Lab2 HMM、CRF、DNN

声明

- 开发语言不限。
- 出现抄袭现象(包括祖传代码)，抄袭双方均按零分计，面试会对代码提问。
- 请严格按照Deadline 提交，延迟一天扣10分，扣完为止。
- 更多问题可在课程群以及助教个人微信进行提问。
- Deadline：2021.12.19 23:59

一、任务一：七位数彩票预测（30分）

1. 给出1万个七位数彩票号码，每一位都是0-9的数字。预测你认为最可能的10个彩票号码，写入handin.utf8并随代码、文档上交elearning。
2. 提示：将彩票号码视为HMM模型中的观测序列；考虑EM算法；一万个彩票号码之间没有时序关系。

二、任务二：CRF、BiLSTM+CRF 实现中文分词（总分70分）

1. 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。
    - 《Bidirectional LSTM+CRF Models for Sequence Tagging 》这是BiLSTM部分的参考文献。
    - 其他资料可以自行寻找。
2. CRF模型必须手写，不得使用深度学习框架；BiLSTM+CRF模型可以使用深度学习框架。
3. 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。
4. 测试说明：check.py是用于测试准确率的文件。文件夹test_dataset中给出了测试样例，input.utf8是输入的句子，gold.utf8是对应句子的正确标注。面试时将给出测试文件，格式与input.utf8、gold.utf8一致。

三、评分标准

1. 将根据上交的handin.utf8中的10个彩票号码，计算这10个彩票号码的概率之和，以此判断性能分数（15分）
2. 任务一实验文档，包括但不限于你预测的HMM模型参数、实验过程等（15分）
3. 实现 CRF 模型，模型能够正确运行并收敛（20分）
4. 实现 BiLSTM+CRF 模型，模型能够正确运行并收敛（10分）
5. 任务二在测试集上的性能（20分）
6. 任务二实验文档（20分）

7. Bonus：CRF模型可以读取额外的特征模板，或使用高阶CRF等（5分）；在BiLSTM+CRF中使用Pretrain Embedding，如word2vec、BERT等。推荐Glove、Huggingface等python库（5分）。想得Bonus必须理解其原理。