分词初步

> Written by Native.S1mple

1. **分词基本原理：现代分词基本都是基于统计的分词，统计的样本内容来自于一些标准的语料库，例如：“小明来到荔湾区”。**
   从统计的角度来讲，期望“小明/来到/荔湾/区”这个分词后出现的概率要较大一些。State it in math:
   有一个句子S,有m种分词选项：

   $A_{11}A_{12}.....A_{1n_1}\qquad A_{21}A_{22}.....A_{2n_2}$

   $..................\qquad A_{m1}A_{m2}.....A_{mn_m}$

   n_i表示第i种分词的个数，如果我们从中选择了最优的第r种方案，那么这种分词方法对应的统计分布概率应该最大，即:$r=argmaxP(A_{i1},A_{i2},\ .....A_{in_i})$
   但是上述概率分布并不好求解，因为涉及到n_i个分词的联合分布，在NLP种，为了简化计算，我们通常使用马尔可夫假设⇒每一个分词出现的概率仅仅和前一个有关，即
   $P(A_{ij}|A_{i1},A_{i2}.....A_{i(j-1)})=P(A_{ij}|A_{i(j-1)})\Rightarrow P(A_{i1},A_{i2}.....A_{in_i})=P(A_{i1})P(A_{i2}|A_{i1})P(A_{i3}|A_{i2}).....P(A_{in_i}|A_{in_{i-1}})$**而**$P(w1|w2)=\frac{P(w1,w2)}{P(w2)}\approx\frac{freq(w1,w2)}{freq(w2)}$ **⇒语料库直接统计即可.找联合分布概率最大⇒最优分词**

2.**N元模型：假如只依赖于前一个模型，确实过于武断。那我们认为现在可以依赖前两个词⇒**
$P(A_{i1},A_{i2},A_{i3}.....A_{in_i})=P(A_{i1})P(A_{i2}|A_{i1})P(A_{i3}|A_{i1},A_{i2}).....P(A_{in}|A_{i(n-2)},A_{i(n-1)})$**这样当然是合理的，但联合分布的计算量就大大增大，我们一般称只依赖于前一个词的模型为二元模型，前三个词⇒三元模型，以此类推。越往后，概率分布的计算复杂度越高，算法原理类似。实际应用，N一般小于4，主要是因为概率分布的空间复杂度为$O(|V|^N)(|V|)$为语料库大小，N为模型元数。**
问题：1.某些生僻词，或者相邻分词联合分布在语料库中没有，概率为0.⇒使用拉普拉斯平滑，即给一个较小的概率值。2.句子长，分词情况多，计算量大。

3.**维特比算法与分词：对于一个有很多分词可能的长句，我们当然可以用暴力方法求解所有分词可能的概率，再找出最优分词方法。但是维特比可以大大简化求出最优分词的时间。(动态规划)**
维特比算法⇒隐式马尔可夫模型HMM模型解码算法，但它其实是一个通用的求序列最短路径的方法，不光可用于HMM也可用于其他的序列最短路算法。(概率图模型算法)

4。**常用的分词工具：**
中文：Jieba　　THULAC-python
Eng: nltk