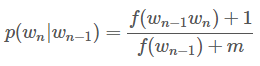
**算法说明文档**

采用的是二元语法（2-gram）

平滑方式是采用的+1 平滑方式，也就是：



*对于f(wn−1wn)=0f(wn−1wn)=0的情况，式子会退化成 p(wn|wn−1)=1f(wn−1)+mp(wn|wn−1)=1f(wn−1)+m。也就是说，如果前词频率越大，则认为后词出现的概率就越小。*

语料库文件名为 北大(人民日报)语料库199801.txt

文件编码格式为无BOM的utf-8

测试集为 testset.txt

文件编码格式为无BOM的utf-8

算法在测试数据集上的切分结果在 2017110758.txt文件里

文件编码格式是gbk

完整算法代码为 test.py