## 一、数据预处理：

1. 数据采用python进行预处理，代码在source目录的data\_format.py文件
2. 预处理后的数据分为两列，第一列为词本身，第二列为该词的标注B代表命名实体，O代表非命名实体。
3. 采用随机方法进行数据划分测试集包含943397个词，训练集包含3830726个词，具体文件见data目录的train\_data.txt和test\_data.txt文件。

## 二、特征模板：

由于电脑性能不足，设置的模板较为简单，如下：

#Unigram

U00:%x[-2,0]

U01:%x[-1,0]

U02:%x[0,0]

U03:%x[1,0]

U04:%x[2,0]

U05:%x[-1,0]/%x[0,0]

U06:%x[0,0]/%x[1,0]

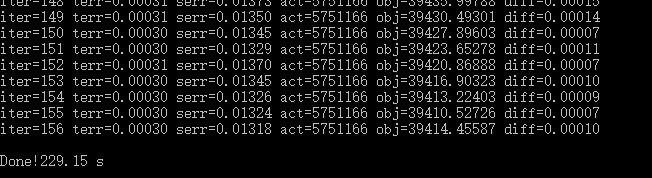
U08:%x[3,0]

U09:%x[-3,0]

可参见data目录下的template.txt文件

## 三、训练与测试

采用crf\_learn.exe template.txt train\_data.txt model\_10.txt 命令训练模型，最终训练结果如下图：

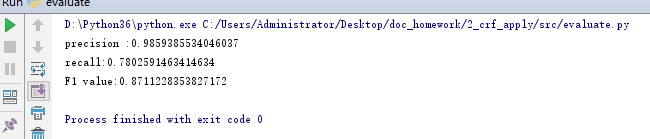


训练生成的模型参见data目录下的model\_10.txt文件。

采用crf\_test.exe -m model\_10.txt test\_data.txt >result.txt命令测试数据。

## 四、实验结果

实验结果生成见source目录evaluate.py文件，具体结果如下图所示：



准确率：98%

召回率：78%

F1值：87%