

# 第一步

名称	修改日期	类型	大小
bilstm	2024/5/10 11:59	文件夹	
cnn	2024/5/10 14:59	文件夹	
cnn-bilstm	2024/5/10 14:59	文件夹	
之前第三模块的代码	2024/5/10 15:00	文件夹	
data.xlsx	2024/4/25 11:56	Microsoft Excel ...	5,286 KB
image.jpg	2024/5/9 9:44	JPG 文件	18 KB
new_data.xlsx	2024/5/9 9:59	Microsoft Excel ...	1,307 KB
stopwords_cn.txt	2024/5/9 11:26	文本文档	16 KB
数据预处理.py	2024/5/9 11:34	JetBrains PyChar...	4 KB
需求.jpg	2024/5/9 9:44	JPG 文件	536 KB

首先运行数据预处理这个步骤 这个代码的目的是把data.xlsx这个原始文件进行数据预处理，生成new\_data.xlsx这个文件，new\_data.xlsx这个文件的作用主要是用于建模需要用到的。

## 建模

建模一共有三个建模 cnn bilstm cnn\_bilstm

首先运行的是cnn这个项目

名称	修改日期	类型	大小
classification_report.txt	2024/5/5 16:53	文本文档	1 KB
cnn_demo.py	2024/5/5 16:52	JetBrains PyChar...	7 KB
cnn_Model loss accuracy_data.xlsx	2024/5/5 16:53	Microsoft Excel ...	6 KB
cnn_PR曲线.png	2024/5/5 16:53	PNG 文件	21 KB
cnn_ROC曲线.png	2024/5/5 16:53	PNG 文件	32 KB
cnn_test.xlsx	2024/5/5 16:53	Microsoft Excel ...	1,491 KB
Model loss accuracy.png	2024/5/5 16:53	PNG 文件	60 KB
情感占比分布.png	2024/5/5 16:53	PNG 文件	272 KB

直接运行cnn\_demo这个py文件就可以自动生成这个文件夹里面的所有文件

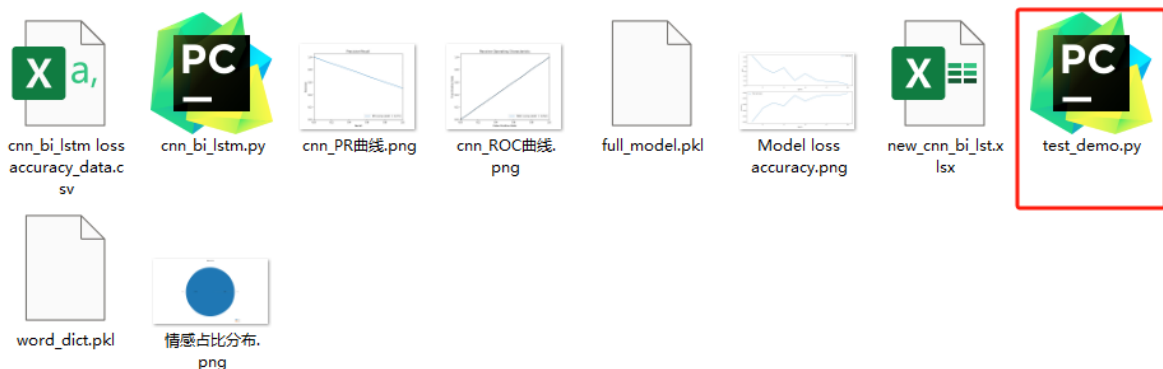
，接着去打开bilstm这个文件

这里建模的文件一共有两个bi\_lstm.py和test\_demo.py，bi\_lstm.py主要是生成full\_model.pkl 这个就是建模的模型，如果你只是说要一些指标和可视化，那就没必要重新运行bi\_lstm.py，而是直接去运行test\_demo.py 这个文件，运行test\_demo.py这个文件就会自动生成各种指标 什么准确率召回率 f1值 roc曲线等等

