README

1. 任务描述

利用课程中学过的**隐马尔可夫模型**实现所给数据的命名实体识别。

您需要从训练集中学习，然后从验证集中验证您所构建的HMM模型效果的优良。最终结果得到模型在验证集上的精确率P、召回率R和F1值。

2. 数据集

数据集分为训练集验证集两个文件，分别是`train.json`、`dev.json`。

数据及大小：

- 训练集：10748
- 验证集：1343

标签类别：
数据分为10个标签类别，分别为: `地址（address）`，`书名（book）`，`公司（company）`，`游戏（game）`，`政府（goverment）`，`电影（movie）`，`姓名（name）`，`组织机构（organization）`，`职位（position）`，`景点（scene）`

例子：

> 💡 {"text": "浙商银行企业信贷部叶老桂博士则从另一个角度对五道门槛进行了解读。叶老桂认为，对目前国内商业银行而言，", "label": {"name": {"叶老桂": [[9, 11]]}, "company": {"浙商银行": [[0, 3]]}}}
> {"text": "生生不息CSOL生化狂潮让你填弹狂扫", "label": {"game": {"CSOL": [[4, 7]]}}}

标签定义与规则：

- 地址（address）: **省**区**街**号，**路，**街道，**村等（如单独出现也标记），注意：地址需要标记完全, 标记到最细。

- 书名（book）: 小说，杂志，习题集，教科书，教辅，地图册，食谱，书店里能买到的一类书籍，包含电子书。
- 公司（company）:**公司，**集团，**银行（央行，中国人民银行除外，二者属于政府机构）, 如：新东方，包含新华网/中国军网等。
- 游戏（game）: 常见的游戏，注意有一些从小说，电视剧改编的游戏，要分析具体场景到底是不是游戏。
- 政府（goverment）: 包括中央行政机关和地方行政机关两级。 中央行政机关有国务院、国务院组成部门（包括各部、委员会、中国人民银行和审计署）、国务院直属机构（如海关、税务、工商、环保总局等），军队等。
- 电影（movie）: 电影，也包括拍的一些在电影院上映的纪录片，如果是根据书名改编成电影，要根据场景上下文着重区分下是电影名字还是书名。
- 姓名（name）: 一般指人名，也包括小说里面的人物，宋江，武松，郭靖，小说里面的人物绰号：及时雨，花和尚，著名人物的别称，通过这个别称能对应到某个具体人物。
- 组织机构（organization）: 篮球队，足球队，乐团，社团等，另外包含小说里面的帮派如：少林寺，丐帮，铁掌帮，武当，峨眉等。
- 职位（position）: 古时候的职称：巡抚，知州，国师等。现代的总经理，记者，总裁，艺术家，收藏家等。
- 景点（scene）: 常见旅游景点如：长沙公园，深圳动物园，海洋馆，植物园，黄河，长江等。

3. HMM

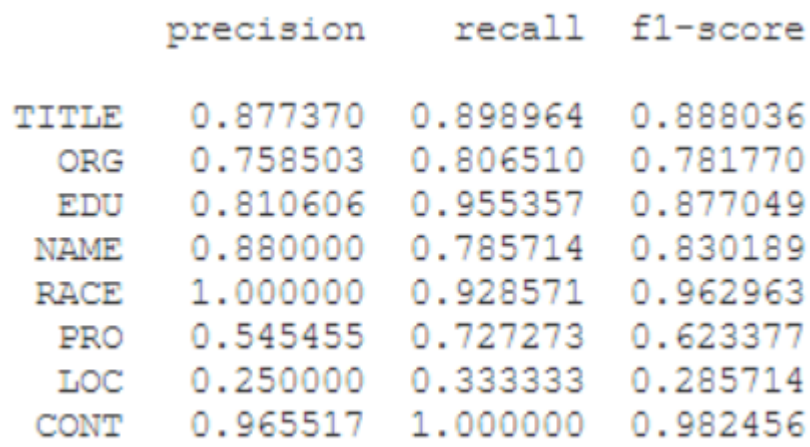
隐马尔可夫模型是关于时序的概率模型，描述由一个隐藏的马尔可夫链随机生成不可观测的状态随机序列，再由各个状态生成一个观测而产生观测随机序列的过程。隐马尔可夫随机生成的状态的序列，称为状态序列；每个状态生成一个观测，而由此产生的观测的随机序列，称为观测序列。

特别注意：**隐马尔可夫模型是生成模型，先生成状态序列，然后由状态序列生成观测序列。即是先$P(Z)$，再$P(O|Z)$，所以拟合的是$P(O,Z)$，也就是联合概率分布**。

流程：

1. 首先根据所给数据集，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. 编写维特比算法对输入的观测序列进行解码，得到状态序列

3. 结果评估。预测序列与验证集中的真实标签序列进行对比，计算出精确率P、召回率R和F1值。类似于下图（该图并不是本实验的结果图）。



	precision	recall	f1-score
TITLE	0.877370	0.898964	0.888036
ORG	0.758503	0.806510	0.781770
EDU	0.810606	0.955357	0.877049
NAME	0.880000	0.785714	0.830189
RACE	1.000000	0.928571	0.962963
PRO	0.545455	0.727273	0.623377
LOC	0.250000	0.333333	0.285714
CONT	0.965517	1.000000	0.982456

TIPS

1. 构造标签字典

```
tagDict = {'O': 0, 'B-ADDRESS': 1, 'I-ADDRESS': 2, 'B-BOOK': 3, 'I-BOOK': 4, 'B-CO
M': 5, 'I-COM': 6, 'B-GAME': 7, 'I-GAME': 8, 'B-GOV': 9, 'I-GOV': 10, 'B-MOVIE': 1
1, 'I-MOVIE': 12, 'B-NAME': 13, 'I-NAME': 14, 'B-ORG': 15, 'I-ORG': 16, 'B-POS': 1
7, 'I-POS': 18, 'B-SCENE': 19, 'I-SCENE': 20}
```

2. 将数据集使用BIO进行标注，这样每一个字都有它自己的标签，标签就是标签字典里的某一个。便于构造初始概率矩阵、发射概率矩和状态转移概率矩阵。