Report of Verification of Zipf’s Law

Weidong Xu
2697782204@qq.com

Abstract

This experiment aims to verify the applicability of Zipf's Law to Chinese text through the statistical data of Chinese Jin Yong’s novels. The latter of the experiment calculates the average information entropy in terms of words and characters, and finally draws a conclusion on the complexity of language text.

Introduction

This experiment focuses on the statistical characteristics of language and explores them via corpora

Zipf’s law, named after linguist George Zipf, is a statistical principle that describes the relationship between the frequency of words in a language and their rank or position in a frequency table. It states that the frequency of any word is inversely proportional to its rank. The most common word appears approximately twice as often as the second most common word, three times as often as the third most common word, etc.

The core idea of information theory is that the “informational value” of a communicated message depends on the degree to which the content of the message is surprising. If a highly likely event occurs, the message carries very little information. On the other hand, if a highly unlikely event occurs, the message is much more informative. For instance, the knowledge that some particular number will not be the winning number of a lottery provides very little information, because any particular chosen number will almost certainly not win. However, knowledge that a particular number will win a lottery has high informational value because it communicates the outcome of a very low probability event. Word-level information entropy reflects the complexity of lexical combinations, and word-level information entropy reflects the degree to which a single Chinese character carries information. The evaluation of word-level and word-level respectively in this experiment has, to some extent, significance for Chinese information processing.

Experimental Studies

Through the introduction of actual corpus data, the experiment verifies whether Zipf's Law is also applicable to Chinese, which is very different from English language characteristics, and visualizes the law by drawing the distribution diagram of frequency and rank. The experiments also included text preprocessing, using regular expressions to remove punctuation, jieba segmentation and stop word filtering. Figure 1 shows the relevant result between two factors: frequency and rank. A dense map is drawn based on the result, which is similar to an inverse proportional function. The words that appear in a descending order of frequency are as follows:

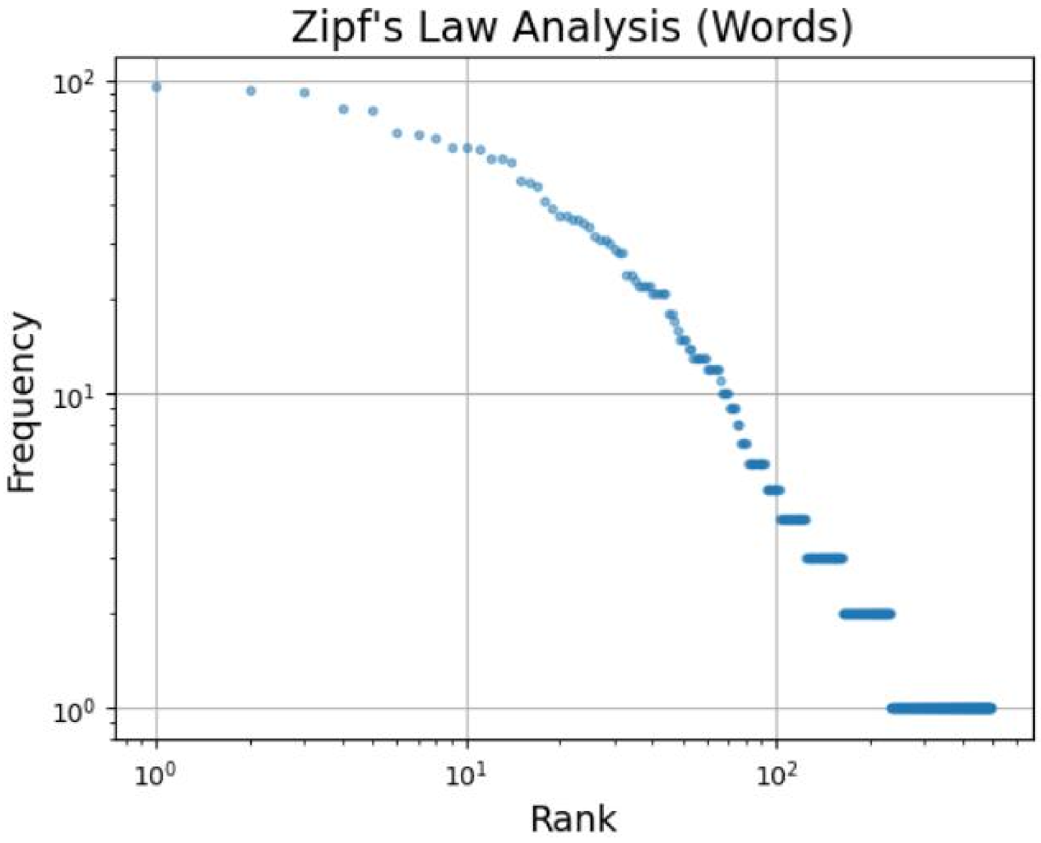

Figure 1 Frequency-Rank relation

In information theory, the entropy of a random variable is the average level of “information”, “surprise”, or “uncertainty” inherent to the variable’s possible outcomes. Given a discrete random variable X, which is distributed according to $p:\mathrm{X}\to[0,1]$， the entropy is

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

The calculation result is shown as follows:(Average Information Entropy)

corpus name	word level AIE(bits)	character level AIE(bits)
白马啸西风	0.0012	0.0038

碧血剑	0.0003	0.0023
飞狐外传	0.0004	0.0026
连城诀	0.0006	0.0029
鹿鼎记	0.0002	0.0021
三十三剑客图	0.0009	0.0033
射雕英雄传	0.0002	0.0020
神雕侠侣	0.0002	0.0022
书剑恩仇录	0.0003	0.0023
天龙八部	0.0002	0.0021
侠客行	0.0004	0.0027
笑傲江湖	0.0002	0.0023
雪山飞狐	0.0008	0.0032
倚天屠龙记	0.0002	0.0022
鸳鸯刀	0.0017	0.0043
越女剑	0.0028	0.0057

Conclusions

This experiment shows that Zipf's Law can be equivalently applicable to Chinese contexts. The essential two factors, frequency and rank are also in an inverse proportional distribution. The distribution pattern of word frequency has rich connotations, and the academic community believes that normal distribution is a typical distribution for describing natural sciences, while Zipf distribution will become a typical distribution for revealing social science laws.