作业1：

基于THCHS30中文语料，尝试采用modified Knser-Ney算法训练相应的语言模型并计算困惑度。

THCHS-30语料：

THCHS-30\data\_thchs30\data:存放所有wav文件和trn文件(trn一共有三行，第一行为分词好的句子，第二行拼音，第三行为音素)

THCHS-30\data\_thchs30\train:存放用于训练的wav文件和trn文件(trn文件存放的是wav对应data里trn文件的路径)，一共10000条。

THCHS-30\data\_thchs30\test:存放用于测试的wav文件和trn文件(trn存放的是wav对应data里trn文件的路径)，一共2495条。

THCHS-30\data\_thchs30\dev同上

除数据集之外还包括训练好的语音模型word.3gram.lm和phone.3gram.lm以及相应的词典。

用SRILM工具建立语言模型分为两步：

①统计语料库生成n-gram统计文件

ngram-count -vocab lexcion.txt -text train.txt -order 3 -write train.count

-vocab lexcion.txt:读取词典lexcion.txt，如果计数文件中或者训练文件中，如果出现了词典之外的词(OOV),这些词会被替换为<unk>

-text train.txt:根据训练文件生成N-gram计数文件，保存到train.out中

②生成语言模型:

ngram-count -read train.count -order 3 -lm.train.lm interpolate -kndiscount

-read train.count 读取我们上一步获得的计数文件

-lm train.lm:生成语言保存至train.lm中

Interpolate -kndiscount:采用插值式的修正Kneser-Ney折扣算法

计算困惑度ppl:

ngram -ppl test.txt -order 3 -lm train.lm

结合测试语料和上一步得到的语言模型，计算困惑度。困惑度越低，对自己越自信，表现更好。

在计算之前发现THCHS语料文本中有很多重复的句子，比如A0-1,A1-1,......的内容都是一样的。经过检测发现，每条句子都至少重复7次，这样会导致，所有词条数目的频率均大于7，没有小于7的词条出现，也就无法进行平滑算法了。因而，分为取出train对应trn里的第一行和test对应trn里的第一行进行去重，原先10000条train文件经过删减剩下750条，2495条test文件剩下250条，分别保存在mytrain.txt和mytest.txt中。

