作业 1 报告:中文语料库上的 Zipf’s Law 验证与平均信息熵计算

朱隽凡
1362992167@qq.com

Abstract

本报告基于作家金庸所著书籍研究其中中文语料是否满足 Zipf’s Law 并计算其中字词的平均信息熵。

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。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。信息熵的提出解决了对信息的量化度量问题。

Methodology

M1: 验证 Zipf’s Law

对给定的数据集，先提取出其中的有效部分，去除其中的广告信息和停词，使用 jieba 库对文本进行分词后，调用 Collect 库中的 Counter 类对词进行统计并排序，得到频数和次序的对应关系。

M2: 计算字词平均信息熵

类似于 M1，首先对数据集进行清洗，并统计文本中字和词的频数，最后套用信息熵计算公式得到计算结果。

Experimental Studies

对数据集中的所有文本，统计词频及其次序，取次序前 100 的词汇，画出二者关系图，如图 1 所示。可以看到词频与其次序基本成反比（即 log(rank)与 log(num)）负相关。二者的皮尔曼相关系数为-0.98451077。

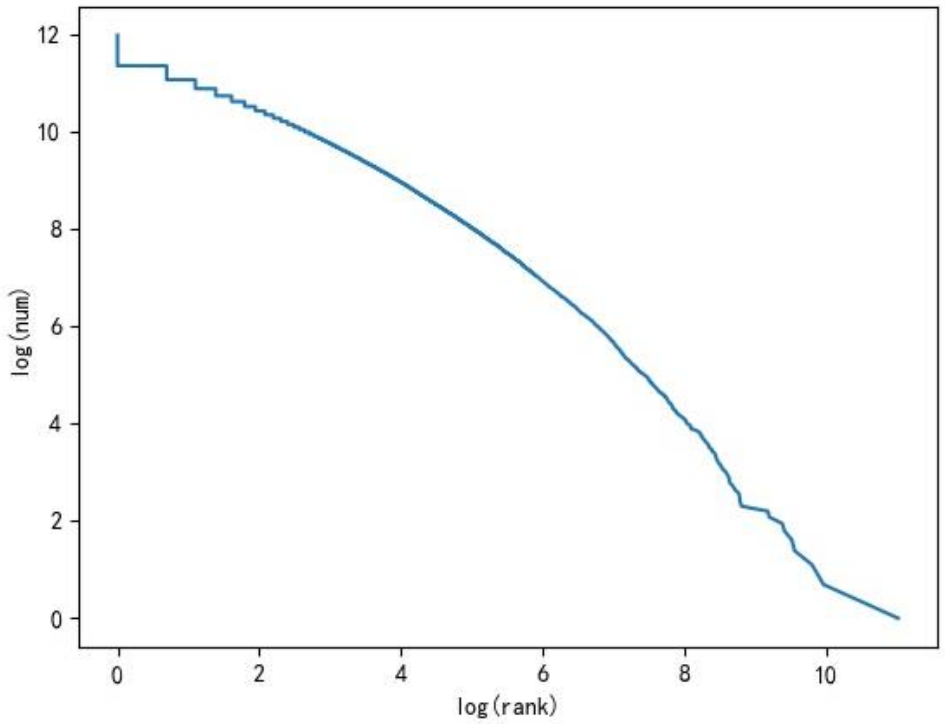

Figure 1: 词频与其次序关系图

对数据集中的所有数据，分别以字和词为单位计算平均信息熵，结果如表 1 所示。

表格 1: 书籍平均信息熵

书籍名称	字平均信息熵	词平均信息熵
白马啸西风	9.221785370877498	11.194709136375897
碧血剑	9.725274594647761	12.885402148212812
飞狐外传	9.569471361964947	12.625918522645847
连城诀	9.494382068272838	12.206626923070681
鹿鼎记	9.587004487555628	12.639245114463204
三十三剑客图	9.947520077851358	12.534805110351403
射雕英雄传	9.717020986702584	13.036156296891088
神雕侠侣	9.620546216268828	12.760658306492935
书剑恩仇录	9.703973225100919	12.715110102847824
天龙八部,	9.720671720044516	13.017355925646886
侠客行	9.418701103865997	12.288036059203385
笑傲江湖	9.472240100051533	12.523986792614927

雪山飞狐	9.453061495642647	12.057556101719474
倚天屠龙记	9.654363478097373	12.893909211087127
鸳鸯刀	9.223561391864322	11.139987017252887
越女剑	8.932779451067786	10.509352271756166

Conclusions

通过本次作业学习了分词的基本原理，深刻理解了齐夫定律和信息熵的含义。