Average Information Entropy of Text Corpora

Xinrui Yan
yanxinrui2021@ia.ac.cn

March 2025

Abstract

This paper presents an entropy analysis of text data from the Gutenberg corpus. Character-level and word-level entropy are computed to evaluate the distribution of information in different literary texts. The results provide insights into the complexity and variability of natural language text.

1 Introduction

Entropy is a fundamental concept in information theory that quantifies the unpredictability or randomness of a dataset. In text analysis, entropy is used to measure the diversity of characters and words in a given corpus. This study employs the Gutenberg corpus, a widely used collection of literary works, to compute and analyze entropy at different levels.

2 Methodology

The analysis involves the following steps:

- Preprocessing the text data by converting it to lowercase.
- Computing character-level entropy based on the frequency distribution of individual characters.
- Tokenizing the text and computing word-level entropy using word frequencies.

- Analyzing the results across different books in the Gutenberg corpus.

Entropy is computed using the formula:

$$H = -\sum_{i} p_i \log_2 p_i \tag{1}$$

where p_i represents the probability of occurrence of a character or word.

3 Experimental Results

The entropy values for different texts in the Gutenberg corpus were computed. Table 1 presents the character-level and word-level entropy values for selected works.

Book	**Character Entropy (bits)**	**Word Entropy (bits)**
austen-emma.txt	4.3683	8.6777
austen-persuasion.txt	4.3119	8.6756
austen-sense.txt	4.3400	8.6883
bible-kjv.txt	4.3920	8.5944
blake-poems.txt	4.2823	8.3667
bryant-stories.txt	4.3636	8.4228
burgess-busterbrown.txt	4.3689	8.0664
carroll-alice.txt	4.3690	8.3006
chesterton-ball.txt	4.3454	9.0015
chesterton-brown.txt	4.3152	9.0124
chesterton-thursday.txt	4.3453	8.8337
edgeworth-parents.txt	4.4230	8.7159
melville-moby_dick.txt	4.4084	9.4311
milton-paradise.txt	4.2928	9.2814
shakespeare-caesar.txt	4.3133	8.4906
shakespeare-hamlet.txt	4.3147	8.7401
shakespeare-macbeth.txt	4.3482	8.7388
whitman-leaves.txt	4.3302	9.1423
Average	**4.3462**	**8.7322**

Table 1: Entropy results for selected books

The results indicate that entropy values vary across different texts, reflecting differences in lexical richness and writing style. For example, works

such as *melville-moby_dick.txt* and *milton-paradise.txt* exhibit higher word-level entropy, suggesting greater lexical diversity. Conversely, texts such as *burgess-busterbrown.txt* have relatively lower entropy, implying simpler vocabulary and structure. The character-level entropy remains relatively stable across all books, suggesting consistent character distributions within English texts.

4 Conclusion

This study demonstrated an entropy-based analysis of text data using the Gutenberg corpus. Character-level and word-level entropy provide useful metrics for evaluating text complexity. Future work may involve comparing entropy values across different languages and genres.

5 References

References

[1] Brown, Peter F., et al. "An estimate of an upper bound for the entropy of English." Computational Linguistics 18.1 (1992): 31-40.

[2] Lieberman, Erez, et al. "Quantifying the evolutionary dynamics of language." Nature 449.7163 (2007): 713-716.