# 通过中文语料库来验证Zipf's Law

**Abstract：**

本文利用已有的中文语料库，先对中文进行提取，再根据哑巴分词，统计汉字词语的频率并将其进行排序，绘制词语的排名与频率的关系图，使用对数坐标轴，当图中的数据点大致落在一条直线上，且符合对数关系，即可验证Zipf's Law。

## Introduction

Zipf's Law是指为：在自然语言的语料库里，一个单词出现的频率与它在频率表里的排名成反比。频率最高的单词出现的频率大约是出现频率第二位的单词的2倍，而出现频率第二位的单词则是出现频率第四位的单词的2倍。即在给定语料中，对于任意一个单词，其频率（Frequency）与频率排序（Rank）乘积大致是一个常数，即：Rank \* Frequency ≈ Constant

## Methodology

### M1 ：验证步骤

利用中文语料库验证Zipf's Law，根据以下步骤：

1、将语料录入。

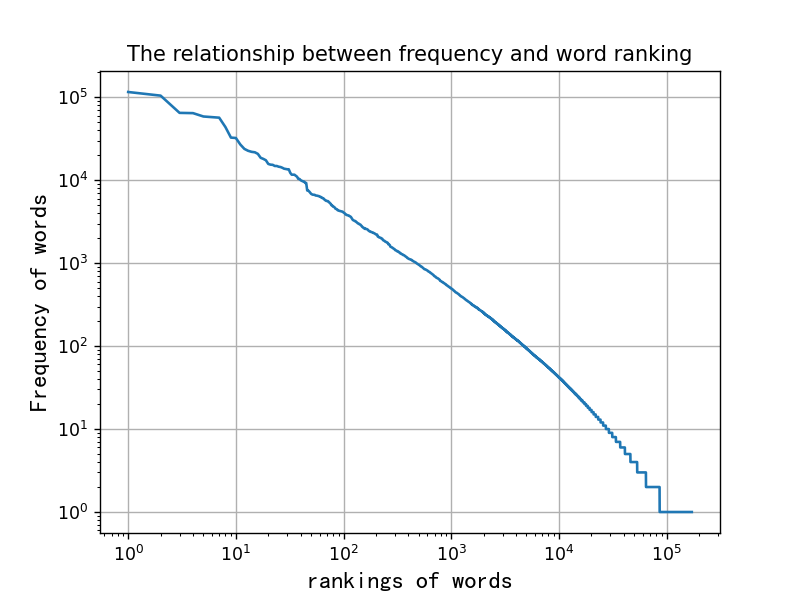
2、对语料库中的词语进行分词，并统计每个词语的出现频率。

3、将词语按照频率从高到低进行排序。

4、绘制词语的排名与频率的关系图，使用对数坐标轴，即横坐标为词语的排名（取对数），纵坐标为词语的频率（取对数）。

5、如果图中的数据点大致落在一条直线上，且符合对数关系，那么就验证了Zipf's Law。

## Experimental Studies



如图所示，频率和词语的排名对应的数值点大致落在一条直线上，即可验证Zipf's Law。

# 计算中文的平均信息熵。

**Abstract：**

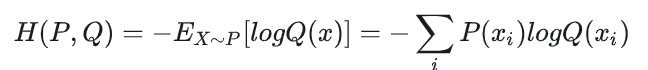
根据文献中提出的多元模型，利用该模型来计算中文(分别以词和字为单位) 的平均信息熵，其中语料库为所提供的中文语料库。

## Introduction

信息熵的概念最早由香农（1916-2001）于1948年借鉴热力学中的“热熵”的概念提出，旨在表示信息的不确定性。熵值越大，则信息的不确定程度越大。其数学公式可以表示为：



但由于事物的实际概率往往是未知的需要估计，而所估计的概率与真实概率又存在差值，因此引入交叉熵的概念，交叉熵（Cross Entropy）是衡量两个概率分布之间差异的一种方法。公式可以表示为：



