中文文本分类作业

参考项目：pytorch_transformer_chinese_text_classification

Overview

项目为对一段文本进行贴标签分类，例如:

> 盖世汽车讯，特斯拉去年击败了宝马，夺得了美国豪华汽车市场的桂冠，并在今年实现了开门红。1月份，得益于大幅降价和7500美元美国电动汽车税收抵免，特斯拉再度击败宝马，蝉联了美国豪华车销冠，并且注册量超过了排名第三的梅赛德斯-奔驰和排名第四的雷克萨斯的总和。根据Experian的数据，在所有豪华品牌中，1月份，特斯拉在美国的豪华车注册量为49，917辆，同比增长34%；宝马的注册量为31，070辆，同比增长2.5%；奔驰的注册量为23，345辆，同比增长7.3%；雷克萨斯的注册量为23，082辆，同比下降6.6%。奥迪以19，113辆的注册量排名第五，同比增长38%。凯迪拉克注册量为13，220辆，较去年同期增长36%，排名第六。排名第七的讴歌的注册量为10，833辆，同比增长32%。沃尔沃汽车排名第八，注册量为8，864辆，同比增长1.8%。路虎以7，003辆的注册量排名第九，林肯以6，964辆的注册量排名第十。|汽车|

> 预测标签：汽车

共有5个类别，分别为体育、健康、军事、教育、汽车。
划分为训练集和测试集，其中训练集每个类别800条样本，测试集每个类别100条样本。

Dependence

- [x] pytorch
- [x] scikit-learn
- [x] gensim
- [x] 中文词向量： https://pan.baidu.com/s/1-l9pdeUOwVzRVT4utvszfQ?pwd=ameb

Run

1. 运行 `preprocessing.py` 处理数据.
2. 将下载好的中文词向量 `sgns.wiki.char.bz2` 放入文件夹 `./Pretrain_Vector/` 中
3. 运行 `model_train.py` 训练模型
4. 运行 `model_eval.py` 测试模型
5. 运行 `model_predict.py` 使用模型进行文本分类

文件说明

- `model.py`：模型实现及部分参数设置
- `model_train.py`：训练模型代码
- `model_eval.py`：测试集上测试代码
- `model_predict.py`：使用模型进行分类
- `params`：训练参数模型参数
- `pickle_file_operator.py`，`preprocessing.py`，`text_featuring.py`： 数据处理脚本

默认参数效果

测试集正确率：0.74

> 修改参数训练至合格正确率大于 80%

提交

以小组为单位，每个小组提交一份报告至邮箱 `22S053079@stu.hit.edu.cn`