Research About Zipf's Law and Information Entropy on Chinese Text

Zhibo Liang

zbliang_sasee@buaa.edu.cn

Abstract. This report performs a research about Zipf's law and information Entropy on Chinese Text. A total of sixteen books are ultilized to operate the calculating and plotting. As a result, the the Zipf's law is verified and the results of information entropy calculating are acquired.

1 Introduction

Text data is the foundation of natural language processing (NLP), so it is crucial to have a good grasp of it for excellent future learning. Language models play a key role in NLP as they are statistical or machine learning models designed to comprehend and generate human language. By learning language rules, grammar, semantics, and context, language models can be applied to various tasks such as text classification, named entity recognition, sentiment analysis, machine translation, and dialogue systems. Thus, a research about Chinese language models is conducted in this report to verify the Zipf's law and calculate the entropy.

2 Methodology

2.1 Zipf's Law

Zipf's law, also known as the Zipfian distribution or the principle of least effort, is an empirical observation[1]. It states that in a large corpus of natural language, the frequency of any word is inversely proportional to its rank in the frequency table. In simpler terms, it suggests that a few words occur very frequently, while the vast majority of words occur infrequently. The Zipf's law can be detailed as follows:

$$n_i \propto \frac{1}{i^\alpha} \tag{1}$$

Also can be expressed as:

$$\log n_i = -\alpha \log i + c \tag{2}$$

2.2 Information Entropy

Information entropy, often referred to simply as entropy, is a concept from information theory that measures the uncertainty or average amount of information contained in a random variable or a set of data. It quantifies the amount of surprise or unpredictability in the data[2]. For text data, if the sample size is large enough, the probability of occurrence for individual characters, words, bigrams, or trigrams is approximately equal to their respective frequencies. Based on this, the information entropy formula for characters and words can be derived as follows:

$$H(X) = -\sum_{x \in X} p(x) \log p(x) \tag{3}$$

3 Experiment Results

3.1 Zipf's Law

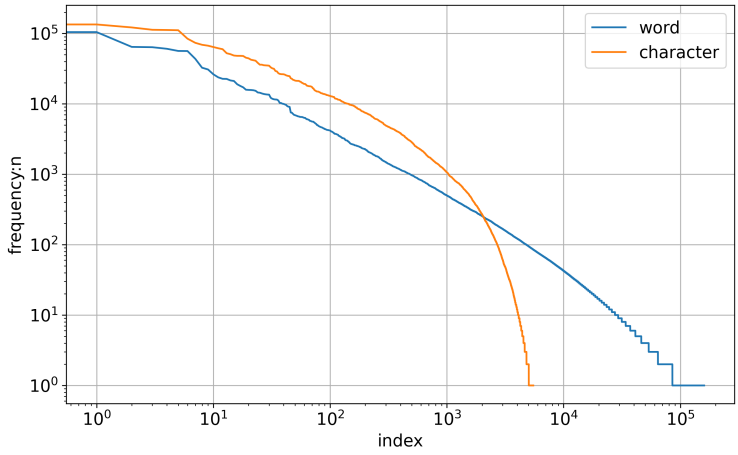

Fig. 1: The frequency-index curve of word.

The result of the verification of Zipf's Law on unigram word and character model is shown as Fig1. From this graph, it can be seen that word frequency rapidly decays in a distinct manner. After removing the first few exceptional words, all the remaining words roughly follow a straight line on a log-log coordinate graph. This indicates that the frequency of word follows Zipf's law. However, it is obvious that the unigram character model of chosen texts does not adhere to Zipf's law.

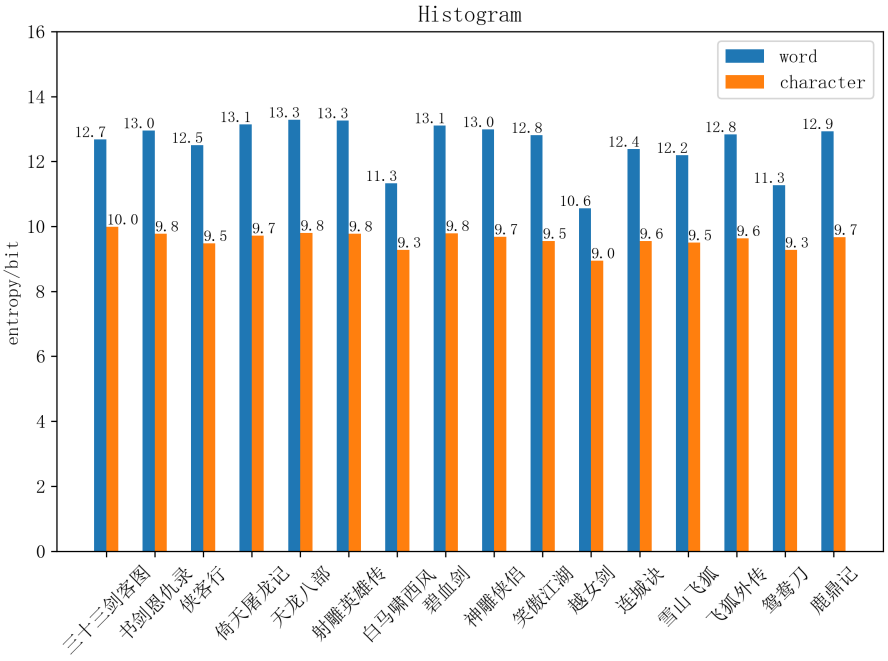

Fig. 2: Information entropy of unigram word and character.

3.2 Information Entropy

In this report, unigram character and word models are used for the calculation of information entropy. The result is shown in Fig.2, from which we can easily acquire that the information entropy of the unigram word model is lower than that of the character model.

4 Conclusion

The verification and calculation results indicate that the unigram word model of selected Chinese texts corresponds to the Zipf's law but unigram character model does not. In practical natural language models, predicting words is generally easier than predicting individual characters. However, the information entropy of the unigram character model is lower than that of the word model.

References

1. W. Li, "Zipf's law everywhere." *Glottometrics*, vol. 5, no. 2002, pp. 14–21, 2002.
2. D. J. MacKay, *Information theory, inference and learning algorithms*. Cambridge university press, 2003.