Report of the Entropy and Perplexity of Chinese and English Using the N-gram Model

Shihui Yang
yangshihui@buaa.edu.cn

Abstract

This is a Report of the Entropy and Perplexity of Chinese and English Using the N-gram Model. It employs unigram, bigram, and trigram models to calculate the average information entropy and perplexity at both the character level (letters or characters) and the lexical level (words or terms) for Chinese and English texts. This approach can reveal the statistical regularities of language, enhance the understanding of the intrinsic nature of linguistic statistics, and may provide data-driven decision-making support for model selection, optimization, and deployment in practical tasks.

Introduction

Information entropy and perplexity are crucial metrics in the field of natural language processing (NLP) for quantifying the statistical properties of language. Information entropy reflects the uncertainty or information content of linguistic symbols (such as words, characters, or letters), while perplexity is used to measure the predictive capability of a language model on a given corpus. These metrics not only reveal the statistical regularities of language but also provide a theoretical foundation for the evaluation and optimization of language models. In recent years, with the rapid development of deep learning techniques, the N-gram model, as a classical approach to language modeling, continues to play a significant role in linguistic statistical analysis and preliminary model evaluation due to its simplicity and efficiency.

This report aims to calculate the information entropy and perplexity at the word and character levels for a Chinese corpus (wiki_zh_2019), as well as at the word and letter levels for an English corpus (Gutenberg Corpus), using unigram, bigram, and trigram models. By doing so, it seeks to uncover the differences in statistical properties between Chinese and English at both the character and lexical levels. Through comparing the information entropy and perplexity across different linguistic units (such as words, characters, and letters), we can gain a deeper understanding of the statistical nature of language and provide data-driven decision-making support for subsequent model selection, optimization, and deployment.

Methodology

N-gram models are a fundamental approach in natural language processing (NLP) for modeling sequences of linguistic units, such as words or characters. These models are based on the Markov assumption, which posits that the probability of a unit in a sequence depends only on a fixed number of preceding units. This assumption allows for the simplification of complex joint probability distributions into products of conditional probabilities, making N-gram models computationally efficient and widely applicable in tasks such as language modeling, text generation, and machine translation.
The core idea of N-gram models is to decompose the joint probability of a sequence $w_1,w_2,\ldots,w_N$ into a product of conditional probabilities, where the context size is

determined by the value of N. The choice of N defines the specific type of N-gram model: unigram, bigram, or trigram, each with increasing levels of contextual information.

$$P(w_1, w_2, \ldots, w_N) = \prod_{i=1}^{N} P\big(w_i \big| w_{i-1}, w_{i-2}, \ldots, w_{i-(N-1)}\big)$$

The unigram model is the simplest form of N-gram, where each linguistic unit w_i is assumed to be independent of other units in the sequence. The probability of a unit w_i is calculated solely based on its frequency in the corpus, without considering any contextual information. The probability of a sequence w_1, w_2, …, w_N is given by:

$$P(w_1, w_2, \ldots, w_N) = \prod_{i=1}^{N} P(w_i)$$

The bigram model extends the unigram model by incorporating contextual information from the immediately preceding unit. It assumes that the probability of a unit w_i depends only on the previous unit w_{i-1}. The joint probability of a sequence is decomposed as:

$$P(w_1, w_2, \ldots, w_N) = \prod_{i=1}^{N} P(w_i | w_{i-1})$$

The trigram model further extends the context by considering the two preceding units. It assumes that the probability of a unit wi depends on the previous two units w_{i-2} and w_{i-1}. The joint probability of a sequence is given by:

$$P(w_1, w_2, \ldots, w_N) = \prod_{i=1}^{N} P(w_i | w_{i-2}, w_{i-1})$$

In the field of Natural Language Processing (NLP), entropy is a pivotal concept. It not only aids in quantifying the uncertainty or randomness of information but also profoundly influences the efficiency of language encoding, storage, transmission, and processing. By analyzing the entropy of language, we can gain a deeper understanding of the complexity of natural language and explore methods to enhance processing efficiency. In NLP, entropy can be utilized to measure the uncertainty of textual information. Its mathematical formulation can be expressed as:

$$H(X) = -\sum_{x \in X} P(x) \log P(x)$$

In the realm of Natural Language Processing, perplexity serves as a metric to evaluate the efficacy of probabilistic language models. A probabilistic language model can be conceptualized as a probability distribution over entire sentences or textual segments. It primarily estimates the likelihood of a sentence's occurrence based on each constituent word, normalized by the sentence's length. The formula is presented as follows:

$$Perplexity(W) = P(W)^{-\frac{1}{N}} = \left(\prod_{i=1}^{N} P(w_i | w_1, w_2, \ldots, w_{i-1})\right)^{-\frac{1}{N}}$$

