**Javaslicer文件夹：**

用于切割java源代码片段的java代码，

ExceptionSlice用于切割throw抛出的异常以及javadoc注释中@throw标签之后的注释部分

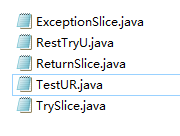
ReturnSlice用于切割返回值语句，返回值相关的语句以及javadoc注释中@return标签之后的注释部分

TrySlice用于切割try-catch语句中，抛出的异常，处理异常的语句以及javadoc注释中@throw标签之后的注释部分

TestUR，RestTryU用于切割java源代码中，除开以上三种代码以及注释，剩下的语句以及不含标签的javadoc注释。

RestTryU针对含有try-catch语句或者throw语句的，但是不含有return语句的java方法，切割出除开try-catch语句或者throw语句的剩下的代码，以及不含标签的javadoc注释。

TestUR针对含有return语句的java方法，切割出除开try-catch语句或者throw语句或者返回值以及返回值相关语句的剩下的代码，以及不含标签的javadoc注释。



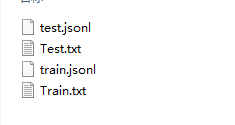
本部分代码逻辑说明可以在论文的第三章查看。

**Datasets文件夹：**

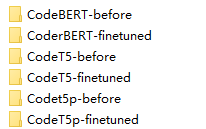
用于微调大模型的数据集文件，分为test集合和train集合，其中train集合用于输入模型进行微调训练，而test集合用于测试模型的输出，同时也测试模型在本集合上的bleu-4分数。

同一集合的Txt和jsonl文件之间仅格式不同，内容是相同的。Jsonl文件为经过分词处理后的代码-注释对。

Train集合中有139766对代码-注释对，test集合中有17470对代码注释对。



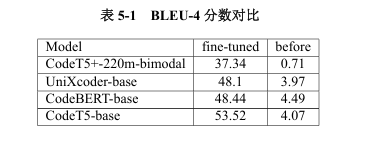
**各种模型checkpoints 文件夹：**

后缀为before的为原模型作者发布的已经在java-code-summarization任务上微调好的下游任务checkpoints，或者是根据原模型作者论文中使用的数据集，在原模型的基础上，之后自己在java-code-summarization任务上微调后的模型。取的是微调过程中bleu分数较高的模型。

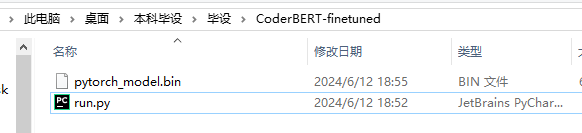
后缀为finetuned的为在原模型基础上，使用自己构建的数据集，微调后的模型或者checkpoints。取的是微调过程中bleu分数最高的模型。

各模型的规格以及表现如下：



**CodeBERT**

官方说明文档：[CodeBERT/CodeBERT/code2nl at master · microsoft/CodeBERT · GitHub](https://github.com/microsoft/CodeBERT/tree/master/CodeBERT/code2nl)



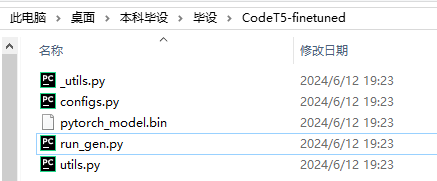
相对于官方给出的微调脚本，此处修改了一些脚本的参数默认值，需要输入的参数以及输入数据集处理函数，修改后的代码为run.py

官方说明文档已经足够详细，其他细节不再赘述。

**CodeT5**

官方说明文档：[CodeT5/CodeT5 at main · salesforce/CodeT5 · GitHub](https://github.com/salesforce/CodeT5/tree/main/CodeT5)

Before中使用的task为summarization

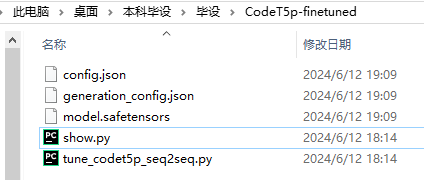


此处给出了相较于原模型作者发布的微调脚本，自己有修改的代码部分。此处修改了一些脚本的参数默认值，需要输入的参数以及输入数据集处理函数。

**CodeT5+**

官方说明文档[CodeT5/CodeT5+ at main · salesforce/CodeT5 · GitHub](https://github.com/salesforce/CodeT5/tree/main/CodeT5%2B)

由于原作者未发布合适的checkpoints，Before使用的是在原模型上使用CodeSearchNet数据金进行微调后的模型，使用该数据集的原因是本模型为CodeT5的进阶版，而codeT5的checkpoints是在CodeSearchNet上微调的。



此处给出了相较于原模型作者发布的微调脚本，自己有修改的代码部分。此处修改了一些脚本的参数默认值，需要输入的参数以及输入数据集处理函数。

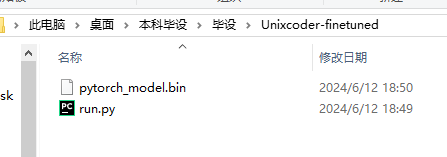
此外，show.py为自己写的用于测试微调后该模型的bleu-4分数的代码。因为关于这个模型，官方并没有给出在微调时就有显示bleu-4分数的脚本，而本文提到的其他模型，在微调时均会显示bleu-4分数。

**Unixcoder**

官方说明文档：[CodeBERT/UniXcoder at master · microsoft/CodeBERT · GitHub](https://github.com/microsoft/CodeBERT/tree/master/UniXcoder)

微调指导文档：

[CodeBERT/UniXcoder/downstream-tasks/code-summarization at master · microsoft/CodeBERT · GitHub](https://github.com/microsoft/CodeBERT/tree/master/UniXcoder/downstream-tasks/code-summarization)



此处给出了相较于原模型作者发布的微调脚本，自己有修改的代码部分。此处修改了一些脚本的参数默认值，需要输入的参数以及输入数据集处理函数