NLP_REPORT_1

基于中文语料库的 Zip's Law 验证及信息熵计算

ZY2343226　朱子航

1010951286@qq.com

Abstract

美国哈佛大学的语言学教授乔治·金斯利·齐夫基于词频分布规律的数学模型，利用大量的统计数据，对自然语言词汇的分布规律作了细致的研究，提出齐夫定律（Zipf's law），核心观点为：如果将语言中的单词按照它们在文本中出现的频率进行排列，那么排在第 n 位的单词的出现频率将近似于排在第 1 位单词频率的 1/n 倍。

信息熵用于描述信息的不确定性或随机性,由美国数学家克劳德·香农（Claude Shannon）在 1948 年提出,通常用来衡量一组信息中所包含的平均信息量或信息的不确定性程度。

本文基于中文语料库进行词频统计验证齐夫定律，并按照一元、二元和三元语言模型分别计算字和词的信息熵。

Introduction

齐夫定律指出在任何一篇文章中，词的出现频率都服从如下规律：如果把一篇较长文章中每个词出现的频次统计出来，按照高频词在前、低频词在后的递减顺序排列，并用自然数给这些词编上等级序号，即频次最高的词等级为 1，频次次之的等级为 2，……，频次最小的词等级为 D。若用 f 表示频次，r 表示等级序号，则有 fr=C(C 为常数)[1]。

信息熵的计算涉及到各个可能事件发生的概率，如果一个事件是确定性的，即概率为 1 或 0，那么它的信息熵为 0，而如果一个事件的概率是随机的，那么它的信息熵会更高。信息熵的公式通常表示为：$H(\mathrm{X}) = E[-log\ pi\ ] = -\sum_{i=1}^{n} pi\ \log pi$ 式中对数一般取 2 为底，单位为比特，$H(\mathrm{X})$是随机变量 X 的信息熵，pi是事件 Xi

发生的概率，对所有可能的事件 Xi进行求和。

Methodology

本文基于中文语料库中（jyxstxtqj_downcc.com）的文本作为验证集，通过结巴分词的方式，去除符号（cn_punctuation.txt），统计对应的词频，获取词-频率字典，并绘制对应图像，即可验证齐夫定律。

在计算过程中发现提供的符号库中缺少“①”，遂添加保存为新的符号库。此问题同样存在于后文中的停词库（cn_stopwords.txt）。

对于中文平均信息熵的计算，如果统计量足够大，字、词、二元词组或三元词组出现的概率大致等于其出现的频率，则可以分为：

一元模型信息熵：

字和词的信息熵计算公式为：

$$H(X) = -\sum_{x \in X} P(x) log_2 P(x)$$

其中P(x)近似等于每个字或词在语料库中出现的频率。

二元模型信息熵：

$$H(X|Y) = -\sum_{x \in X, y \in Y} P(x,y) log_2 P(x|y)$$

其中联合概率P(x, y)近似等于每个二元词组在语料库中出现的频率，条件概率P(x|y)近似等于每个二元词组在语料库中出现的频数与以该二元词组的第一个词为词首的二元词组的频数的比值。

三元模型信息熵：

$$H(X|Y,Z) = -\sum_{x \in X, y \in Y, z \in Z} P(x,y,z) log_2 P(x|y,z)$$

其中联合概率近似等于每个三元词组在语料库中出现的频率，条件概率近似等于每个三元词组在语料库中出现的频数与以该三元词组的前两个词为词首的三元词组的频数的比值。

Experimental Studies

对中文语料库中所有文本合并后进行词频统计，由于语料库中字词数量超过一万，考虑到其庞大性，此处仅展示一部分词频，如表 1 所示。根据词频表绘制对应图像如图 1 所示。

表 1 中文语料库部分词频

Number	Word(s)	Frequency	Number	Word(s)	Frequency
1	的	115629	6	你	56405
2	了	104502	7	我	56244
3	他	64390	8	在	43609
4	是	63747	9	也	32606
5	道	60551	10	这	30781

