# 切词说明文档

训练语料：北大(人民日报)语料库199801.txt

语法：2-gram语法

训练过程：

先初始化一个字典嵌套字典*dic = defaultdict(lambda: defaultdict(lambda: 0))，*作为语料库中每个词左邻词的记录，之后对第一层字典中的每项取*value/sum(value)*作为每个词左邻词的概率，以此方法构建好词典。

之后分别使用HMM、FMM两种方式读入训练语句，以此找到训练语句分词的分歧。如果HMM与FMM得到的分词结果相同，则直接返回此分词结果；如果产生分歧，则从分歧处开始计算两种分词策略出现的概率，*P(w1,w2,w3)=P(w1｜Start)P(w2|w1)P(w3｜w2）P（End｜w3）*，选取概率较大的一种切分策略。