**作业3 - 概率图模型练习**

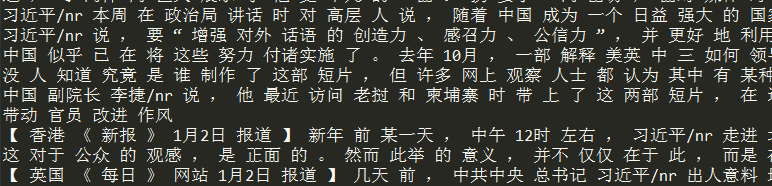
**一、任务3：中文人名识别**

**人名识别** 是命名实体识别的一部分，是将中文序列中的人名识别出来。

**要求：**用最大熵或条件随机场模型实现中文人名识别。

**数据资源：**语料来自1998年1月份人民日报语料。

**语料格式：**



其中，人名已显式标注。

**语料使用方法：**

按照8:2比例，将数据划分为训练数据和测试数据；训练数据部分用来训练模型参数，测试部分用来验证模型效果。

**模型验证：**

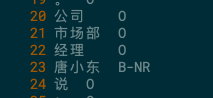
采用交叉验证方式，**指标：**准确率、召回率、F1值。

**二、模型说明**

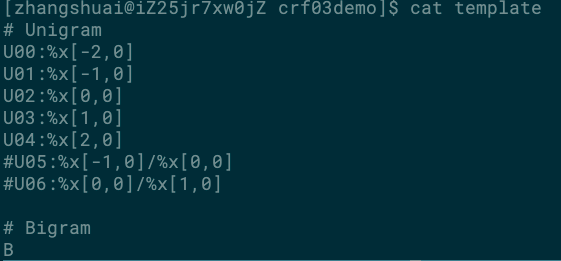
本次实验通过使用crf++工具包来进行名字实体识别。主要过程如下：

1. 首先将数据集按8：2划分成训练集、测试集，并对数据进行处理，保存在文件中。

python3 data\_util.py

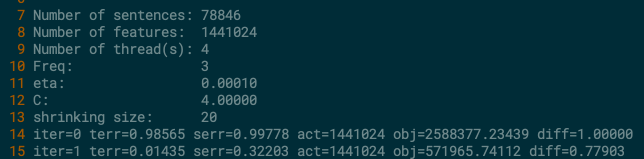


1. 编写template，设置特征函数



2)使用crf++命令进行学习，并将学习的模型进行保存

nohup CRF++-0.58/crf\_learn -f 3 -p 4 -c 4.0 template train\_data.txt model > train.result &



1. 使用crf++命令进行标注,并保存标注结果

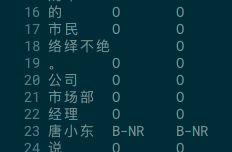
nohup CRF++-0.58/crf\_test -m model test\_data.txt > test.result &

1. 编码实现标注结果评估

python3 evaluate.py

**四、结果**

标注结果（test.result）：



评估结果：



词总数： 941246 预测正确标注数目：936123

名字实体总数：12321 预测总数：8074 预测准确数目： 7636

准确率：0.995 F1值：0.749 正确率：0.946 召回率：0.620