作业：基于最大概率的汉语切分

作业内容：采用基于语言模型的最大概率法进行汉语切分。其中： n-gram语言模型要求n >1，至少采用一种平滑方法；或采用神经语言模型。

11月29日：发布测试数据： test.txt

12.月2日： 提交作业，提交内容包括如下3个部分：

完整的算法代码、完整的算法说明文档(用到的算法描述、训练语料描述等)、算法在测试数据上的切分结果

切分结果文件说明：

1) 结果输出到一个纯文本文件，文件名为：学号.txt

2）切分单元之间用一个字节的空格“ ”分隔，如：我们 在 学习 。

3）标点符号算一个切分单元。

注：请严格按此格式输出，以便得到正确评测结果

评分：

|  |  |
| --- | --- |
|  | 分 |
| 切分结果评测 | F\*15, F=2P\*R/(P+R) |

特别注意：代码雷同问题

评估方法：

比较切分位置，两个连续的切分位置一致时，表明该处的切分正确，否则错。

评估方法执行示例：

以评估如下结果为例：

标准答案：本报 讯 春节 临近

模型输出：本 报 讯 春节 临近

1．首先给句子按字编号，如 0本1报2讯3春4节5临6近7

2．依据该标号，标准答案中的切分位置在2、3、5、7位，即标准答案的切分序列为2357（序列长度为4也表明切分结果包含4个词），而模型输出的切分位置12357，比较这两个序列里的相同的长度为2的连续数字串的数目就是正确切分的数。例如，标准答案2357中长度为2的连续字串为23，35，57，而模型输出12357中长度为2的连续字串为12，23，35，57，相同的有3个，即切分正确了3个。

模型输出了5个词，标准答案含4个词， 因此：

P=3/5 (切了5个词对了3个)，

R=3/4 (一共4个正确的词，找到了3个)。

又如：

输出：本 报讯 春节 临近

则有2个对，P=2/4(切了4个词对了2个)， R=2/4(一共4个正确的词，找到了2个)。