# 环境

Ubuntu18.04

Python3.6

torch1.7.0

transformers>=4.2.1

pycharm

4路tesla V100

# 预训练模型:

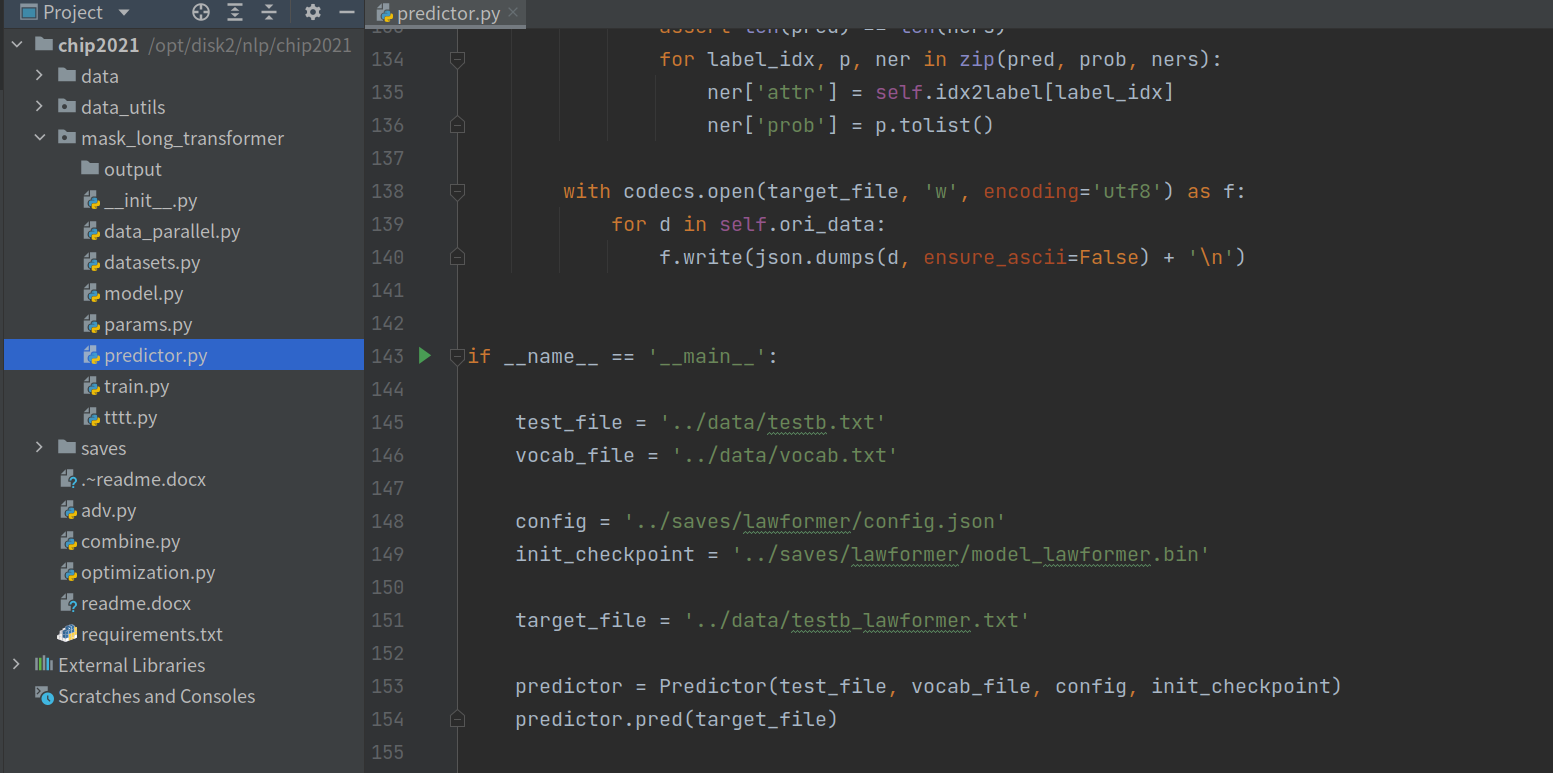
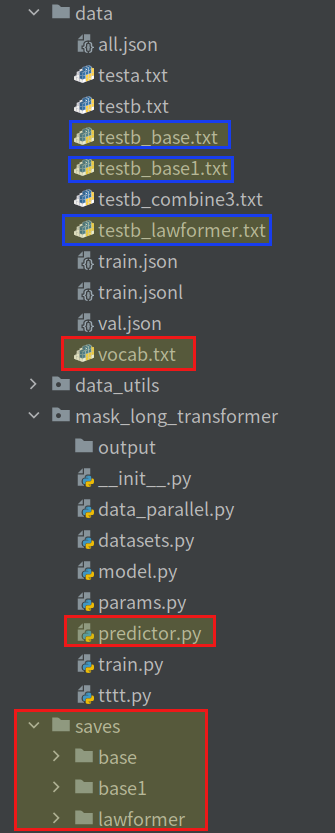
采用longformer，预训练参数采用huggingface官方库参数，分别如下：

