编程作业 2：基于 Transformer 的命名实体识别

作业说明：

1、语料：

1.1 训练语料：train.txt 为字(符号)序列，train_TAG.txt 为对应 train.txt 中每一个字(符号)的实体标签。例如：

train.txt 第一句：人 民 网 1 月 1 日 讯 据 《 纽 约 时 报 》 报 道 ，

train_TAG.txt 中对位的标签为：O O O B_T I_T I_T I_T O O O B_LOC I_LOC O O O O O O O

即：

train.txt 中的字(符号)	train_TAG.txt 中对位的标签	说明
人	O	非命名实体中的字(符号)
民	O	非命名实体中的字(符号)
网	O	非命名实体中的字(符号)
1	B_T	时间实体的开头字
月	I_T	时间实体的中间字
1	I_T	时间实体的中间字
日	I_T	时间实体的中间字
讯	O	非命名实体中的字(符号)
据	O	非命名实体中的字(符号)
《	O	非命名实体中的字(符号)
纽	B_LOC	地点实体的开头字
约	I_LOC	地点实体的中间字
时	O	非命名实体中的字(符号)
报	O	非命名实体中的字(符号)
》	O	非命名实体中的字(符号)
报	O	非命名实体中的字(符号)
道	O	非命名实体中的字(符号)
,	O	非命名实体中的字(符号)
……	……	……

1.2 发展集 dev.txt 及其标注 dev_TAG.txt（标注规范和 train_TAG.txt 中的相同），可用于训练过程中进行模型选择。

1.3 测试语料：test.txt：用于测试模型。

2、基于 train.txt 和 train_TAG.txt 数据训练一个以 Transformer 为基础结构的命名实体识别模型，进而为 test.txt 进行序列标注，输出标签文件，标签文件输出格式与 train_TAG.txt 相同。即保持 test.txt 中的行次序、分行信息以及行内次序，行内每个字的标签之间用空格分隔。输出文件命名方式：学号.txt。

■ 关于模型设计：

模型必须以 Transformer 为基础结构，可以单独采用 Transformer 编码器(典型的如 BERT)，也可以尝试单独采用 Transformer 解码器(典型的如 GPT)，或者编码器+解码器的组合(典型的如 T5)。Transformer 的层数等超参数自选，字向量可以随机初始化后训练，也可以采用已有开放词向量作为初始值，维度自选。是否采用类 CRF 模块建模序标关联可选。

3、所有输出文本均采用 Unicode(UTF-8)编码、算法采用 Python (3.0 以上版本) 实现

4、此次作业需要提交的材料：
4.1、基于 Transformer 结构的序标模型和训练算法的文本说明，提交 doc(或 pdf)文件，文件命名方式：学号；文本说明至少包括：
a)给出标注集(标签集)。序列标注的标签集是 train_TAG.txt 中的所有不同的标签组成的集合，请自行统计获得，标签集是后续训练模型和标注的基础，请注意统计完整。
b)对模型和执行细节进行说明。模型和执行细节应至少包含 :模型的结构及其结构参数(如用什么 transformer、堆叠层数、输入 token 数等等)、所用初始字向量的来源、向量维数、训练算法以及学习率、训练批次大小、训练轮数等；
c)给出训练损失和发展集性能随时间变化的曲线：每轮记录训练 loss，同时每一轮在发展集上进行测试，获得其标注性能(准确率)。给出所选择用来进行测试的训练轮次。
d)参考文献。给出参考的论文、网站、代码链接等等，说明是否使用了大模型辅助，辅助程度如何。
4.2、提交完整的实现代码，其中关键部分需要进行注释说明：与文本说明中的参数和执行细节对应。
4.3、对 test.txt 进行序列标注得到的标注文件
提交 txt 文件，文件命名方式：学号.txt；

注：评分时 4.1 占 40%、4.2 占 30%、4.3 占 30%

5、提交时间和方式：
提交方式：提交到教学云平台
提交时间：按教学云平台上规定的截止时间