实验题目：文本标注与命名实体识别（**NER**）

实验目标

1. 理解文本标注的基本概念和流程。
2. 掌握命名实体识别（NER）任务的基本方法。
3. 使用标注工具对文本数据进行标注，并训练一个简单的NER模型。

实验内容

1. 数据准备
    - 根据座号从附件中对应的数据集文件中自行摘取完整句子进行文本标注。
    - 根据数据集内容自行定义至少四个类型的命名实体，例：人名（PER）、地名（LOC）、组织名（ORG）、时间（TIME）。
    - 摘取的句子需包含一个及以上自定义的命名实体类型，标注数据总量不少于500条。
2. 文本标注
    - 标注工具：使用标注工具（如Label Studio、BRAT）对文本中的命名实体进行标注。

      Label Studio：

      安装：`pip install label-studio`

      启动：`label-studio start`

      访问：打开浏览器，输入 http://localhost:8080。

      BRAT：

      安装：

      ```
      sudo apt-get install apache2
      git clone https://github.com/nlplab/brat.git
      cd brat
      ./install.sh
      ```

      访问：打开浏览器，输入 http://localhost/brat。

    - **标注格式**：使用BIO格式（Begin, Inside, Outside）标注实体。

      示例：

      文本：2023年10月，马云在阿里巴巴总部发表了演讲。

      标注：

      2023年10月：B-TIME I-TIME

      马云：B-PER

      阿里巴巴：B-ORG

      总部：B-LOC

- 标注要求：
  - 确保标注的前后一致性。
  - 标注完成后，将数据分为训练集、验证集和测试集（如8:1:1）。

3. 模型训练
   - 数据预处理：
     i. 将标注数据转换为模型可接受的格式（如JSON、CoNLL格式）。
     ii. 使用分词工具对文本进行分词（如Jieba、PKU分词工具）。
   - 模型选择（任选一个模型，可自定）：
     i. 传统机器学习模型：例如CRF（条件随机场）等。
     ii. 深度学习模型：
         1. 例如BiLSTM-CRF（双向LSTM + CRF）。
         2. 预训练模型：BERT、RoBERTa、XLNet等。
   - 模型训练：
     i. 使用Python和深度学习框架（如PyTorch、TensorFlow）进行模型训练。
     ii. 示例代码（使用Hugging Face Transformers库）：

```
from transformers import AutoTokenizer, AutoModelForTokenClassification
from transformers import pipeline


# 加载预训练模型和分词器
tokenizer = AutoTokenizer.from_pretrained("bert-base-chinese")
model = AutoModelForTokenClassification.from_pretrained("bert-base-chinese", num_labels=len(label_list))


# 创建NER pipeline
nlp = pipeline("ner", model=model, tokenizer=tokenizer)


# 测试模型
text = "2023年10月，马云在阿里巴巴总部发表了演讲。"
ner_results = nlp(text)
print(ner_results)
```

   - 训练细节：
     i. 设置超参数：学习率、批量大小、训练轮数等。
     ii. 使用验证集进行模型调优。

4. 模型评估
   - 评估指标：准确率（Precision）、召回率（Recall）、F1分数（F1-Score）。
   - 结果分析：分析模型在总体和不同类型实体上的表现（如PER、LOC、ORG等）。

实验要求

1. 提交材料：
    - 标注好的文本数据集（原始文本和标注文件）。
    - 模型训练代码和预测结果。
    - 实验报告，重命名为“学号姓名.docx”。
    - 截止日期：2025年3月16日23:55。
2. 实验报告内容：
    - 数据集描述：数据来源、数据量、实体类型分布。
    - 标注过程：标注工具、标注标准、标注一致性检查方法。
    - 模型训练：模型选择、超参数设置、训练过程。
    - 结果分析：评估指标、模型表现、错误案例分析。