Report on Text Entropy Analysis

Menghuan Li
tshu603@gmail.com

Abstract

This report presents two experiments that analyze the "information entropy" of texts in both English and Chinese. Experiment 1 utilizes NLTK and the Gutenberg corpus to compute letter-level and word-level entropies of English texts. Experiment 2 processes Chinese texts from a local wiki_zh dataset using jieba segmentation and regular expressions, calculating both word-level and sentence-level entropies. The results demonstrate that the computed entropy values reflect the underlying structure and information distribution in the texts, providing useful indicators for natural language processing research.

Introduction

Information entropy is a key measure for the uncertainty of a random variable and has been widely applied in linguistics and natural language processing. By calculating the entropy of fundamental units (such as letters, words, and sentences) in a text, one can intuitively gauge the information density and complexity of the text. This report aims to compute the entropy for both English and Chinese texts through programming experiments, thereby exploring the characteristics of information distribution at different linguistic levels.

Methodology

The study consists of two parts, each using distinct text processing methods:

Experiment 1: Entropy Calculation for English Texts

This experiment uses the Gutenberg corpus provided by NLTK as the data source. The procedure begins by extracting raw text data using the *gutenberg.raw()* function. For each text, all alphabetic characters are filtered and converted to lowercase, and the frequency of each letter is counted. The letter entropy is then computed using the formula $H = -\sum p_i \log_2 p_i$. Similarly, the list of words is obtained using *gutenberg.words()*, where the words are converted to lowercase (keeping only alphabetic words) and their frequencies are calculated to determine the word entropy. Finally, the program prints both the letter entropy and word entropy for each text and computes the average entropies across all texts.

Experiment 2: Entropy Calculation for Chinese Texts

This experiment uses Chinese text files located in the local folder “wiki_zh” that follow the naming pattern “wiki_” as the data source. The procedure begins by reading the content of each file using proper encoding, after which jieba is applied for word segmentation to compute word frequencies and calculate the word-level entropy. In addition, regular expressions are used to split the text into sentences based on Chinese punctuation marks (。 ！ ？) to compute the sentence-level entropy. Exception handling is implemented to skip files that cannot be processed. Finally, for each file, the program outputs both the word-level and sentence-level entropies, and it computes the overall average entropies from all valid files.

Experimental Studies

We calculated the information entropy of Chinese Wikipedia in terms of characters and words, and the information entropy of Gutenberg Corpus in terms of letters and words, and the results are shown in Table 1.

Table 1: Average Information Entropy

Laguage	Unit	Entropy(Bits)
English	Letter	4.163
English	Word	9.174
Chinese	Character	12.775
Chinese	Word	10.239

We also drew the long tail graph between frequency and rank. The log-log graph of rank and frequency of characters/letters and words is shown in Figure 1 and Figure 2, where Figure 1 shows the long tail graph between Chinese characters and words, and Figure 2 shows the long tail graph between English letters and words

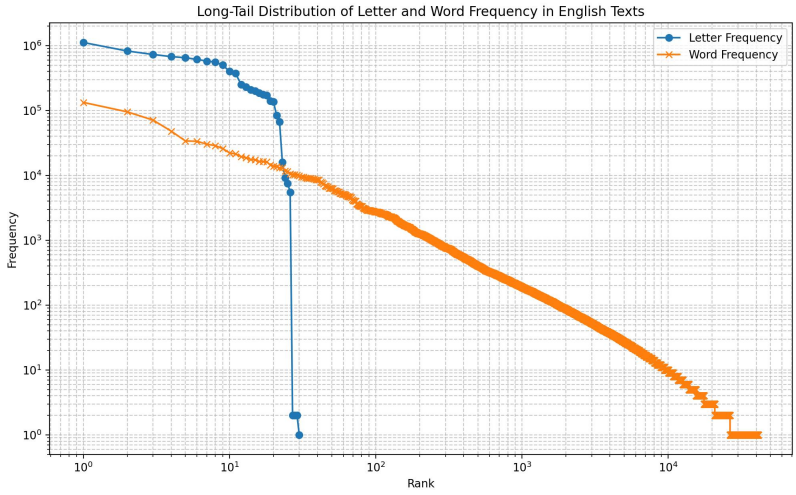

Figure 1： Log-log Graph of Gutenberg

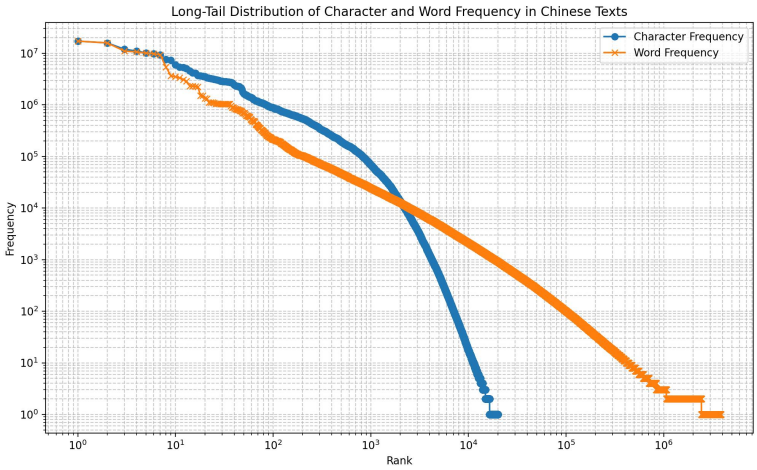

Figure 2： Log-log Graph of wiki_zh

Conclusions

In this study, we successfully computed the information entropy of texts in both English and Chinese, focusing on different linguistic levels such as letters, words, and sentences. The results indicate the following key insights:

Entropy Values Reflect Linguistic Complexity: The computed entropy values align with the inherent complexity of the respective languages. English letter-level entropy (4.163 bits) is significantly lower than Chinese character-level entropy (12.775 bits), highlighting the greater diversity and complexity of Chinese characters compared to English letters. However, English word-level entropy (9.174 bits) is close to Chinese word-level entropy (10.239 bits), suggesting that the structural complexity of words in both languages is relatively similar.

Zipf's Law Holds Across Languages: The log-log plots of frequency versus rank for both English and Chinese data sets exhibit a clear long-tail pattern, consistent with Zipf's law. This suggests that both English and Chinese follow a power-law distribution, where a small number of high-frequency units account for most of the text content, while a large number of low-frequency units form the long tail.

Differences Between Chinese and English: The higher entropy for Chinese characters and words indicates that Chinese texts have a higher information density and complexity. This can be attributed to the non-phonetic nature of Chinese characters and the larger vocabulary set compared to English.

Implications for Natural Language Processing: The entropy values and long-tail distribution patterns provide useful insights for language modeling and text analysis. The higher entropy in Chinese suggests greater challenges for language modeling, while the long-tail structure indicates the importance of handling low-frequency words effectively.

References

[1] Brown, P. F., Della Pietra, V. J., Mercer, R. L., Della Pietra, S. A., & Lai, J. C. (1992). An Estimate of an Upper Bound for the Entropy of English. Computational Linguistics, Vol. 18(1): pp. 31-40.
[2] Zipf, G. K. (1949). Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology. Addison-Wesley Press, Vol. 1: pp. 1-573.
[3] Powers, D. M. W. (1998). Applications and Explanations of Zipf's Law. Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning, Vol. 1: pp. 151-160.