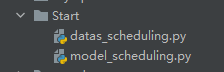
# 关于文本分类模型：

1：代码文件格式介绍已在lab的md文件中提及这里不再赘述。

2：流程介绍

### 2.1 开关：

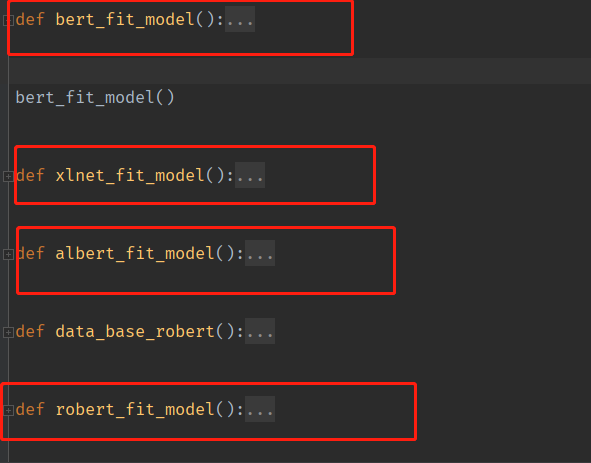
Start文件包含了两部分第一部分是开启数据清洗和特征工程：



Data开头的文件是开启数据处理流程：



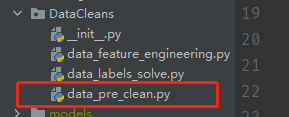
model开头的文件是启用模型的开关，直接根据调用对应的函数名即可，它会自动对传入模型的数据，进行传入准备处理。



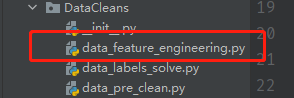
**直接调用xx\_fit\_model即可**。

### 2.2 数据处理

data\_clean是数据清洗文件对应：



而datas\_engineer是数据特征工程文件，对应



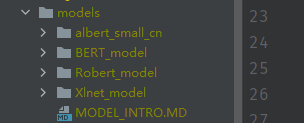
Data\_labels\_solve文件是之前设计的单词重要性程度排名逻辑，**用于缩减文本，后续由于数文本策略的改变因此，放弃使用**。

1609722437(1)

Mysql文件是调用对应的数据库，获取数据，这一块后续需要同@钟俊伟俊伟哥处沟通

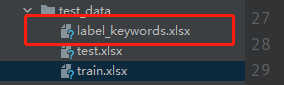
### 2.3模型

模型文件：models



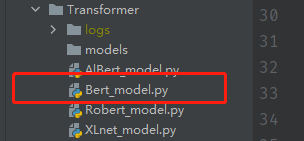
**只需要关注bert\_model即可**，xlnet和albert同bert一样是进行文本分类处理的文件，但是在效果上面和效率上面都不如bert因此，放弃，但是这两个文件都是可以直接使用的，会因此如果后续有需要可以自行调整传入的数据，进行训练，robert则是下一阶段的准备文件，只是写了逻辑未进行测试和使用。

Test\_data:测试数据



**红框的数据文件，是数据处理流程必须的文件，不能忘记**，test和train文件则是训练数据和测试数据。

Transformer文件：模型逻辑文件。

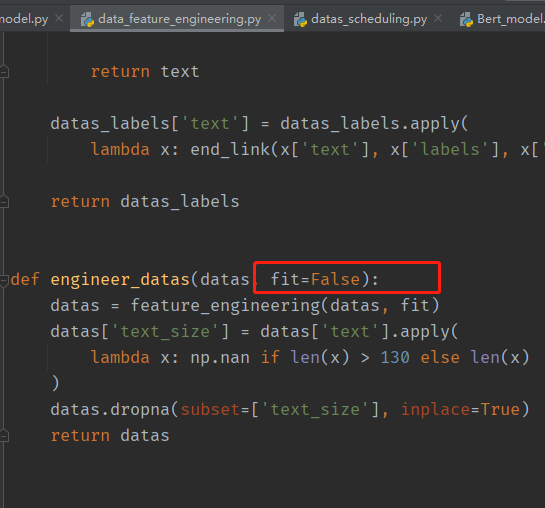


同样重点关注bert文件，开关在start文件中，直接根据对应函数名调用即可。

**Config文件和models文件则就是对应的一些配置和模型原始文件。**

### 2.4 开启数据清洗

关于start的模型开启训练：



Fit参数默认为false。

Fit代表的是训练，这里的意思是，在经过了**统一需要进行的数据清洗**后，开始进行特征工程了，**开启fit则意味着数据是按照训练模式的数据进行处理**，**关闭fit则是按照待预测的数据进行了处理**，这两种的不同点就是，训练数据是**有正确标签的且有是否解决这两列**的，而待预测数据则是**杂乱标签和没有是否解决这两列**的，这一点需要注意。