## 基于N-gram的最大概率中文切分实验

实验实现语言：python

运行环境：Python 2.7.10

实验要求：

实现一个中文切分程序。

实验步骤：

1. 首先读入训练语料corpus\_for\_ass2train.txt，n-gram中n的取值，最大匹配长度maxlen；对原始语料进行预处理，清除回车换行字符，按空格生成词典集合；
2. 然后打开测试语料corpus\_for\_ass2test.txt，对其进行一句一句的处理，对于每一句分别用最大匹配算法对其进行双向扫描，若有多种切分方式，则计算不同切分方式对应的概率，选取切分概率最大的一种方式作为该句最终的切分路径；依此循环读取，直至将测试语料读取并处理完毕，将结果写入相应的结果文件中
3. 将切分结果同正确答案corpus\_for\_ass2GS.txt比对，进行评估，得出正确率、召回率、F-1指数。

实验操作说明：

1. 输入相应的参数，运行代码



入参有5个：

第一个是切分程序：Segmentation.py

第二个是训练数据：corpus\_for\_ass2train.txt

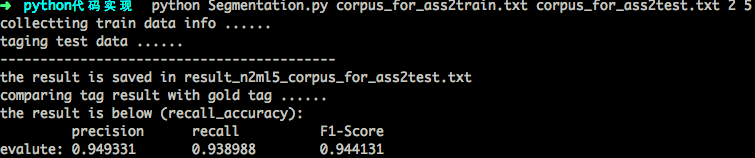
第三个是测试数据：corpus\_for\_ass2test.txt

第四个是n-gram中n的值：2

第五个是最大匹配时选取的maxlen：5

其中第四、第五个参数的值可以取不同的值，此处分别取2、5是对应准确率最高的参数。

1. 测试数据的结果会存入，result\_n2ml5\_corpus\_for\_ass2test.txt,其中2、5是根据入参得来的，选取不同的参数会得到不同的结果文件名，corpus\_for\_ass2test.txt对应于测试文件名；
2. 实验结果展示



#### OOV的能力

由于训练语料的局限性，本次实验中将没有出现在训练语料corpus\_for\_ass2train.txt中的未登录词作为一个汉字独立切分。

那么这样切分的过程中就会有一些词并不在训练语料中，那么依据最大似然估计，这些词的概率将会为0，但是实际上这些词的概率并不是0，对于此类数据稀疏的问题导致句子的切分概率为0的情况，我们需要对其进行平滑处理，本实验采用Add-one拉普拉斯加法平滑，即每类词在训练语料中出现的次数都加一：

Count\*(word) = count(word) + 1

即规定没有出现过的n-gram在训练语料中出现了一次，没有出现过的n-gram的概率不再是0；

#### 处理歧义

本实验采取双向最大匹配算法来消除部分歧义：

首先分别根据前向后向最大匹配法，得到多种切分结果；

其次对于不同的切分结果，基于马尔可夫假设计算句子的概率，取最大概率的那个切分路径。

#### 评估方法及评估结果

采用的评估方法为： 将语料测试文本的切分结果与正确答案进行比对。采用准确率(P)、召回率(R)及F1值进行评估；

A:实验结果中正确切分的词数

B:实验结果中词总数

D:黄金分词的词总数

1）正确率

2）召回率

3）F值

是正确率和召回率的调和平均数，用来综合平测系统的综合效果：

本次实验取,即：

本次切分实验结果为：

正确率：94.9331%

召回率：93.8988%

F-1值：0.944131