对于两个概率分布P和Q，其中P为真实分布，Q为预测分布。

本论文利用多元模型预测概率分布，进而求得信息熵。

## Methodology

### M1 ：信息熵

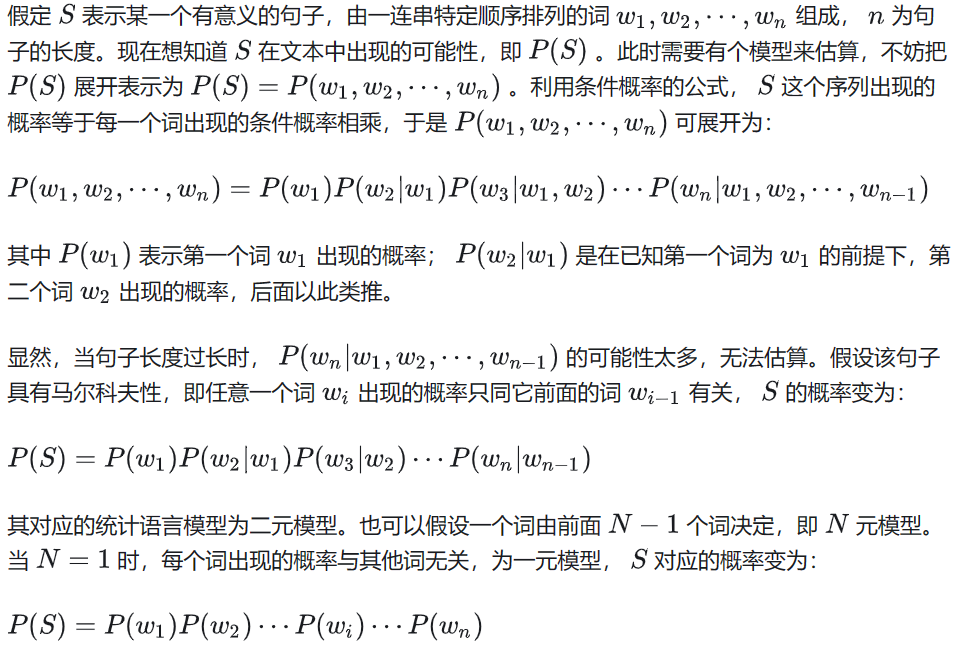
信息熵是信息论中的一个重要概念，用于衡量一个随机变量的不确定性或者信息量的大小。信息熵最初由香农在他的《通信的数学理论》中提出，被广泛应用于通信、数据压缩、密码学、统计学等领域。

在信息论中，一个离散型随机变量的信息熵表示为：

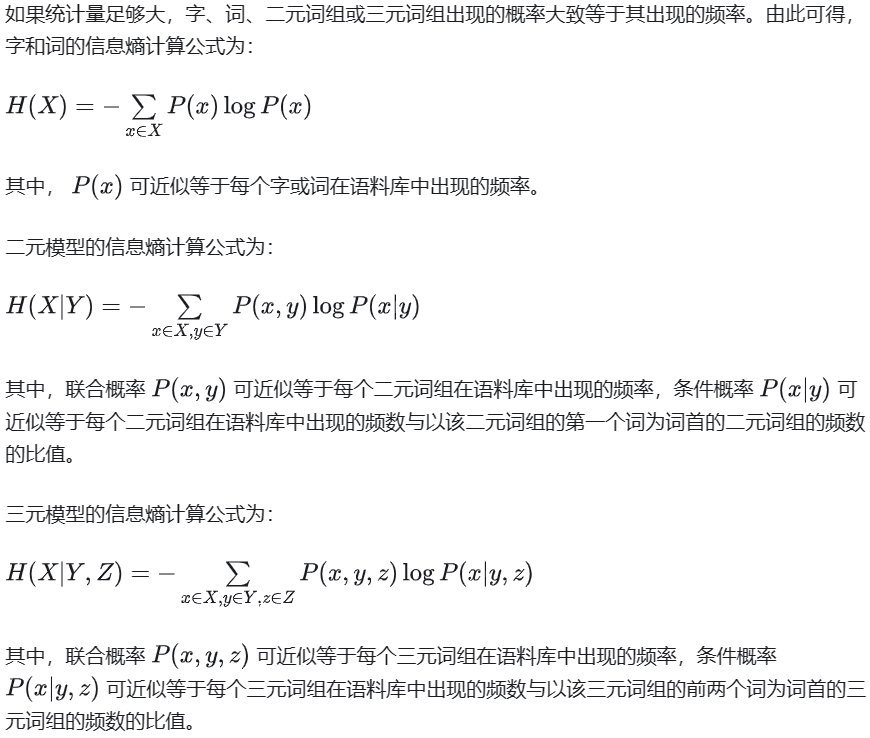


信息熵的单位通常是比特（bit），表示信息的量。信息熵越高，表示随机变量的不确定性越大，需要更多的信息来描述。

## M2 ：统计语言模型



### M3 ：信息熵计算



### M4 ：实验步骤

1. 将文件夹里的语料枚举输入
2. 分别读取txt文件，并进行预处理，删除文章内的所有非中文字符，以及和小说内容无关的片段，得到字符串形式的语料库。
3. 按照”分词“和”分字符“两种不同模式生成词频字典，在”分词“模式下，用jieba库中的cut函数对原始语料库进行处理，在”分字符“模式下，使原始语料库
4. 根据一元、二元、三元的相关公式获取得到词频表
5. 根据词频表的概率，利用相关公式，求解各个文件的以字和词两种形式的信息熵

白马啸西风,碧血剑, 飞狐外传, 连城诀,鹿鼎记, 鹿鼎记,射雕英雄传,神雕侠侣,书剑恩仇录, 天龙八部,侠客行,笑傲江湖,雪山飞狐,倚天屠龙记,鸳鸯刀,越女剑

## Experimental Studies

通过python的计算，获得了各个txt文件以字和词，在一元、二元、三元模型下的平均信息熵