名称	修改日期	类型	大小
neg	2024/5/9 22:28	文件夹	
pos	2024/5/9 22:23	文件夹	
bi_lstm_loss_accuracy_data.csv	2024/4/30 8:03	Microsoft Excel ...	1 KB
bi_lstm.py	2024/4/29 22:29	JetBrains PyChar...	8 KB
bi_lstm_PR曲线.png	2024/5/5 16:58	PNG 文件	19 KB
bi_lstm_ROC曲线.png	2024/5/5 16:58	PNG 文件	33 KB
classification_report.txt	2024/5/5 16:58	文本文档	1 KB
full_model.pkl	2024/4/30 8:03	PKL 文件	167 KB
LDA建模.py	2024/5/10 11:59	JetBrains PyChar...	6 KB
Model loss accuracy.png	2024/5/5 16:58	PNG 文件	61 KB
neg.xlsx	2024/5/10 9:38	Microsoft Excel ...	513 KB
neg-top200-词云图.html	2024/5/9 20:22	Chrome HTML D...	49 KB
neg-高频词Top30.html	2024/5/9 20:22	Chrome HTML D...	7 KB
neg-高频词Top200.csv	2024/5/9 13:01	Microsoft Excel ...	3 KB
new_bi_lstm.xlsx	2024/5/5 16:58	Microsoft Excel ...	1,485 KB
pos.xlsx	2024/5/10 11:06	Microsoft Excel ...	916 KB
pos-top200-词云图.html	2024/5/9 20:22	Chrome HTML D...	49 KB
pos-高频词Top30.html	2024/5/9 20:22	Chrome HTML D...	7 KB
pos-高频词Top200.csv	2024/5/9 13:03	Microsoft Excel ...	3 KB

名称	修改日期	类型	大小
bi_lstm_PR曲线.png	2024/5/5 16:58	PNG 文件	19 KB
bi_lstm_ROC曲线.png	2024/5/5 16:58	PNG 文件	33 KB
classification_report.txt	2024/5/5 16:58	文本文档	1 KB
full_model.pkl	2024/4/30 8:03	PKL 文件	167 KB
LDA建模.py	2024/5/10 11:59	JetBrains PyChar...	6 KB
Model loss accuracy.png	2024/5/5 16:58	PNG 文件	61 KB
neg.xlsx	2024/5/10 9:38	Microsoft Excel ...	513 KB
neg-top200-词云图.html	2024/5/9 20:22	Chrome HTML D...	49 KB
neg-高频词Top30.html	2024/5/9 20:22	Chrome HTML D...	7 KB
neg-高频词Top200.csv	2024/5/9 13:01	Microsoft Excel ...	3 KB
new_bi_lstm.xlsx	2024/5/5 16:58	Microsoft Excel ...	1,485 KB
pos.xlsx	2024/5/10 11:06	Microsoft Excel ...	916 KB
pos-top200-词云图.html	2024/5/9 20:22	Chrome HTML D...	49 KB
pos-高频词Top30.html	2024/5/9 20:22	Chrome HTML D...	7 KB
pos-高频词Top200.csv	2024/5/9 13:03	Microsoft Excel ...	3 KB
test_demo.py	2024/5/5 16:57	JetBrains PyChar...	8 KB
word_dict.pkl	2024/4/30 8:03	PKL 文件	176 KB
可视化.py	2024/5/9 20:22	JetBrains PyChar...	5 KB
情感占比分布.png	2024/5/5 16:58	PNG 文件	267 KB

接着再去打开cnn\_bilstm这个文件，同样也是运行test\_demo.py这个文件就会生成一系列的图片和指标



## 可视化

根据上面的模型效果，我们可以得知，bilstm这个模型效果是最好的，所以我们专攻bilstm这个模型即可，打开bilstm这个文件

点击可视化，这个文件的生成逻辑如下：首先基于new\_bi\_lstm.xlsx 这个文件去生成top200的高频词，然后再基于这些高频词文件去生成词云图和柱状图

这边做了一个注释处理，当你运行的可视化的这个文件，直接跳过第一步，去读取top200高频词这两个文件去生成词云图和柱状图，所以如果你想要词云图和柱状图，你只需要修改top200这两个文件就可以改变对应词云图和柱状图的状态

