

NLP 作业 1：齐普夫定律验证与中文信息熵计算

顾韬

1312855584@qq.com

Abstract

首先介绍齐普夫定律来源，并根据文献[1]介绍了 N 元语言统计模型估计中文信息熵的方法，然后通过实验验证齐普夫定律与计算语料库中字和词为单位的中文信息熵。

Introduction

Zipf's Law 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。其表述为：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。所以，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。

信息熵：信息量是对信息的度量，事件发生概率越小信息量越大（且不为负）。信息量度量的是一个具体事件发生了所带来的信息，而熵则是在结果出来之前对可能产生的信息量的期望——考虑该随机变量的所有可能取值，即所有可能发生事件所带来的信息量的期望。

Methodology

M1: 信息熵

通常，一个信源发送出什么符号是不确定的，衡量它可以根据其出现的概率来度量。概率大，出现机会多，不确定性小；反之不确定性就大。

不确定性函数 f 是概率 P 的减函数；两个独立符号所产生的不确定性应等于各自不确定性之和，即 $f(P_1, P_2) = f(P_1) + f(P_2)$ ，这称为可加性。同时满足这两个条件的函数 f 是对数函数：

$$f(p) = \log \frac{1}{p} = -\log p$$

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： $U_1 \dots U_i \dots U_n$ ，对应概率为： $P_1 \dots P_i \dots P_n$ ，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 $-\log P_i$ 的统计平均值（ E ），可称为信息熵，即：

$$H(X) = E[-\log p_{x_i}] = - \sum_{i=1}^n p_{x_i} \log p_{x_i}$$

M2: 信息熵计算公式

在样本足够大的情况下，字与词出现的概率约为其出现的频率，因此，字与词的信息熵计算

公式为:

$$H(X) = - \sum_{x \in X} p(x) \log p(x)$$

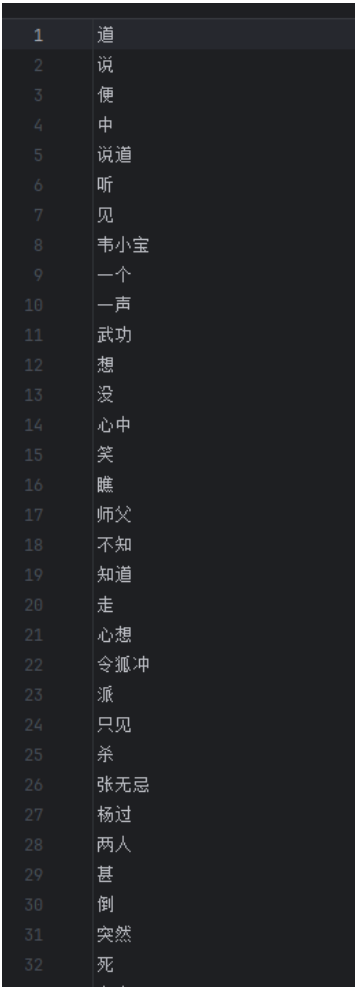
其中 $p(x)$ 近似每个字/词出现频率。

Experimental Studies

实验 1:验证 Zipf's Law

在验证之前，需要对原始数据进行处理，包括去除停用词，英文词等。停用词通过读取停用词 txt 文档获取，英文词、字符通过提取中文词去除。

使用 jieba 分词对语料库分词，就算各词的频率，排序后绘制图像。分词保存在 words.txt 中。



1	道
2	说
3	便
4	中
5	说道
6	听
7	见
8	韦小宝
9	一个
10	一声
11	武功
12	想
13	没
14	心中
15	笑
16	瞧
17	师父
18	不知
19	知道
20	走
21	心想
22	令狐冲
23	派
24	只见
25	杀
26	张无忌
27	杨过
28	两人
29	甚
30	倒
31	突然
32	死

Figure 1: 出现频率较高的词

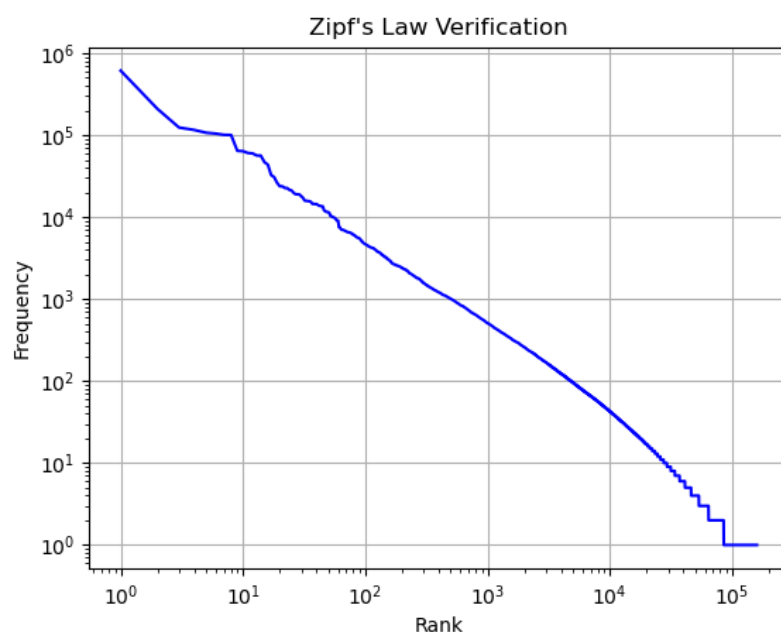


Figure 2: Zipf's Law 验证图

实验 2:中文信息熵计算

首先处理语料库数据，同实验 1 一致，在完成处理后分别计算中文字，词的出现频率，并由此计算对应的信息熵。本次实验分别计算一元模型、二元模型、三元模型的字和信息熵，实验结果如下：

```
一元词的信息熵: 13.639649930183577
二元词的信息熵: 6.505614703061618
三元词的信息熵: 1.150652701127206
一元词的信息熵: 9.84372831784633
二元词的信息熵: 6.781249477999459
三元词的信息熵: 3.6791929243341737
```

Figure 1: 各模型信息熵

Conclusions

在本次实验中，成功验证了 Zipf's law，并且使用了不同模型计算了中文语料库的信息熵，综合来看，一元模型的信息熵更高。

References

[1] Brown P F , Pietra V J D , Mercer R L ,et al.An estimate of an upper bound for the entropy of English[J].Computational Linguistics, 1992.DOI:10.5555/146680.146685.