**作业2 - 概率图模型练习**

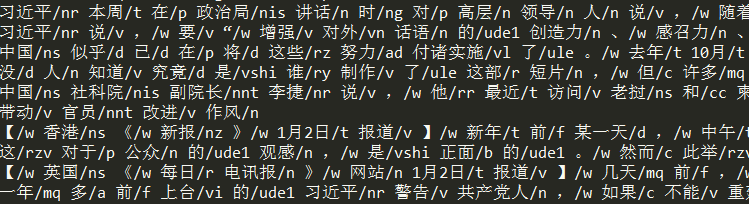
**一、任务：词性标注**

**词性标注**是为分词后的每个单词标注一个正确的词性。

**要求：**用隐马尔科夫模型实现词性标注。

**数据资源：**语料来自1998年1月份人民日报语料。

**语料格式：**



**语料使用方法：**

（1）采用合理的方式对数据进行预处理；

（2）按照8:2比例，将数据划分为训练数据和测试数据；训练数据部分用来训练模型参数，测试部分用来验证模型效果。

**模型验证：**

采用交叉验证方式，**指标：**准确率。

**二、模型说明**

本次实验通过编写hmm相关代码进行词性标注。主要分为以下几个部分：

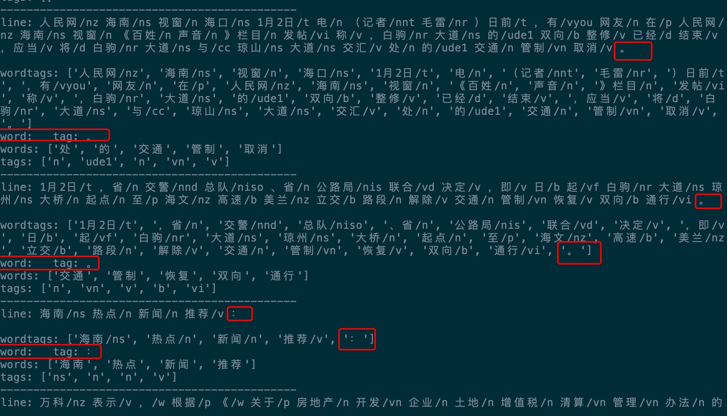
1）读取数据，建造字典，以及字典映射，，并存储在数据结构中

2）HMM学习过程：统计获取发射概率矩阵、状态转移矩阵和观测矩阵

3）HMM预测过程：使用viterbi算法进行预测，并使用准确率评估预测结果

**三、问题与解决**

**问题1**：语料中有许多没有被标注的词以及多余的空格，如下图所示：

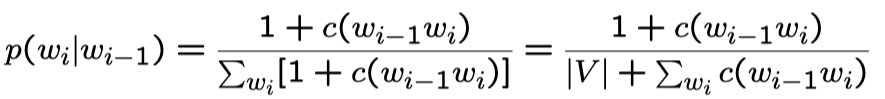


**解决方法**：经过对这些异常的数据输出发现，都是因为没有标记造成的，而且这些没有标记的要么是空格，要么是标点符号，所以我们都使用统一的标注‘w’进行标注。

**问题2**：由于某些字在统计发射概率时，这个词没有在训练语料中的每段话的第一位出现，所以统计计算后概率为0，这就使得测试预料中以该词开始的那句话无法进行正确词性标注（概率全为0），这种情况同样出现在状态转移矩阵和观测矩阵。

**解决方法1**：做平滑，给每个概率为0的值赋一个很小的值（本文赋值为字典大小的倒数），此外也有其它的平滑技术，如add one、goodturing等等。

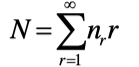
**解决方法2**：add one平滑算法



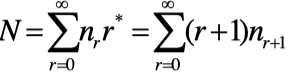
**解决方法3**：good turing平滑算法

I.J.Good1953年引用Turing的方法来估计概率分布。假设N是样本数据的大小，nr是在样本中正好出现r次的事件的数目（在这里，事件为n-gramw1,w2,…,wn）即：出现1次的n1个，出现2次的n2个,……。

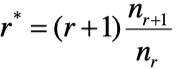
那么，



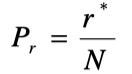
由于，



所以，



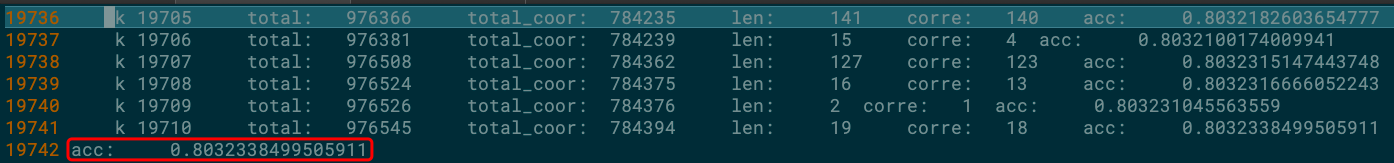
那么，Good-Turing 估计在样本中出现r次的事件的概率为：



**四、结果**

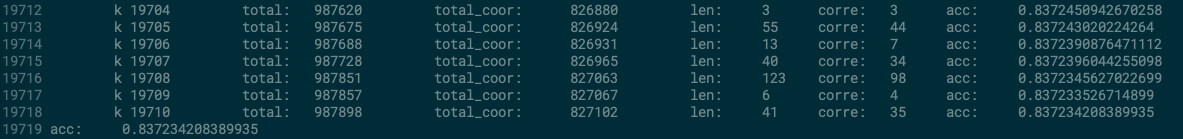
采用8:2交叉验证的方法，在分训练集和测试集之前，对数据shuffle。

1）使用第一种平滑策略：



准确率为：acc: 0.803

2) Add-one smoothing



准确率为：acc: 0.837

可以发现使用addone平滑之后，其评测正确率有了明显提高。

3) Good-Turing

目前深入理解了这个方法的原理，在实现的过程中发现一些问题，造成最终的预测结果很糟糕，但理论上该平滑策略好优于以上两种策略。