基于中文语料库的齐夫定律验证及中文平均信息熵计算

ZY2303118　Yisong Wang
2740416657@qq.com

Abstract

Zipf's law（中文名**齐夫定律**）是由哈佛大学的语言学家乔治·金斯利·齐夫于 1949 年发表的实验定律[1]。这是自然语言处理领域的一个十分有趣的定律，但严格来说称之为规律更为贴切，因为这是一个经验性的结果，是通过统计数据得出的近似规律[2]。即在一个自然语言的语料库中，一个词的出现频数和这个词在这个语料中的排名成反比。应用场景一般为：可以假设一组数符合齐夫分布且第 N 个排名频率未知的情况下估计第 N 个排名频率。

信息熵在信息论中式接受每条信息中包含信息的平均量，又被称为信息熵、信源熵、平均自信息量[3]。

本文基于中文语料库中的《白马啸西风》及《天龙八部》文本作为验证集，再将所有文本集合成一个整体，通过结巴分词方式，统计对应的词频，获取词-频率字典，最终计算得到以字和词为单位下的平均信息熵，并绘制频率-排次表格及图像，经观察对比发现，齐夫定律成立。

Introduction

对于齐夫定律而言，频率最高的单词（排名第一）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，这个定律被作为任何与幂定律概率分布有关的事物的参考。类似 80/20 原则，即 20%的内容会占有 80%的访问量。

对于信息熵而言，依据 Boltzmann's H-theorem，香农把随机变量 X 的熵值 H 定义如下[4]，其值域为 $x_1,...,x_n$：

$$H(x)=E[I(X)]=E[-\ln(P(X))]$$

其中，P 为 X 的概率质量函数，E 为期望函数，而 $I(X)$ 是 X 的信息量，$I(X)$ 本身是个随机变数。

当取自有限的样本时，熵的公式可以表示为：

$$H(x)=\sum_i P(x_i)I(x_i)=-\sum_i P(x_i)\log_b P(x_i),$$

在这里，b 是对数所使用的底，通常为 2，此时熵的单位是 bit；当 $\mathrm{b=e}$，熵的单位是 mat；而当 $\mathrm{b=10}$，熵的单位是 Hart。

Methodology

由于上文提到对于齐夫定律而言，频率最高的单词（排名第一）出现的频率大约是出现频率第二位的单词的 2 倍，根据这种现象，不难设计出验证该定律的方法：即针对某个中文语料库中的文本，统计出分词出现的频率及对应的频率排次，绘制图像观察其是否符合反比关系。

本文首先采用结巴分词对文本进行分词处理，在进行频率排次时，为避免特殊字符对统计结果产生不良影响，仅对非特殊字符的字词进行排次和统计频次，最终根据统计得到的字典绘制相应的频率-排次表及图像。对于平均信息熵而言，其计算依赖于上述的统计量，但是由于熵的计算要考虑字和词分别出现的频率，因此计算前需将两者区分开统计。

Experimental Studies

由于文本性质缘故，选择中文语料库中的两个文本《白马啸西风》和《天龙八部》进行分析，根据上述方法，得到以下结果：

表 1:《白马啸西风》频率-排次表

Rank	Frequency	C=r×f	C×10
10	350	3500	35000
20	172	3440	34400
30	145	4350	43500
40	113	4520	45200
50	91	4550	45500
100	46	4600	46000

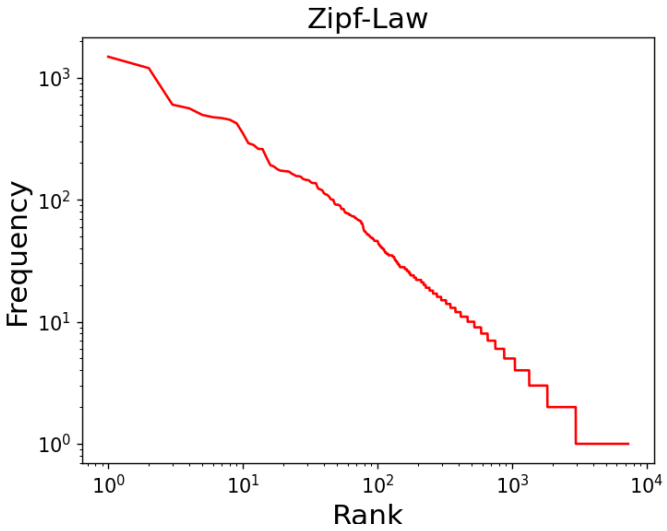

图 1:《白马啸西风》频率-排次图=

表 2:《天龙八部》频率-排次表

Rank	Frequency	C=r×f	C×10
10	5499	54990	549900
20	2980	59600	596000
30	2044	61320	613200
40	1735	69400	694000
50	1323	66150	661500
100	615	61500	615000

