**古德图灵估计**

# 简介

在很多语言相关的任务中，知道一个句子在一个文档中的概率是非常有用的。然而，我们往往没有足够的数据量来统计所有的句子。因此，N-Gram模型通常用来进行估计一个句子的概率。对于N-Gram模型来说，即使进行一些独立性的假设可以减少数据稀疏问题，但是数据稀疏问题是不可避免的。比如：谷歌有10亿英文句子的语料。在这些语料中，词频超过200的词语有1300万，但是在这些语料中没有出现的二元词语至少还有1690亿，远远多于语料中的10亿。对于数据的稀疏问题，一般是使用平滑策略来解决。在这篇论文中，我们介绍**古德图灵平滑**算法。

# 假设

假设所有可能句子的集合为：。这些句子可能是N-Gram，但是为了简单起见，我们假设它们都是一元词语。比如：。

我们还有个独立的句子样本,其中。