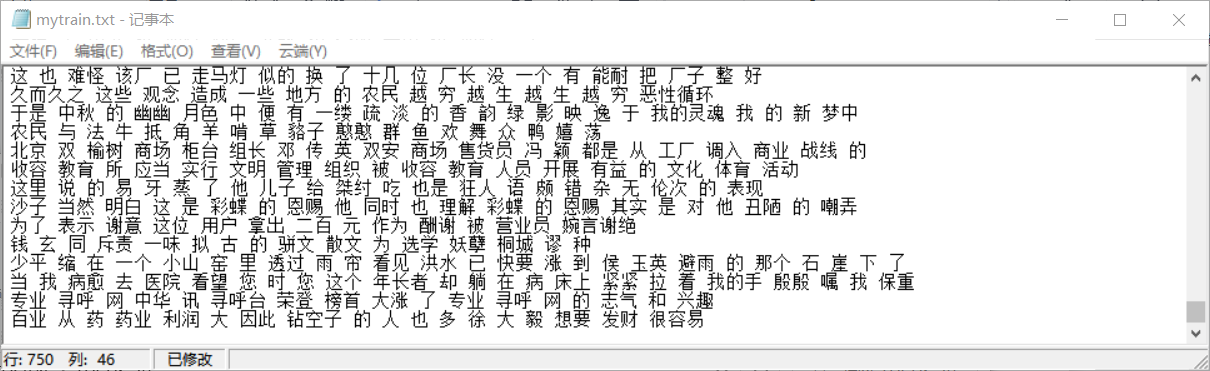


图1. mytrain.txt

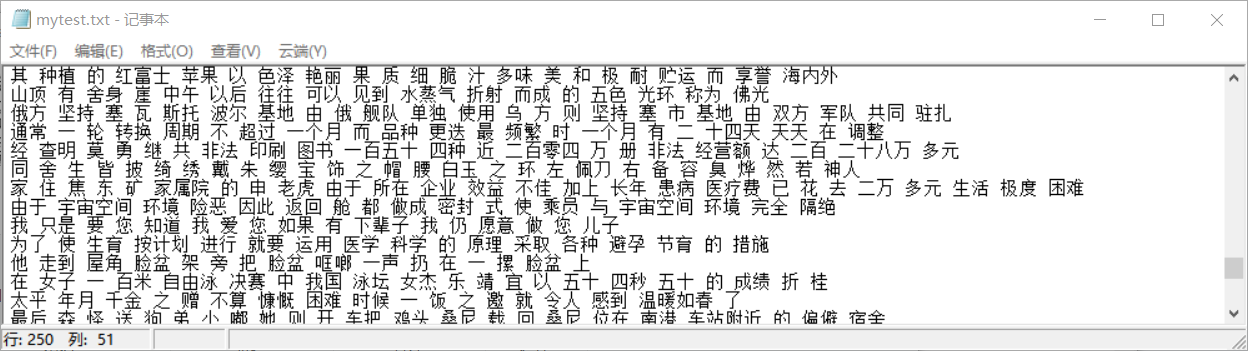


图2. mytest.txt

下面进行语言模型训练:  
①生成count文件:

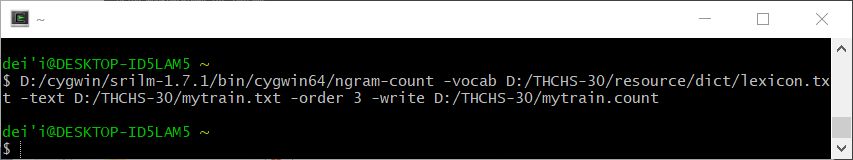


图3. mytrain.count

②生成语言模型:

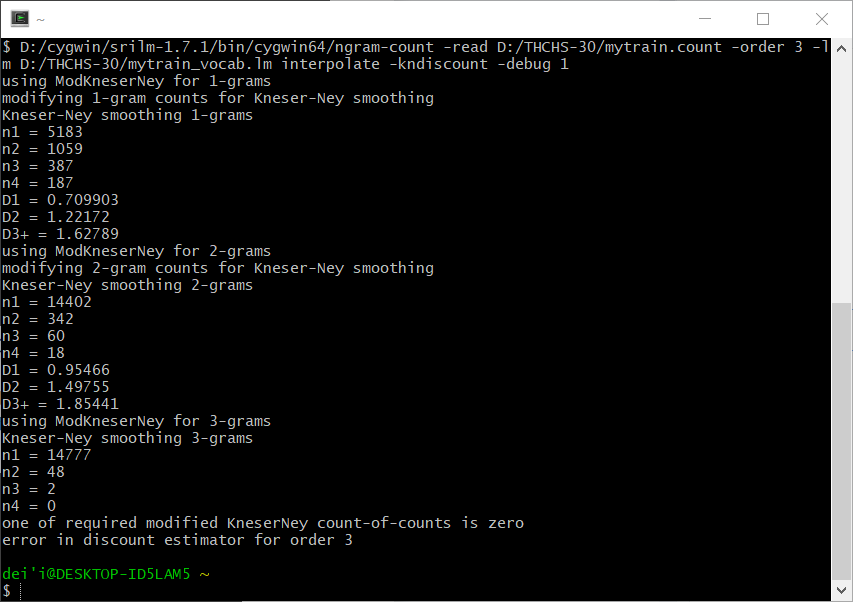


图4. 生成语言模型

可以看到，因为3gram的参数n4 = 0，频率大于4次的3gram的词条不存在，从而出现错误，无法采用插值式的修正Kneser-Ney折扣算法，这是语料太少导致的。

