**统计语言模型中的平滑算法**

# 平滑算法

## 古德图灵Good-Tuning

### 基本思想

古德图灵平滑的基本思想是“劫富济贫”，把从训练集中统计得到的词的频数分一部分给没有在训练集中出现的词语。

### 平滑流程

我们假设训练集中所有词的频数为：，词语的频数为：，频数的频数为： （表示频数为r的词语的个数为）

1. 聚类。是对词语的频数进行聚类。统计完词语的频数以后，对词语的频数再进行统计。也就是统计出出现1次的词语有多少，出现两次的词语有多少，以此类推。
2. 平滑。聚类结束以后，按照如下公式进行平滑：

通过上面的公式，就把原来的频数降低为。

但是古德平滑有一个问题：

假如其中某一个为0，此时就没法算了。那么此时怎么办呢？需要使用其它算法方法进行配合。

### 古德平滑实例

语料为

Sam I am I am Sam I do not eat

语料统计结果：

do 1次

not 1次

eat 1次

am 2次

Sam 2次

I 3次

总频数：N = 10

聚类：

1. 3
2. 2
3. 1

总频数：N=3+2\*2+3\*1=10

古德平滑：

1次的平滑：（1+1）\*2/3 = 1.33

2次的平滑：（2+1）\*1/2=1.5

3次的平滑：3

总频数：1.33 \* 3 + 1.5 \* 2 + 1\*3 = 9.99

这样就相当于把0.01的频数给了不可见的词语。

## Katz