作业三

命名实体识别（Named Entity Recognition，简称NER）是自然语言处理中的一个重要任务，其目的是从给定的文本中识别出具有特定意义的实体，将它们分类为预定义的类别，如人名、地名、组织机构名等。NER技术可以应用于很多场景，如信息抽取、文本分类、知识图谱构建等。在本次作业中，同学们将通过在中文法律文本上实践NER任务，熟悉一般NLP任务的基本流程与工具使用。最终须完成一份书面技术报告。

数据说明

本次任务所使用的数据集主要来自于网络公开的若干罪名起诉意见书，仅涉及盗窃罪名的相关信息抽取。针对本次任务，我们会提供包含案件情节描述的陈述文本，同学们需要识别出文本中的关键信息实体，并按照规定格式返回结果，存储在相应文件夹中。

发放的文件为`train_data.json`和`test_data.json`。请注意，test_data.json 并不是最终的评测数据，只是展示评测时的数据格式。两个文件均为字典列表，字典包含字段为：

- `id`：案例中句子的唯一标识符。
- `context`：句子内容，抽取自起诉意见书的事实描述部分。
- `entities`：句子所包含的实体列表。（该字段信息在`test_data.json`中被隐去）
  -- `label`：实体标签名称。
  -- `span`：实体在`context`中的起止位置。

其中`label`的十种实体类型分别为：

label	含义
NHCS	犯罪嫌疑人
NHVI	受害人
NCSM	被盗货币
NCGV	物品价值
NCSP	盗窃获利
NASI	被盗物品
NATS	作案工具
NT	时间
NS	地点
NO	组织机构

人名是指出现在案例文本中的自然人的姓名、昵称、社交媒体账号，该实体进一步细分为两种类型的实体，即“犯罪嫌疑犯”、“受害者”。

物品是指《中华人民共和国刑法》第九十一条、第九十二条规定的案件中的公私财产。为了准确区分项目，物品中还包括物品的属性（数量、颜色、品牌和编号等）。该实体进一步细分为“被盗物品”、“作案工具”。

货币是指国家法律认可的法定货币，包括贵金属货币、纸币、电子货币等。货币属性（人民币、美元等）也需要标注，以区分货币类型。该实体细分为“被盗货币”、“物品价值”和“盗窃获利”

案发时间是指案件发生期间的时间表达，包括日历时间（年、月、日等）和非日历时间（上午、下午、晚上、清晨等）。

案发地点是指案例中涉及的地理位置信息，应尽可能详细标注。它包括行政区名称、街道名称、社区名称、建筑编号、楼层编号、地标地址或自然景观等。此外，它还应包含位置指示，例如：“在房子前面”或“在建筑物后面”。

组织是指涉案的行政组织、企业组织或者非政府组织。

提交须知

我们在 `model` 中提供了你需要提交代码的样例（`model/submit_sample.zip`）。请注意，在提交的时候，你需要将你所有的代码和训练好的模型**压缩**为一个 `zip` 文件进行提交。该 `zip` 文件**内部顶层**必须包含`main.py`。

我们会在该文件夹下调用python main.py 来做inference，你的程序需要从`./input/test_data.json`中读取数据进行预测，并将预测的结果输出到`./output/output.json`中。预测结果文件为一个json格式的文件，包含两个字段，分别为`id`和`entities`，具体可参考`evaluate/result_sample.json`。

注意，使用预训练模型和不使用预训练模型会是两个平行的赛道，分别评分。如果使用预训练模型，只允许使用base大小的模型（参数不允许超过220M），例如bert-base, T5-base。

评测脚本

我们在 `evaluate` 文件夹中提供了评分的代码和提交文件样例，以供参考。

系统环境限制

torch1.8 ./envs/torch1.8.txt； **torch2.0** ./envs/torch2.0.txt

以上是评测时可以使用的环境，建议使用上述环境实验。
如果**一定**需要使用其他环境，需要在教学网提交作业后，再进行线下评测。我们会当场提供评测数据，在你的环境上做完inference之后，将output再提交给我们。