# 互联网新闻情感分析复现文档

组名：我们都上哈工深 复赛A榜：0.82330281 复赛B榜：0.81697905

### 文件结构说明

本项目下共有三个文件夹，分别为**data**，**model**，**source**以及**复赛融合**。

1. **data**文件夹下包含初赛和复赛的训练数据及复赛的测试数据。另外还包括数据处理代码。
2. **model**文件夹下包含运行的各个模型保存及预测的结果。
3. **source**文件夹下包含本次比赛的代码以及模型运行的jupyter notebook文件，共12个。
4. **复赛融合**文件夹下包含了融合的代码和融合前的12个中间文件。

### 数据处理

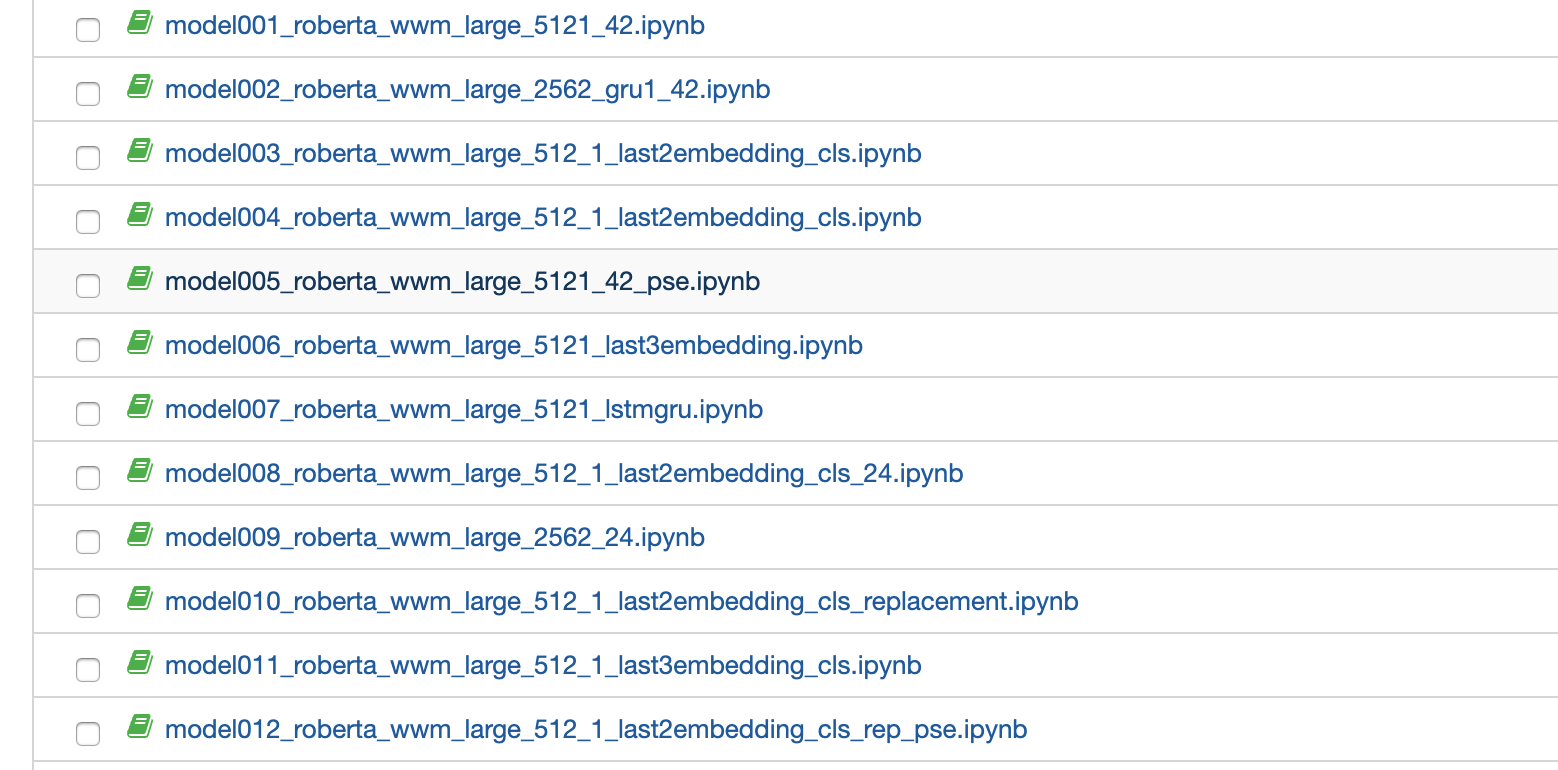
先将初赛和复赛数据保存在data目录中，其中初赛的训练数据为Train\_DataSet.csv和Train\_DataSet\_Label.csv，复赛的训练数据重命名为Second\_DataSet.csv和Second\_DataSet\_Label.csv，而复赛的测试数据集则重命名为Second\_TestDataSet.csv。

