# Lab2 HMM vs. CRF vs. DNN

## 一、 任务

## 使用 3 种模型实现中文分词任务。

(1) 实现 Hidden Markov Model (HMM)。

(2) 实现 CRF 模型。

参考文献：

《Conditional random fields: probabilistic models for segmenting and labeling sequence data》这篇论文是提出 CRF 模型的首篇论文，主要搞清楚 CRF 的思想和方法，对于模型训练算法可以忽略（因为作者提出的两种算法都并不是很好，后人经过了许多改进）。

Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms》

这是一篇训练 CRF 模型常用的算法之一，想法简单，实现容易。

*此部分的 CRF 模型**不可**使用 torch 等框架实现的 CRF 模型。*

(3) 实现 Bidirectional LSTM+CRF 模型。

参考文献：

《Bidirectional LSTM-CRF Models for Sequence Tagging》

*此部分的 CRF 模型**可**使用 torch 等框架实现的 CRF 模型。*

(4) 数据集：在 2 个数据集上进行实验以体验不同数据集的影响。

数据集文件说明：dataset 文件夹下是两个数据集。train.utf8 是训练集、template.utf8 是特征模板、labels 是标注集合（B 表示词首字、I 表示词中字、E 表示词尾字、S 表示单字词）。对于 template 文件，在训练 CRF 模型时可以自己进行调整以达到较佳性能。

(5) 测试说明：稍后给出

~~dataset/validation 数据集给出了测试样例（面试时会给出新的测试集）。测试时，输入为 input.utf8 文件，模型预测输出到 output.utf8 文件中，gold.utf8 为正确标签。output.utf8 和 gold.utf8 文件中的每行为某个汉字对应的标签。checker.py 文件评估模型输出结果的精确率、召回率与 F1 分数，用法：~~

~~python checker.py --gold=真实结果文件 --output=模型输出结果文件~~

面试时最好提前保存好模型。

(6) 提交较详细的实验报告。

## 二、评分

(1) 实现 HMM 模型，模型能够正确运行并收敛（15%）

(2) 实现 CRF 模型，模型能够正确运行并收敛（15%）

(3) 实现 BiLSTM + CRF 模型，模型能够正确运行并收敛（10%）

(4) 在另外给出的最终测试集上的性能（30%）

(5) 实验文档（30%）

## 三、截止日期

2020 年 12 月 20 日 23: 59 截止。每超过一天，扣 10 % 分数。