隐马尔科夫模型分词任务

概述

- 利用隐马尔科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，是中文分词统计的常用资料。可以在语料库基础上构建词典、进行统计、机器学习等。

任务说明

- 中文信息处理是自然语言处理的分支，是指用计算机对中文进行处理。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。因此对中文进行处理的第一步就是进行**分词，即将字串转变成词串**。通过确立状态集合(B, M, E, S)，四个字母分别代表一个字在词语中的开始/中间/结尾/或者单字成词，这样可以将输入的中文句子编为一段状态序列，然后计算初始状态概率、转移概率及发射概率实现整个算法过程。具体过程及细节可以**参考Tips中的链接**。
- 在人民日报分词语料库上统计语料信息，对**隐马尔科夫模型**进行训练。利用训练好的模型，对以下语句进行分词测试：
  1) 今天的天气很好。
  2) 学习模式识别课程是有难度的事情。
  3) 我是东南大学的学生。

Tips

- 推荐语言：Python（可采用Numpy，Pandas，Matplotlib等基础代码集成库）、Matlab、C++。
- 不得使用集成度较高，函数调用式的代码库（如Python环境下的sklearn，PyTorch，Tensorflow等）。
- 关于分词任务的简介可参考此链接 https://www.cnblogs.com/llhthinker/p/6323604.html
- HMM用于分词任务的示例 https://blog.csdn.net/taoyanqi8932/article/details/75312822
- 此外，提供两篇文献作为参考与学习：

[1] 魏晓宁. 基于隐马尔科夫模型的中文分词研究[J]. 电脑知识与技术(学术交流), 2007, 4(21):885-886.

[2] 吴帅, 潘海珍. 基于隐马尔可夫模型的中文分词[J]. 现代计算机(专业版), 2018(33):27-30.

作业提交格式要求

- 需提供完整的**代码文件**，将以上内容打包压缩，**压缩文件命名格式：学号-姓名-隐马尔科夫模型分词任务实验。**
- 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。
- **代码若有雷同，一律按0分处理。**