数学之美

n-gram model：
markov assumption: the probability of a word occurring only depends on its previous n words

Limitation of n-gram model:
- computational difficulty: dimension increases with N, operation time O(V)**(n-1) for V number of vocabulary
- cannot capture the relations of words even after increasing N—some are related across paragraphs (long-distance dependency)
- cases of unseen word or words that occurred only a few times

Good- Turing estimate: to tackle unseen words:
From the total mass, choose a small mass and assign to the unseen words. This is done by reducing the relative frequency of the seen words. (page 69)

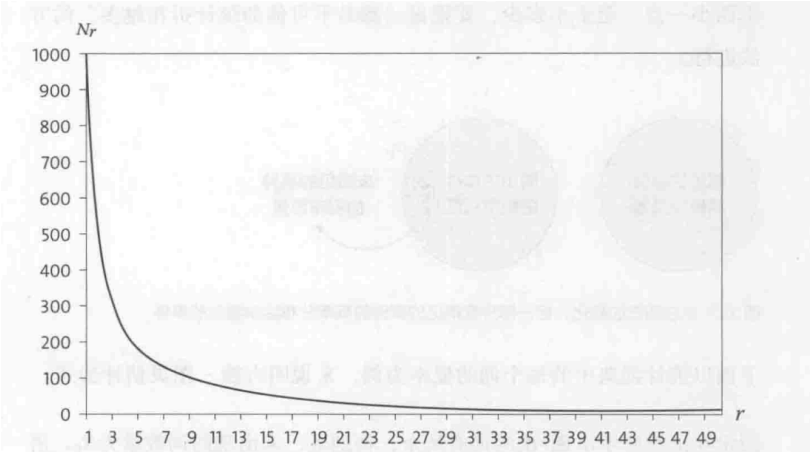

图 3.2 Zipf 定律：出现 r 次词的数量 N_r 和 r 的关系

可以看出 r 越大，词的数量 N_r 越小，即 $N_{r+1} < N_r$。因此，一般情况下 $d_r<r$，而 $d_0>0$。这样就给未出现的词赋予了一个很小的非零值，从而解决了零概率的问题。同时下调了出现频率很低的词的概率。当然，在实际的自然语言处理中，一般对出现次数超过某个阈值的词，频率不下调，只对出现次数低于这个阈值的词，频率才下调，下调得到的频率总和给未出现的词。

The relative frequency of high-occurring words will not be affected while low-occurring words will be reduced

Good- Turing estimate: to tackle single occurring pairs:
- reduce the effect

中文分词:
- 最长匹配
- 统计语言模型: possibility of one sentences occurring as compared to other word tokenization results.

- 词的颗粒度：实现不同层次的词的区分：清华大学 and “清华”“大学”

文章分类：
tf-idf，用 vector 计算不同文章的相似度，相似的合成一个 subclass，再往上合成更少的 class

- stopwords
- 文章开头、结尾，topic sentences 加权