等）。
- 尽量以相对路径的形式索引数据集，便于我们对代码进行复现。

实验报告格式

- 需要提供完整的可运行代码文件，结果文件和实验报告，将以上内容打包压缩，压缩文件命名格式：学号-姓名-xxx 实验。实验报告和代码注释应尽量详细。
- 实验报告内容参照报告模板，包括问题描述、实现步骤与流程、实验结果与分析、每个实验的心得体会（谈谈你自己的实现和 MindSpore 实现的差异、你在使用 MindSpore 平台过程中遇到的问题，以及想对平台改进提出的建议）、一个总的心得

体会（谈一谈你对这门课程理论及实验的感悟与体会）。

- 代码和报告若有雷同，一律按 0 分处理。
- 若存在疑问，可以联系：seu_pr_2023@163.com