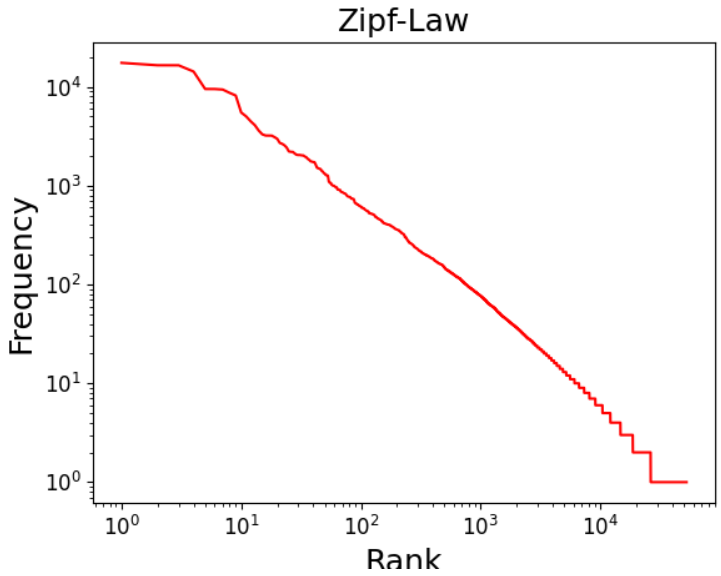

图 2:《天龙八部》频率-排次图

所有文本统计分析后的结果如下:

表 3: 整体频率-排次表

Rank	Frequency	C=r×f	C×10
10	43602	436020	4360200
20	18972	379440	3794400
30	14060	421800	4218000
40	11163	446520	4465200
50	7541	377050	3770500
100	4265	426500	4265000

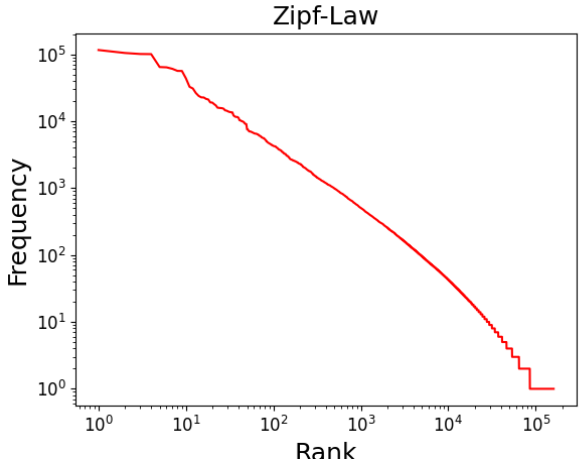

图 3：整体频率-排次图

观察分析可知，表格中分词排次 r 及对应的频率 f 之积大致处于一个较为稳定的状态，且通过频率-排次图可以发现，二者近似一条直线，因此可以验证满足齐夫定律，也即通过统计方法验证了齐夫定律的合理性。

对字频和词频分别统计，再依据信息熵公式计算得到《白马啸西风》的平均信息熵为（单位：bit）：

```
Character Entropy: 8.642208270763236
Word Entropy: 10.286307201354083
```

《天龙八部》的平均信息熵为（单位：bit）：

```
Character Entropy: 9.314677894095295
Word Entropy: 11.431480503447462
```

全部文本的平均信息熵为（单位：bit）：

```
Character Entropy: 9.464134347542979
Word Entropy: 11.911744789514241
```

References

[1] 梦家.齐夫定律(Zipf'law)理论及其应用场景[EB/OL].2022:[2024-4-7].https://dreamhomes.top/posts/202204221003/.

[2] historyasamirror.Zipf's law[EB/OL].2008:[2024-04-07].https://blog.csdn.net/historyasamirror/article/details/3125223.

[3] Kanglei Zhou.中文平均信息熵[EB/OL].2021:[2024-04-07].https://kangleizhou.github.io/nlp/2021/04/08/Chinese_entropy/.

[4] Brown P F, Della Pietra S A, Della Pietra V J, et al. An estimate of an upper bound for the entropy of English[J]. Computational Linguistics, 1992, 18(1): 31-40.