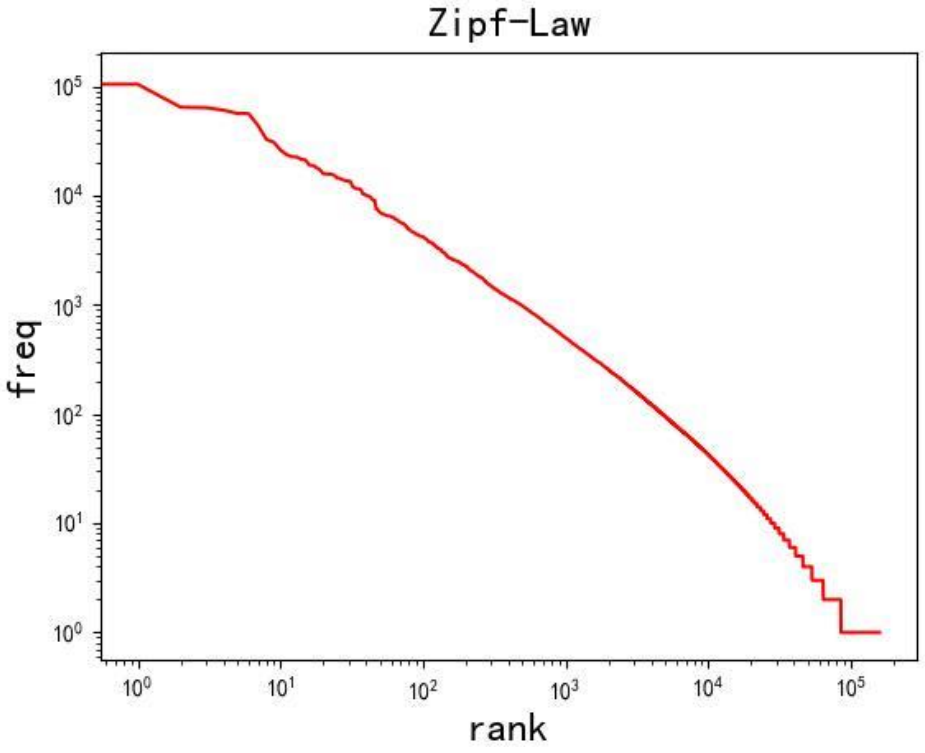

图 1 中文语料库词频统计图

由上图分析可得：中文语料库中字词的排序（Rank）及对应的频率（Frequency）的乘积接近不变，所画曲线近似为一直线，即可验证齐夫定律。

对中文语料库中的 16 部文本分别计算一元、二元和三元的字/词信息熵，得到结果如表 2 所示，绘制各文本信息熵柱状图如图 2、3 所示。

表 2 中文语料库信息熵

	名称	信息熵（比特/字词）					
		一元字	二元字	三元字	一元词	二元词	三元词
1	白马啸西风	9.2209	4.0907	1.2126	11.1920	2.8818	0.3543
2	碧血剑	9.7547	5.6755	1.7956	12.8849	3.9622	0.4307
3	飞狐外传	9.6301	5.5691	1.8650	12.6259	4.0404	0.4609
4	连城诀	9.5152	5.0902	1.6390	12.2066	3.5890	0.3685
5	鹿鼎记	9.6583	6.0200	2.4097	12.6390	4.9929	0.8344
6	三十三剑客图	10.0067	4.2845	0.6507	12.5335	1.8096	0.0912
7	射雕英雄传	9.7411	5.9700	2.1995	13.0359	4.6006	0.5341
8	神雕侠侣	9.6628	6.0025	2.2828	12.7604	4.6874	0.6265
9	书剑恩仇录	9.7448	5.6043	1.8654	12.7148	4.1461	0.4986
10	天龙八部	9.7807	6.1155	2.3522	13.0172	4.8399	0.6636
11	侠客行	9.4349	5.3800	1.8197	12.2875	3.9929	0.5127
12	笑傲江湖	9.5158	5.8565	2.3614	12.5238	4.8388	0.7955
13	雪山飞狐	9.5010	4.8015	1.3032	12.0564	3.0659	0.2906
14	倚天屠龙记	9.7066	5.9830	2.2760	12.8937	4.6851	0.6426
15	鸳鸯刀	9.2108	3.6565	0.8961	11.1362	2.1465	0.2323
16	越女剑	8.7824	3.1092	0.8421	10.5021	1.7351	0.2331

