Lab2 HMM vs. CRF

一、任务

(1) 实现 Hidden Markov Model (HMM)。
(2) 阅读《Conditional random fields: probabilistic models for segmenting and labeling sequence data》论文，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）。
(3) 阅读《Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms》论文，这是一篇训练 CRF 模型常用的算法之一，想法简单，实现容易。
(4) 根据上述两篇论文实现 CRF 模型。
(5) 将 HMM 应用于中文分词的任务。其中：train.utf8 是训练集、test.utf8 是测试集、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）。
(6) 将 CRF 应用于中文分词的任务（着重体会 CRF 较 HMM 可以任意添加特征来提高模型性能的能力），在训练集进行训练模型，并且测试集测试模型。其中：train.utf8 是训练集、test.utf8 是测试集、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 文件可以自己进行调整以达到较佳性能。
(7) 提交训练和解码（使用 Vertibi 算法）源代码
(8) 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（20%）
(2) 实现 CRF 模型，模型能够正确运行并收敛（40%）
(3) 在另外给出的最终测试集上的性能（20%）
(4) 代码风格和文档（20%）

三、其他

(1) 实验时间是四周，2017 年 12 月 1 日截止。
(2) 这次实验有一定难度，请大家努力尝试和完成。
(3) 如有问题，随时向助教和任课教师询问。