名称	修改日期	类型	大小
bi_lstm_PR曲线.png	2024/5/5 16:58	PNG 文件	19 KB
bi_lstm_ROC曲线.png	2024/5/5 16:58	PNG 文件	33 KB
classification_report.txt	2024/5/5 16:58	文本文档	1 KB
full_model.pkl	2024/4/30 8:03	PKL 文件	167 KB
LDA建模.py	2024/5/10 11:59	JetBrains PyChar...	6 KB
Model loss accuracy.png	2024/5/5 16:58	PNG 文件	61 KB
neg.xlsx	2024/5/10 9:38	Microsoft Excel ...	513 KB
neg-top200-词云图.html	2024/5/9 20:22	Chrome HTML D...	49 KB
neg-高频词Top30.html	2024/5/9 20:22	Chrome HTML D...	7 KB
neg-高频词Top200.csv	2024/5/10 9:38	Microsoft Excel ...	3 KB
new_bi_lstm.xlsx	2024/5/5 16:58	Microsoft Excel ...	1,485 KB
pos.xlsx	2024/5/10 11:06	Microsoft Excel ...	916 KB
pos-top200-词云图.html	2024/5/9 20:22	Chrome HTML D...	49 KB
pos-高频词Top30.html	2024/5/9 20:22	Chrome HTML D...	7 KB
pos-高频词Top200.csv	2024/5/9 13:03	Microsoft Excel ...	3 KB
test_demo.py	2024/5/5 16:57	JetBrains PyChar...	8 KB
word_dict.pkl	2024/4/30 8:03	PKL 文件	176 KB
可视化.py	2024/5/9 20:22	JetBrains PyChar...	5 KB
情感占比分布.png	2024/5/5 16:58	PNG 文件	267 KB

# LDA建模

LDA建模主要就是运行lda.py这个文件即可

名称	修改日期	类型	大小
neg	2024/5/9 22:28	文件夹	
pos	2024/5/9 22:23	文件夹	
bi_lstm_loss_accuracy_data.csv	2024/4/30 8:03	Microsoft Excel ...	1 KB
bi_lstm.py	2024/4/29 22:29	JetBrains PyChar...	8 KB
bi_lstm_PR曲线.png	2024/5/5 16:58	PNG 文件	19 KB
bi_lstm_ROC曲线.png	2024/5/5 16:58	PNG 文件	33 KB
classification_report.txt	2024/5/5 16:58	文本文档	1 KB
full_model.pkl	2024/4/30 8:03	PKL 文件	167 KB
LDA建模.py	2024/5/10 11:59	JetBrains PyChar...	6 KB
Model loss accuracy.png	2024/5/5 16:58	PNG 文件	61 KB
neg.xlsx	2024/5/10 9:38	Microsoft Excel ...	513 KB
neg-top200-词云图.html	2024/5/9 20:22	Chrome HTML D...	49 KB
neg-高频词Top30.html	2024/5/9 20:22	Chrome HTML D...	7 KB
neg-高频词Top200.csv	2024/5/9 13:01	Microsoft Excel ...	3 KB
new_bi_lstm.xlsx	2024/5/5 16:58	Microsoft Excel ...	1,485 KB
pos.xlsx	2024/5/10 11:06	Microsoft Excel ...	916 KB
pos-top200-词云图.html	2024/5/9 20:22	Chrome HTML D...	49 KB
pos-高频词Top30.html	2024/5/9 20:22	Chrome HTML D...	7 KB
pos-高频词Top200.csv	2024/5/9 13:03	Microsoft Excel ...	3 KB

```

177
178
179 ► if __name__ == '__main__':
180     list_emotion = ['neg', 'pos']
181     for e in list_emotion:
182         df = pd.read_excel(f'{e}.xlsx')
183         data = df[df['情感type'] == e]
184         lda(data, e)
185
186
187

```

这里是读取两个文件，一个是neg.xlsx文件，一个是pos.xlsx文件

以上便是整个项目的运行过程，个人建议，你自己最好提前运行一遍，按照这个流程来运行，避免有什么卡点，你可以提前知晓

这里有一些额外的库你需要安装

pip install pyLDAvis -i <https://pypi.tuna.tsinghua.edu.cn/simple>

pip install gensim -i <https://pypi.tuna.tsinghua.edu.cn/simple>

pip install pyecharts -i <https://pypi.tuna.tsinghua.edu.cn/simple>