base: <https://huggingface.co/schen/longformer-chinese-base-4096>

base1: <https://huggingface.co/ValkyriaLenneth/longformer_zh>

lawformer: <https://huggingface.co/thunlp/Lawformer>

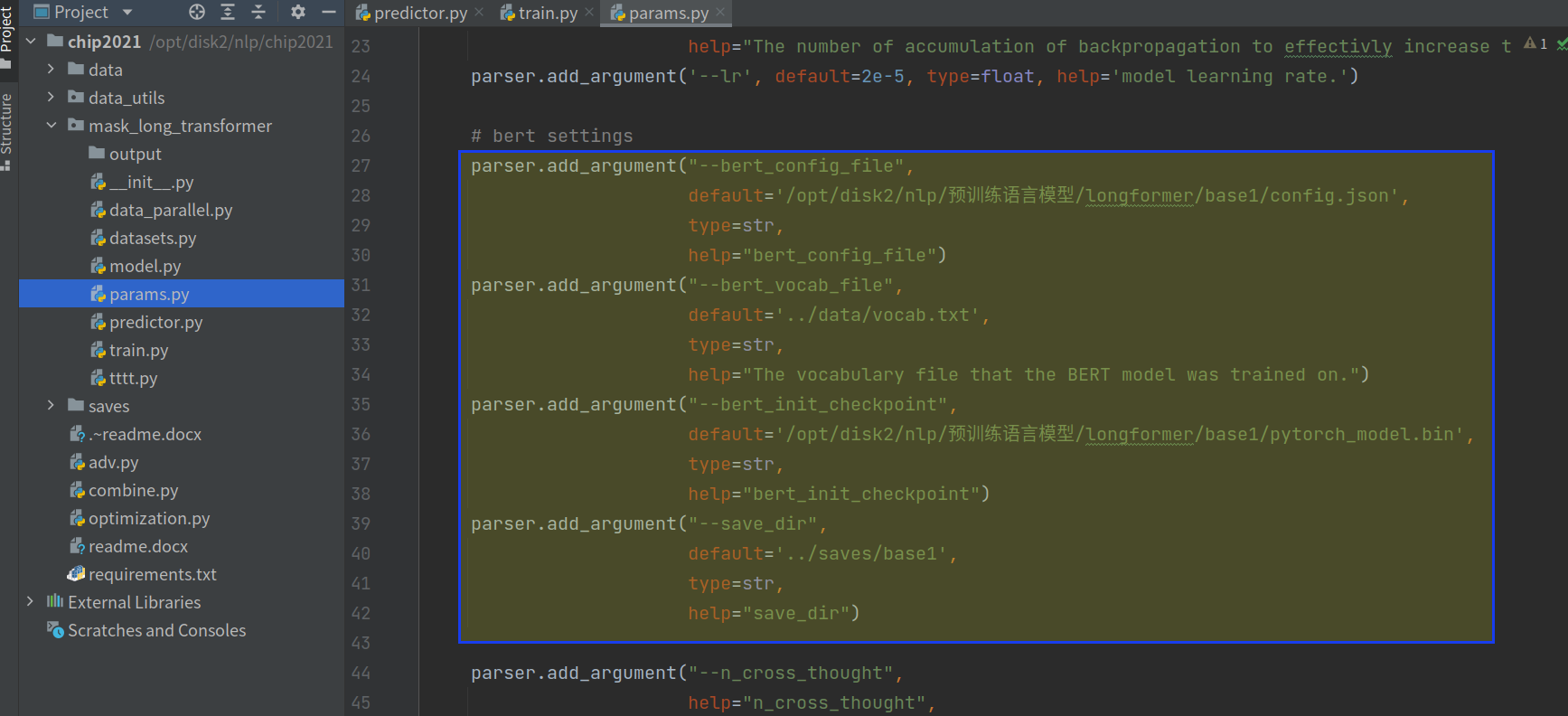
# 预测



使用saves下的参数预测，使用data目录下的vocab.txt文件，在predictor.py修改相关文件路径。这样可以得到3个预测的文件testb\_base.txt，testb\_base1.tx，

testb\_lawformer.txt。再运行combine.py融合上面3个预测文件得到testb\_combine3.txt

# 训练



修改params.py中上面的路径参数，其他参数默认。然后直接运行train.py即可。分别使用3个预训练参数训练3个模型。

# 基本模型构架

利用MLM模型，把实体标签分类看作mask预测任务，mask预测为4类。

原句：

[XLS][PATIEND]打了五联疫苗后， 发 烧 38.2度，医生诊断上呼吸道感染，可以吃感冒药，消炎药吗？[DOCOTOR]您好...

MLM语言模型任务：

[XLS][PATIEND]打了五联疫苗后，[MASK][MASK]38.2度，医生诊断上呼吸道感染，可以吃感冒药，消炎药吗？[DOCOTOR]您好...

本模型任务（并不是直接把entity全部遮蔽，而是输入原句，通过attention mask控制[entity\_MASK]可见性）：

[XLS][PATIEND]打了五联疫苗后，[entity\_MASK]38.2度，医生诊断上呼吸道感染，可以吃感冒药，消炎药吗？[DOCOTOR]您好...