解决办法：①采用不修正也就是原先的Knser-Ney算法：

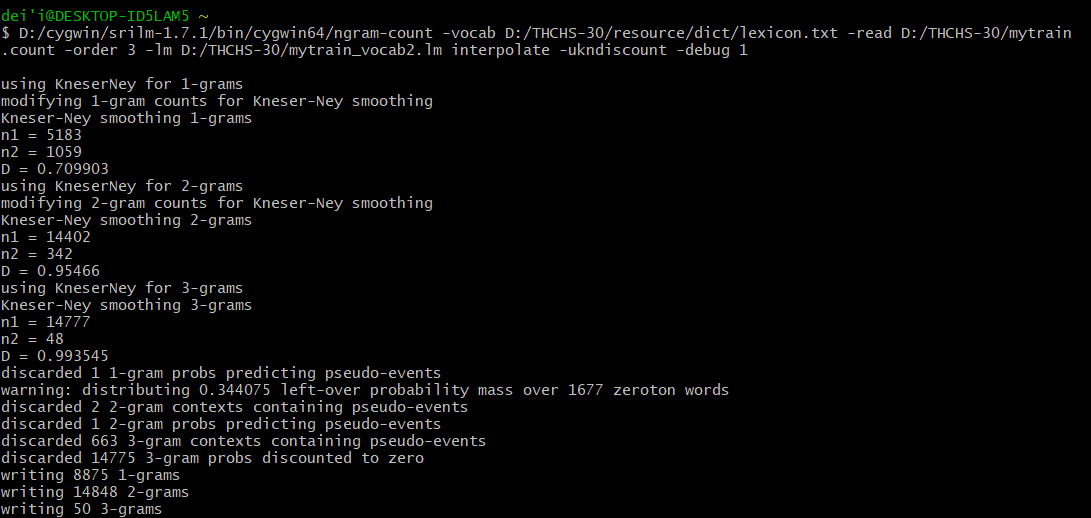


图5. Knser-Ney

语言模型保存为mytrain\_vocab2.lm

②根据ppt提示：有的时候数据集太小，通常会发现对于高阶N-gram出现没有n\_x统计技术，这个时候常换为witten bell。

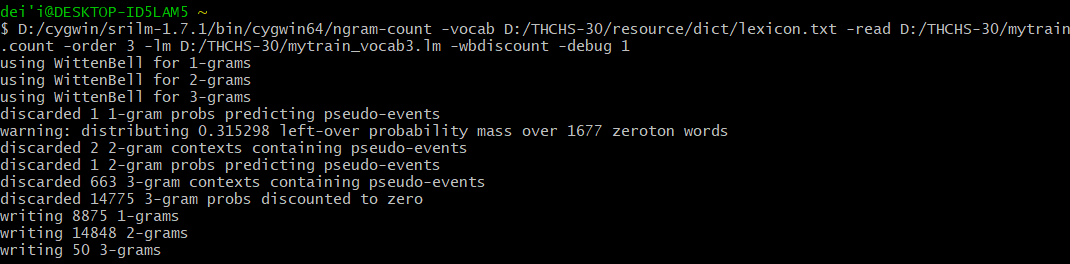


图6. 语言模型

语言模型保存为mytrain\_vocab3.lm

针对俩个语言模型计算困惑度:

①Knser-Ney：

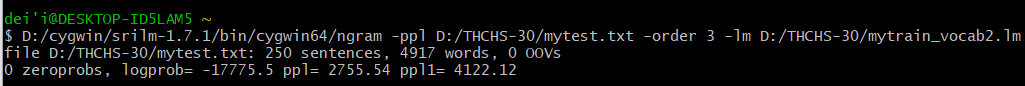


图7. Knser-Ney

ppl = 2755.54 ppl1 = 4122.12

②witten bell

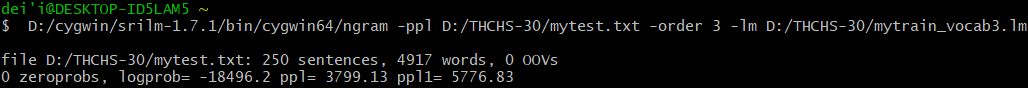


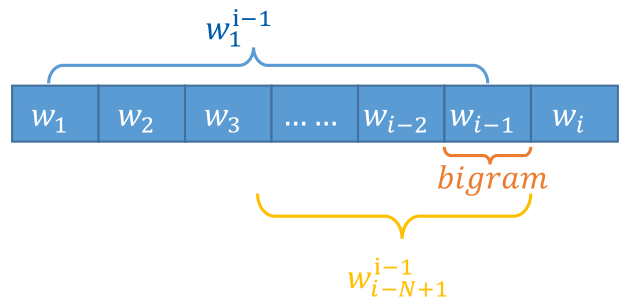
图8. Witten bill

ppl = 3799.13 ppl1 = 5776.83

结果表明，通过原始Knser-Ney算法训练得到的模型在测试集上表现好于witten bell。当然由于训练集和测试集样本较小，困惑度的比较很大的参考意义，只是通过一个小的实验用于展示SRILM工具包的使用方法。

作业2：写出插值语言模型的递归格式，并根据递归规格是给出插值系数的EM算法公式：

递归公式：





......

递归格式为:





