**Verification of Zipf Law in Natural Language Processing and Calculation of Chinese Average Information Entropy**

**自然语言处理中的zipf law的验证和中文平均信息熵的计算**

Ruofei Wang

wrf15539223579@163.com

**Abstract**

本报告旨在验证自然语言处理中 Zipf 定律，并计算给定中文语料库的平均信息熵。本报告验证了在大规模语料库中，词频与其排名的乘积近似为常数，计算了语料库的字，词的一元、二元、三元平均信息熵，为中文文本的信息理解提供了参考。

This report aims to verify Zipf's law in natural language processing and calculate the average information entropy of a given Chinese corpus. This report verifies that in large-scale corpora, the product of word frequency and its ranking is approximately constant. The average information entropy of words in the corpus is calculated, which provides a reference for understanding information in Chinese texts.

**PART 1: Zipf’s Law**

**Introduction**

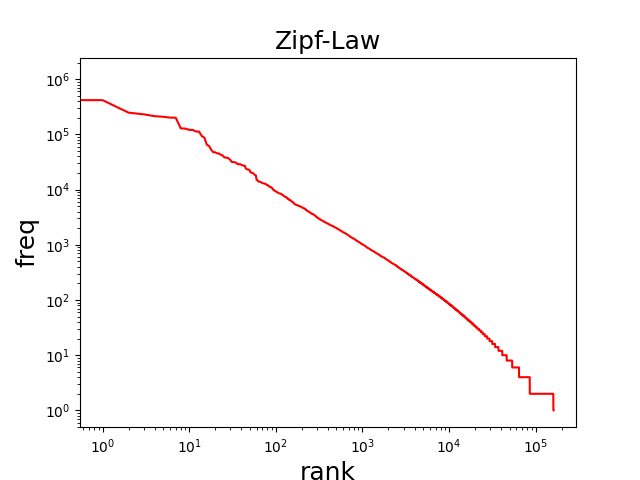
齐夫定律是由哈佛大学的语言学家乔治·金斯利·齐夫（George Kingsley Zipf）于1949年发表的实验定律。它可以表述为：在自然语言的语料库里，一个单词出现的频率与它在频率表里的排名成反比。

**Methodology**

在这部分的验证过程中，首先将不同文本导入到同一个列表中，主要利用jieba函数对文本进行了分词，统计词频并进行排序，使用正则匹配结果并计算乘积，最后使用matplotlib绘图。

**Experimental Studies**

最终绘制的图像如下图。图像显示，在对数坐标轴下，词频关于排名大约是一条斜率为负一的直线，这说明词频与其排名的乘积近似为常数，齐夫定律成立。



**PART 2: Entropy**

**Introduction**

信息熵是信息论的基本概念。描述信息源各可能事件发生的不确定性。20世纪40年代，香农（C.E.Shannon）借鉴了热力学的概念，把信息中排除了冗余后的平均信息量称为“信息熵”，并给出了计算信息熵的数学表达式，即。

在信源中，考虑的不是某一单个符号发生的不确定性，而是要考虑这个信源所有可能发生情况的平均不确定性。若信源符号有n种取值：U1…Ui…Un，对应概率为：P1…Pi…Pn，且各种符号的出现彼此独立。这时，信源的平均不确定性应当为单个符号不确定性-logPi的统计平均值（E），可称为信息熵，即。

**Methodology**

在这部分的验证过程中，对文本进行的预处理与part1相似，之后提取字，词，得到一元、二元、三元模型的词频表。计算信息熵的公式分别为：

(一元模型)

