Project2 HMM、CRF、BiLSTM+CRF

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。
- 三个部分统一Deadline：2023.11.30 23:59:59
- 为完成此Project，推荐阅读李航的《统计机器学习》第十、十一章

一、Part1：HMM实现命名实体识别（NER）任务

1. 手写HMM模型，不能使用机器学习框架。
2. 总体思路：

   将实体标签看成隐藏状态，文字看成观测结果。从而把整个问题转化为两步：

   （1）先通过train.txt文件（大量的"观测结果——隐藏状态"序列），估计出隐马尔可夫模型的三个参数：初始概率矩阵、发射概率矩阵、状态转移概率矩阵。（2）利用估计出的HMM模型，使用维特比算法对观测序列进行解码，得到隐藏状态序列，即实体标签序列。

二、Part2：CRF 实现命名实体识别（NER）任务

1. 参考文献：

   （1）《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）。

   （2）《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。

   （3）其他资料可以自行寻找。
2. 可以使用机器学习框架，但是必须理解CRF完成NER的原理，面试时会提问。

三、Part3：BiLSTM+CRF 实现命名实体识别（NER）任务

1. 参考文献：

   （1）《Bidirectional LSTM+CRF Models for Sequence Tagging 》

   （2）其他资料可以自行寻找。
2. 在BiLSTM+CRF模型中，BiLSTM部分可以使用Pytorch等深度学习框架，CRF部分必须手写完成。

四、数据说明

1. 在2个数据集上进行实验，一个中文，一个英文，以体验不同数据集的影响。中文数据集有33种tag标签，英文数据集有9种tag，详细解释见数据集目录下的tag.txt
2. 数据集：train.txt、validation.txt左侧是（单词），右侧是对应的实体标签，中间用空格隔开。
3. 测试说明：NER任务的测试分数计算方式在check.py中给出，以micro avg的f1-score分数为准。运行check.py需要安装sklearn包：

```
pip install scikit-learn -i https://pypi.mirrors.ustc.edu.cn/simple
```

4. 面试：面试时会给出与训练集、验证集格式完全相同的test.txt，要求测试后生成与example_my_result.txt格式相同的结果文件，方便check.py运行。

五、评分标准

1. 实现HMM模型，模型能够正确运行并收敛（25分）
2. 实现CRF模型，模型能够正确运行并收敛（25分）
3. 实现 BiLSTM+CRF 模型，模型能够正确运行并收敛（25分）
4. 面试时的代码解释与问题回答（15分）
5. 实验文档（10分）
6. Bonus：尝试理解老师上课讲过的使用CRF模型进行中文分词的例子，并应用到NER任务上。使用给出的template.utf8中的模板，手写CRF模型（10分）