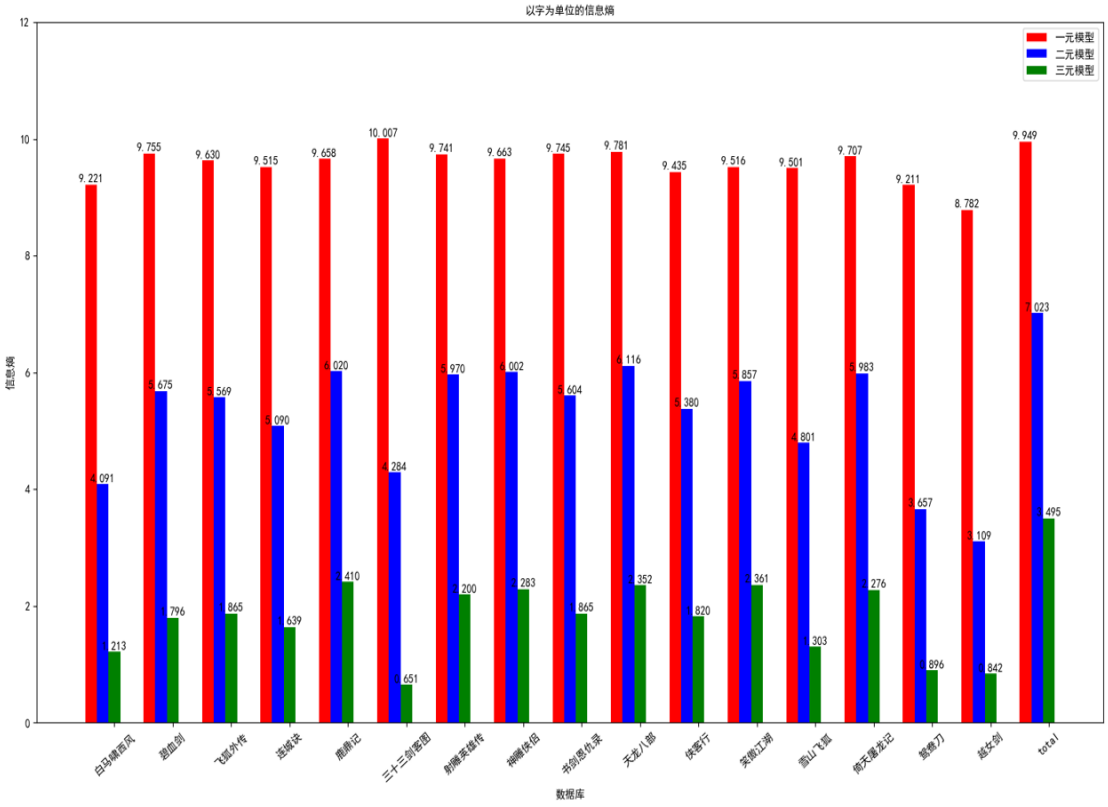

图 2 中文语料库信息熵（以字为单位）

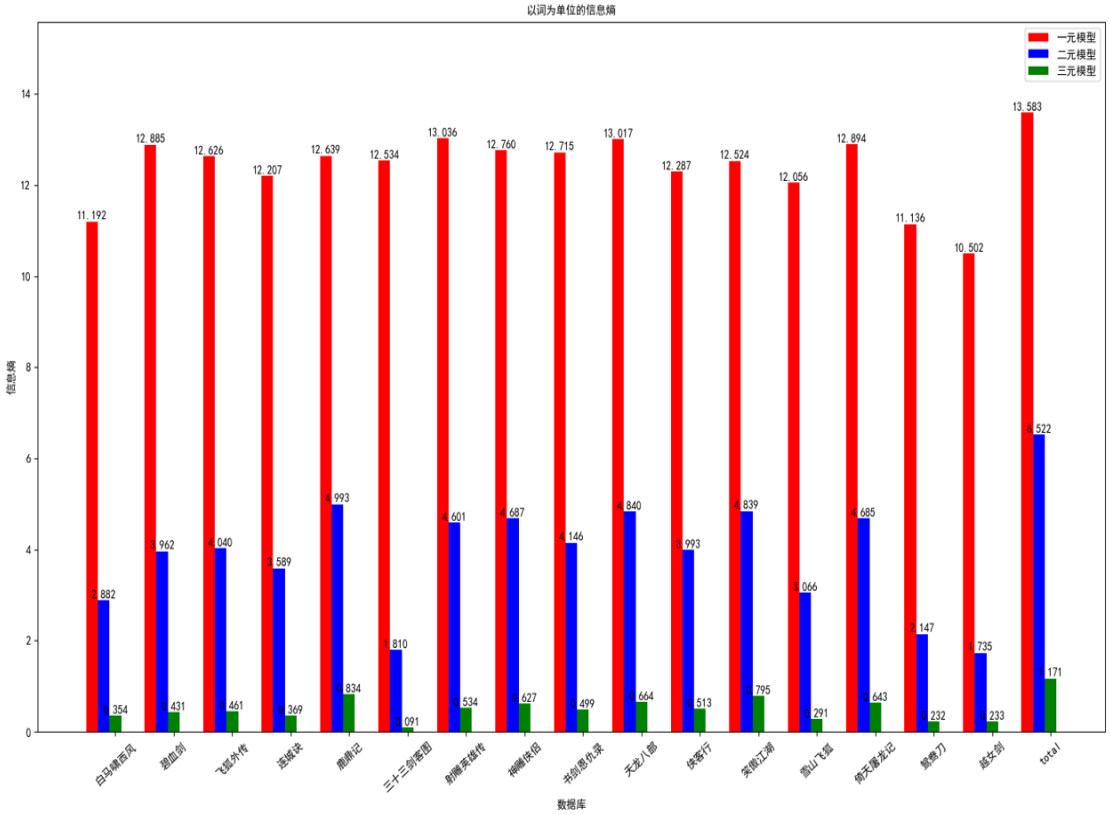

图 3 中文语料库信息熵（以词为单位）

由上图分析可得：中文语料库中 16 部文本字/词在相同语言模型下的信息熵基本为同一水平，且在每部文本中字/词的信息熵变化趋势相同；在不同的语言模型下，16 部文本字/词的信息熵呈现：一元模型 > 二元模型 > 三元模型的趋势，这是由于分元数的增大导致文本中分得的固定词数量增加，使文本分得更有序，其含义更确定。

References

[1] 方世敏,赵金金.基于齐夫定律的红色旅游景区旅游流扩散研究——以延安为例[J].延安大学学报（社会科学版）,2010,32(1):67-74

[2]https://www.gelbukh.com/CV/Publications/2001/CICLing-2001-Zipf.htm