中文情感分析

本次作业为某宝收集到的商品客户评价，1为负面，2为正面。

训练集20w，测试集 20w，除了训练集和测试集，我们还给出了无标记数据集100w。

要求根据训练集训练网络，并进一步使用无标记数据采用semi-supervised learning方法判断测试集合的商品评价情感。

semi-supervised 建议采用其中一种比较好实作的方法 Self-Training

Self-Training:

把训练好的模型对无标记数据做预测，并将这些预测后的值转成该笔无标记数据的label，并加入这些新的做训练。你可以调整不同的 threshold，或是多次取样来得到比较有信心的 数据。e.g., 设定 pos\_threshold = 0.8，只有 prediction > 0.8 的数据会被标上 1

训练好的model请上传至github-classom

评分系统为kaggle

Python>3 开放frame为 numpy, pandas, standatd lib, pytorch, tensorflow, keras

其他frame请群@我

Github上请包涵

report.pdf

python files

train.sh

test.sh

model param

res.csv

Do not upload data! Do not upload data!Do not upload data!

补充

预测结果为res.csv，title为label，value为预测结果

bash train.sh <training data>

training data: path of train.csv

bash test.sh <testing data> <prediction file>

testing data: path of test.csv

prediction file: path of res.csv

Report

学号 姓名 专业

1. 请分别采用RNN（LSTM），Bert（Bert family）架构模型进行训练模型，并预测，说明你的RNN/Bert 的模型架构、训练过程 (learning curve) 和准确率为何？
2. 请叙述你如何 improve performance（preprocess、embedding、架构等等），并解释为何这些做法可以使模型进步，并列出准确率与 improve 前的差异。
3. （选做）请描述你的semi-supervised方法是如何标记label，并比较有无semi-supervised training对准确率的影响并试着探讨原因（如果效果不明显，可以尝试减少training date数量）