**作业三：序列标注作业**

**任务内容：** 利用目录 data 中的数据训练模型来完成序列标注任务（命名实体识别），识别出文本中的人名、地名和组织机构名。

**数据说明：（此数据集为小规模中文数据集，来自于98年人民日报NER语料）**

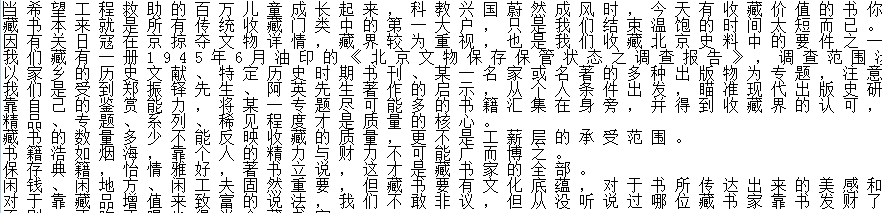
1. **标签说明：**

{ **B-PER**：人名开始； **I-PER**：人名中间 ；**B-LOC**：地名开始， **I-LOC**：地名中间； **B-ORG**：机构名开始 ；**I-ORG**：机构名中间；**O**：其他 }

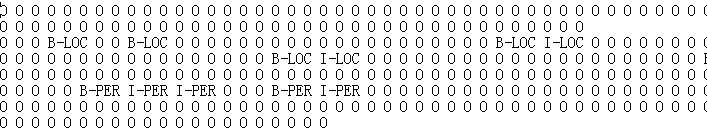
1. **数据集格式说明：**

**训练语料**

目录data 中 train\_corpus.txt 文件为训练文本，其格式为：



目录data 中 train\_label.txt 文件为训练文本对应的每个字的标签，其格式为：

如周恩来这个人名实体对应的标签B-PER, I-PER, I-PER。

**测试语料**

目录data 中 test\_corpus.txt 文件为测试文本，test\_label.txt 文件为对应的标签 （格式同训练语料）

1. **评价指标**

采用准确率、召回率以及F1值来进行模型评价。（由于标签中 “O”占非常大的比例，因此在计算指标时，采用两种方式，一是直接计算所有的指标，二是去掉“O”这个类别后再计算所有指标）

**任务说明：**

**1. 模型要求：**作业的基本要求为采用神经网络+条件随机场模型实现，例如Bi-LSTM+CRF模型，其他细节可自己设计或参考相关论文。

**2. 完成作业方式：** 使用任意一种深度学习框架（推荐pytorch），利用目录 data 中的数据训练模型来完成序列标注任务。

**3. 作业提交要求**：需提交代码部分（去掉数据和训练好的模型），同时写一个实验报告来说明自己所使用的模型结构以及最终测评结果等。由于邮箱的空间有限，如果愿意提交训练好的模型，可以将其用百度云分享链接的形式放在报告中。