作业报告一

Weijia Long
wjial@buaa.edu.cn

Abstract

第一部分：通过中文语料库来验证 Zipf's Law.

第二部分：阅读 Entropy Of English，计算中文(分别以词和字为单位) 的平均信息熵。

Part 1

Zipf's Law（齐普夫定律）是一种经验定律，描述了自然语言中词频分布的规律。该定律由美国语言学家乔治·齐普夫（George Zipf）在 20 世纪 30 年代提出。

齐普夫定律主要表述了这样一个现象：在大多数自然语言的语料库中，某一特定词的出现频率与其在词频排名上的倒数成反比关系。换句话说，排名第一的词出现的频率大约是排名第二的词的两倍，排名第三的词则是排名第二的词的三分之一，以此类推。

在中文语料库中，本文使用 python 中 jieba 库对 16 个文本进行分词，并计算词频，根据词频排序，提取词频和排名，并绘制频率与排名的对数图，若对数图为线性，则证明齐普夫定律。

Experimental 1

为验证齐普夫定律，本文将 16 个文本合并并绘制对数图，具体结果如图 1 所示，可以验证齐普夫定律

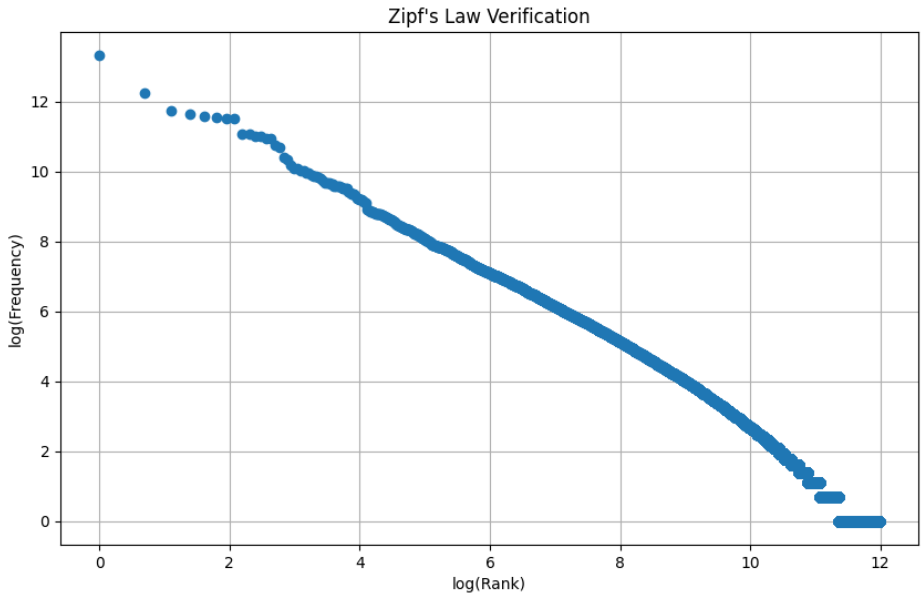

图 1. 频率与排名的对数图

Part 2

An Estimate of an Upper Bound for the Entropy of English 这篇论文提出了估算印刷英语字符熵上限的方法及其结果。作者使用了布朗语料库中的 5.96 百万字符，并利用从 5.83 亿个训练文本中构建的词三元语言模型来测量其交叉熵。通过这个方法，他们得出每个字符的上限估计值为 1.75 比特。文章还指出，自从香农在 1951 年发表论文以来，关于英语熵的估计已经有了很多，作者的方法与之前的工作不同，主要体现在使用了更大规模的英语文本样本，利用了语言模型来近似字符字符串的概率，并预测了所有可打印的 ASCII 字符。

根据信息熵公式，使用中文语料库进行计算中文的平均信息熵如下所示。

Experimental 2

通过计算 16 个文本库下中文平均信息熵如下所示:

语料库	词单位平均熵	字单位平均熵
三十三剑客图	4.0169	1.4516e-04
书剑恩仇录	6.5209	1.7429e-05

侠客行	5.5937	2.3277e-05
倚天屠龙记	5.5708	9.2013e-06
天龙八部	5.7437	7.2407e-06
射雕英雄传	4.6050	9.5999e-06
白马啸西风	4.4300	1.1135e-04
碧血剑	5.7974	1.8320e-05
神雕侠侣	4.6121	9.0809e-06
笑傲江湖	6.0149	8.9107e-06
越女剑	4.3667	4.61644-04
连城诀	5.5127	3.6845e-05
雪山飞狐	5.8225	6.4367e-05
飞狐外传	6.1616	2.0022e-05
鸳鸯刀	4.65036	2.1994e-04
鹿鼎记	5.2134	7.0355e-06
总语料库	5.1750	1.0354e-06