第二次作业数据集介绍

文件

- **train.in**, **train.out**
  共6000条训练数据。输入在‘in’文件中，每行一条数据，大概为30~100词长度的英文段落。这些英文段落出自两个相似作者的文学著作中，对应每条输入数据按顺序在‘out’文件中有一行输出，分别为0或者1，分别代表此输入数据的文本出自作者0或是作者1的笔下。
- **test.in**, **test.out**
  共600条测试数据。输入输出格式与训练集相同。

提交文件

- 文件1：
  要求格式与train.out相同，每行第一个字符为0/1，第二个字符为回车/换行，共600行。对应下面的**精准率**指标
- 文件2：
  要求格式与train.out相同，但每行为一个[0.0,1.0]的小数，表示预测结果为作者1的概率。对应下面的**交叉熵**指标
- 要求提交结果必须包括文件1和文件2，如果实现的模型仅能提供一种结果，可以按下列方式转换结果：
  - 文件1->文件2，将每行的0/1替换为**0.1/0.9**。
  - 文件2->文件1，将每行的小数按**0.5**为阈值转换为0（小于**0.5**）/1（大于**0.5**）
  - 上述两种转换方法中加粗的**阈值参数**均可根据需要自行调整。实现方法也可自行选择。只需提供最终的文件1和文件2，以及在报告中注明所采用的方法即可
- 最终结果将综合以下两种指标评价

评测指标

精准率

$$P = correct/total$$

correct为提交的结果与标准答案匹配的行数
total=600

交叉熵

wiki link
对于此二分类的计算公式：

$$CE = -\frac{\sum_{i=1}^{total} y_i * log(y_i') + (1 - y_i) * log(1 - y_i')}{total}$$

其中对于第i组数据，y_i 为标准答案（表示作者为作者1的概率），y_i' 为预测结果（表示预测为作者1的概率）
total=600

评测方法

此次作业已给出测试集的正确答案, 采取自评方式，在提交的报告中给出自评的结果（结果仅供打分参考，请勿弄虚作假）。
运行 `python judge.py -gold $PATH_TO_GOLD_STANDARD -acc $PATH_TO_FILE_1 -ce $PATH_TO_FILE_2` ，即可得到在测试集上的结果。（python 2或3均可）