所有涉及数据处理的ipynb文件也都在data目录中。

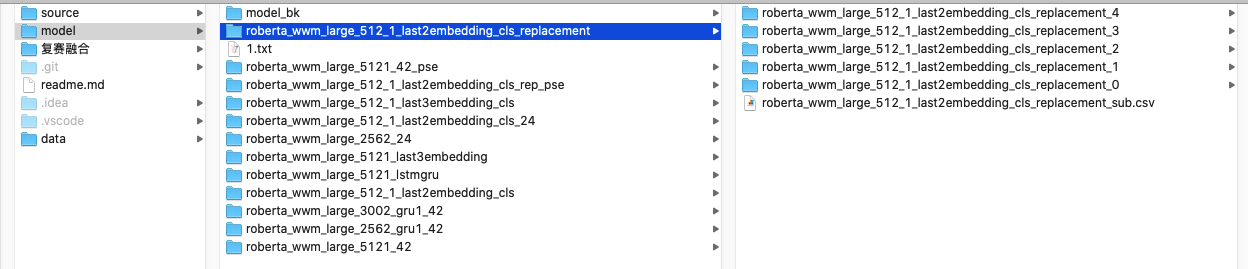
1. 先运行“001\_[初赛复赛数据集合并.ipynb](user_cancel)”将初赛和复赛训练数据进行合并。
2. 运行两个ipynb文件“002\_[原始数据分层抽样划分5折.ipynb](user_cancel)”和“003\_[原始数据分层抽样划分5折-随机种子24.ipynb](user_cancel)”分别以24和42的随机种子对合并后数据文件进行五折划分，用于后续的交叉验证。
3. 运行“004\_数据预处理（替换英文字符).ipynb”生成新的预处理后的数据，并进行五折划分，用于后续的模型训练

### 前9个模型运行

一共使用了12个模型，12个模型的运行代码所在目录为“sentiment-code-last/source”，如下图所示：



先运行前9个模型（model001-model009），可以得到模型的结果文件，所在目录为“sentiment-code-last/model”，如下图所示：

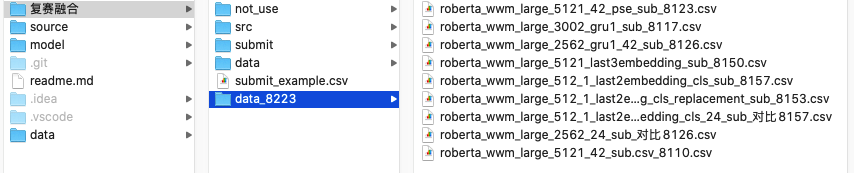


### 前9个模型结果融合用于伪标签

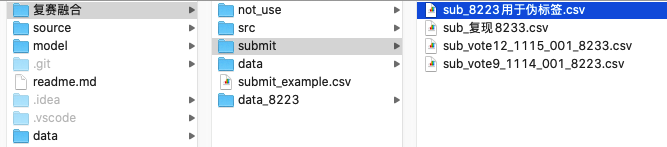
model012使用到了伪标签数据（方法是，将模型对测试集的预测结果当作训练集，用于训练）。

