作业 1 报告：Zipf’s Law 在中文语料库上的验证与平均信息熵计算

吴頔
wd_0509@163.com

Abstract

本文利用金庸创作的 16 部武侠小说作为中文语料库，验证 Zipf’s Law 在中文文本中的适用性。分别以字和词为单位，获取对应的频率，并按照排名绘制图像，最后计算其平均信息熵。观察对比发现，在误差允许的范围内，齐夫定律成立。

Introduction

齐夫定律（Zipf’s Law）是由美国语言学家乔治·齐夫在 20 世纪提出的经验定律之一。齐夫定律描述了自然语言中词语使用频率的分布规律，即某一词语的频率与其在频率表中的排名成反比关系，如排名第二的词语出现频率大约是排名第一的词语频率的一半，词频和词序之间存在着负相关关系。

齐夫定律在自然语言处理、信息检索、文本挖掘等领域具有广泛的应用价值，例如在文本压缩、信息检索、语音识别等方面。

信息熵是信息论中的一个重要概念，用于衡量一组数据的不确定性或信息量。信息熵越高，数据的不确定性就越大，含有的信息量也就越多。通过计算信息熵，能够计算词语表意的精确程度，信息熵越小，表意越精确。因此信息熵常被用来衡量信息的不确定性程度，以及信息编码的效率。

Methodology

Part 1:齐夫定律的验证

语料库的选择：本文用金庸创作的 16 部武侠小说作为中文语料库，其中包含丰富的词汇和句式，适合用于自然语言处理任务的训练和研究。

数据预处理：利用字符的 Unicode 编码对文件中的每个字符进行检查，滤除中文字符以外的乱码、无效内容以及符号等，并采用了 jieba 分词库进行分词处理。jieba 分词库是一款用于中文文本处理的 Python 库，提供了高效而准确的中文分词功能，它支持三种分词模式：精确模式、全模式、搜索引擎模式。在本文中，jieba 采用了精确模式，试图将句子最精确地切开，适合文本分析。

绘图：对字频与排序以及词频与排序取对数画图，绘制了所有 txt 文件的统计结果。

Part 2：平均信息熵的计算

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，即 $f(P)=\log\frac{1}{p}=-\log p$。

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：$U_1,...,U_i,...,U_n$，对应概率为：$P_1,...,P_i,...,P_n$，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：

$H(U)=E[-\log p_i]=-\sum_{i=1}^{n}p_i\log p_i$，式中对数一般取 2 为底，单位为比特。

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。由此可得，字和词的信息熵计算公式为：

$H(X)=-\sum_{x\in X}P(x)\log P(x)$。

其中，$P(x)$ 可近似等于每个字或词在语料库中出现的频率。

三元模型的信息熵[1]计算公式为：

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

其中，联合概率 $P(x,y,z)$ 可近似等于每个三元词组在语料库中出现的频率，条件概率 $P(x\mid y,z)$ 可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。

本文在对数据预处理后，分别对字和词进行平均信息熵的计算。采用三元词组计算，考虑三个相邻词之间的关系，捕捉更多的上下文信息。

Experimental Studies

Part 1:齐夫定律的验证

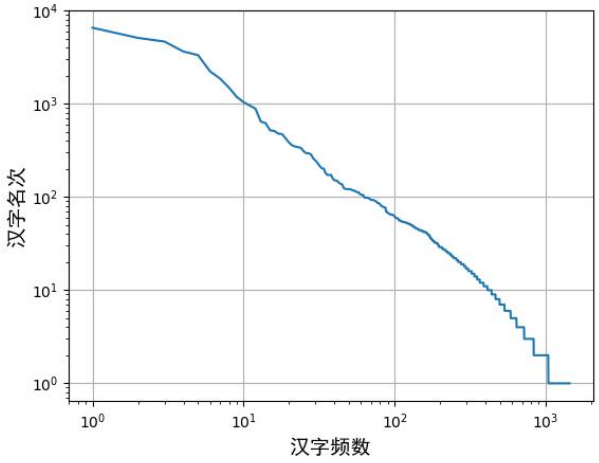

Figure 1：汉字频数与排序的关系

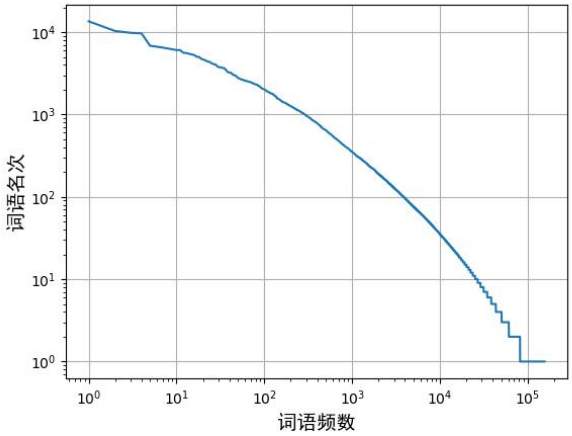

Figure 2：词语频数与排序的关系

Part 2:平均信息熵的计算

Table 1:各篇与总和的信息熵

篇名	以字为单位的信息熵	以词为单位的信息熵
白马啸西风	5.39574571659673	8.293883933192562
碧血剑	6.337301027249504	11.385374967043573
飞狐外传	5.744094095366788	11.095319950378036
连城诀	6.148944727646049	10.532660411038055
鹿鼎记	6.287714748341334	12.437028515446064
三十三剑客图	6.511861021550735	8.984700643547676
射雕英雄传	6.119319193111376	12.110432431274383
神雕侠侣	6.099892331614897	12.194352110707916
书剑恩仇录	6.42813782629828	11.6454356838577

天龙八部	6.381070996697945	12.619224392818701
侠客行	5.814943728255635	10.927059451263398
笑傲江湖	6.347289038563586	12.346738062991346
雪山飞狐	5.938712064573106	9.647766619250273
倚天屠龙记	6.09587839104598	12.216182279375221
鸳鸯刀	5.127186596517024	7.5156998382840365
越女剑	4.730914298173786	5.7548875021634665
总和	5.618176811858985	9.982749811331317

Conclusions

由 Figure 1 与 Figure 2 可知，在误差允许的范围内，汉字与词语的频数与排名成反比关系，且齐夫定律在词语上的表现更明显。

由 Table 1 可知，以字为单位的信息熵整体低于以词为单位的信息熵，猜测主要原因为词语之间的关联性可以提供更多的上下文信息，且词语通常具有一定的结构和组合规律，因此以词为单位的信息熵会更高。同时，对比 16 篇语料，可知以字为单位的信息熵波动较小，以词为单位的信息熵波动较大，猜测最主要原因为自得稳定性与词语的多样性，且词语有语境依赖性。

References

[1] 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.