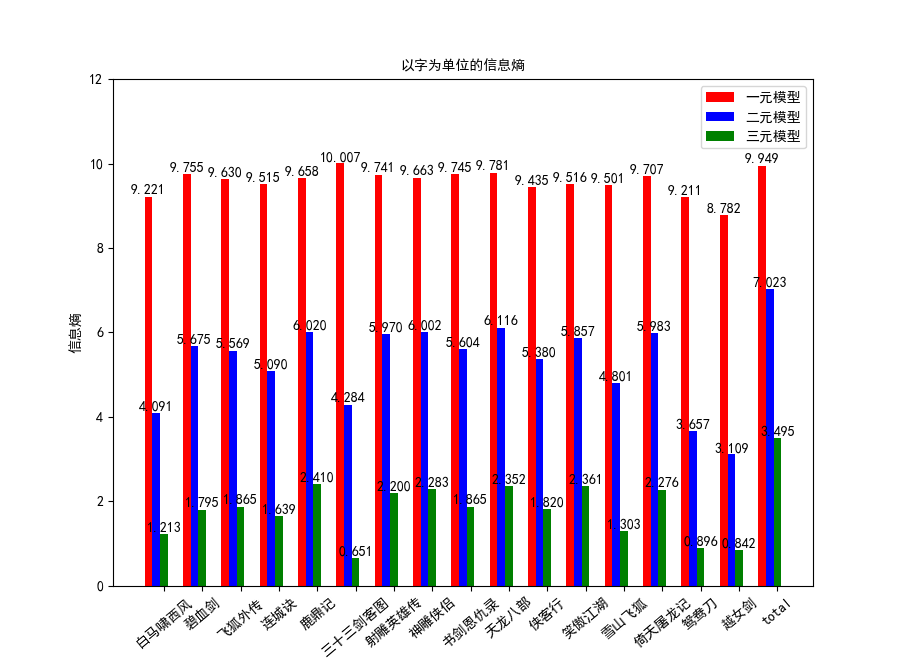
(二元模型)

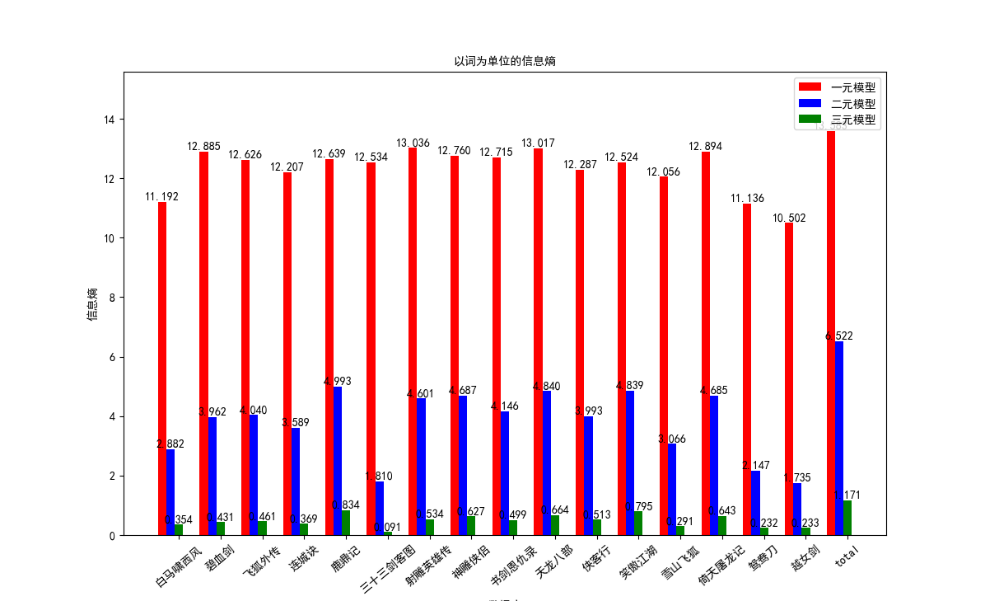
(三元模型)

最后使用matplotlib绘制语料库的信息熵柱状图。

**Experimental Studies**

最终绘制的图像如下图。对比一元模型、二元模型、三元模型的信息熵可以看到，N取值越大，即考虑前后文关系的长度越大，文本的信息熵越小，这是因为N越大，组成该词组的词越多，其冗余度也就越小，使用的特定场景越小，出现在文章中的不确定性越小。

****



**Conclusions**

在这个作业中，我成功地运用课堂上学到的理论知识，通过编写代码进行验证。通过实际的编码实践，我能够将课堂上学到的理论知识转化为可执行的代码，从而巩固了对知识的掌握。在编码和实验的过程中，我遇到了各种挑战，这促使我进行问题解决和批判性思考，进一步提升了我的编程技能。总的来说，这个作业为我提供了一个宝贵的机会，帮助我将理论与实践相结合，巩固了我的学习成果，同时提升了我的编程和算法思维能力。