伪标签数据的生成过程：

首先，将前9个模型（model001-model009）生成的可提交结果文件，放入目录“sentiment-code-last/复赛融合/data\_8223”中，如下图所示：



然后，运行“sentiment-code-last/复赛融合/src/czx\_001\_投票融合4伪标签.ipynb”，可得到融合后的结果文件“sentiment-code-last/复赛融合/submit/sub\_8223用于伪标签.csv”。

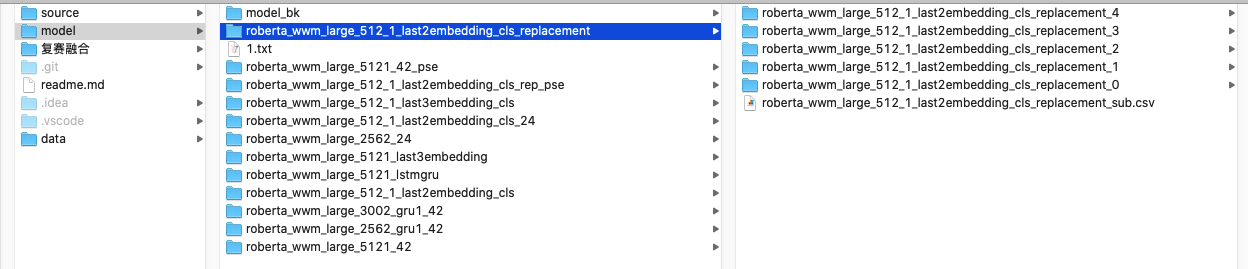


### 伪标签数据处理

在**完成第一次模型融合**后，运行"sentiment-code-last/data"目录下"005\_原始数据伪标签处理.ipynb"和"006\_替换数据伪标签处理.ipynb"两个文件，分别生成基于原始数据和预处理好后数据的带伪标签的五折划分数据，用于最后的模型训练

### 后3个模型运行

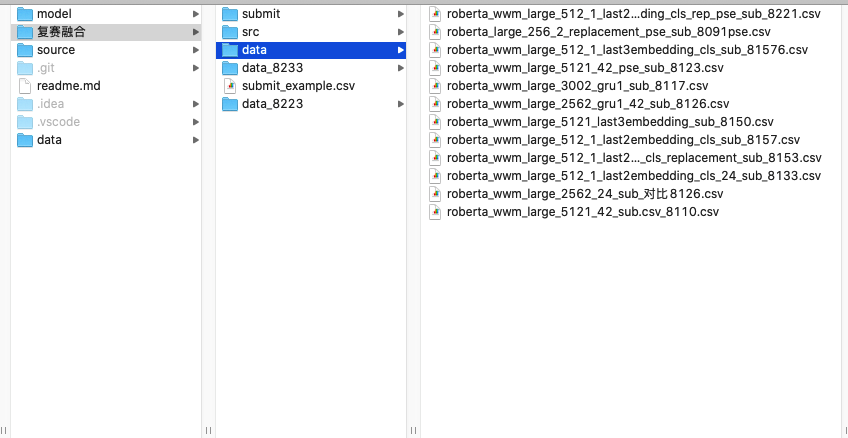
依次运行model010-model012，可以得到模型的结果文件，所在目录为“sentiment-code-last/model”，如下图所示：



### 最终模型融合

模型融合采用简单投票。

首先，将12个模型生成的可提交结果文件（例如：“sentiment-code-last/model/roberta\_wwm\_large\_5121\_42/roberta\_wwm\_large\_5121\_42\_sub.csv”），放入目录“sentiment-code-last/复赛融合/data”中，（12个模型生成结果，我们已经提供）如下图所示：



然后，运行“sentiment-code-last/复赛融合/src/czx\_001\_投票合.ipynb”可得到融合后的结果文件“sentiment-code-last/复赛融合/submit/sub\_复现8233.csv”，提交即为复赛A榜最高分0.8233和B榜最高分0.8169。