This report will calculate and analyze the character-level and word-level information entropy and perplexity of the Chinese corpus (wiki_zh_2019) and the English corpus (Gutenberg Corpus) based on Unigram, Bigram, and Trigram models. To achieve this goal, we will sequentially preprocess the corpus data, perform frequency statistics, compute information entropy and perplexity, and conduct data visualization.

Step1: Data Preprocessing

In this phase, the process entails the tokenization and textual purification of both corpora, which involves the elimination of extraneous data such as punctuation, numerals, and stop words.

```
import re
def clean_and_join_words(word_list):
    pattern = re.compile(r'[^\w\s]')
    cleaned_words = [word for word in word_list if not
pattern.match(word)]
    result_string = ' '.join(cleaned_words)
    result_list = [result_string]
    return result_list
```

Additionally, for the English textual data, we have implemented stemming procedures to further refine the dataset. The corresponding code example is provided below.

```
from nltk.stem import PorterStemmer
```

```
def stem_tokens(token_list):
    stemmer = PorterStemmer()
    stemmed_tokens = [stemmer.stem(token) for token in token_list]
    return stemmed_tokens
```

Step2: Frequency Calculation

In this phase, we quantified the corpus by enumerating the total word count, calculating the average word length, and determining the frequency of the top 16 high-frequency words for each corpus. Subsequently, these results were subjected to visualization for analytical scrutiny. The relevant code example is provided as follows.

```
def calculate_average_word_length(text):
    words = text[0].split()
    if not words:
        return 0
    total_length = sum(len(word) for word in words)

    average_length = total_length / len(words)
    return average_length,len(words)

import matplotlib.pyplot as plt
import seaborn as sns
from collections import Counter
import numpy as np
plt.rcParams['font.sans-serif'] = ['SimHei']
plt.rcParams['axes.unicode_minus'] = False

def get_top_16_words(nested_list):
    words = [word for sublist in nested_list for word in sublist]
    word_counts = Counter(words)
    top_16_words = word_counts.most_common(16)
    top_words_list = [word for word, count in top_16_words]
    return top_words_list, top_16_words
```

Step3: Entropy and Perplexity Calculation

In this phase, the computations were performed utilizing the formula $H(X) = -\sum_{x \in X} P(x) \log P(x)$ to Entropy and the formula $Perplexity(W) = P(W)^{-\frac{1}{N}} = \left(\prod_{i=1}^{N} P(w_i|w_1, w_2, \dots, w_{i-1})\right)^{-\frac{1}{N}}$ (to prevent issues related to excessively large orders of magnitude or data overflow, it is common practice to first apply a logarithmic transformation followed by an exponential operation.) to compute the perplexity. The average information entropy was calculated according to the following formula. In our calculations of information entropy and perplexity, we also implemented smoothing techniques to mitigate the problem of data sparsity and enhance the accuracy of probability estimation. The corresponding code example is provided below (utilizing the trigram model to compute the information entropy and perplexity at the word level in English).

```
import math
from collections import defaultdict, Counter
def calculate_entropy(text):
    words = text[0].split()
    bigram_counts = defaultdict(Counter)
    trigram_counts = Counter()
    for i in range(len(words) - 2):
        bigram = (words[i], words[i+1])
        trigram = (*bigram, words[i+2])
        bigram_counts[bigram][words[i+2]] += 1
        trigram_counts[trigram] += 1
    total_trigrams = sum(trigram_counts.values())
    entropy = 0
```

```
    for trigram, count in trigram_counts.items():
        bigram = trigram[:2]
        next_word = trigram[2]
        joint_prob = count / total_trigrams
        if bigram in bigram_counts and next_word in
bigram_counts[bigram]:
            cond_prob = bigram_counts[bigram][next_word] /
sum(bigram_counts[bigram].values())
        else:
            cond_prob = 1e-10
        entropy -= joint_prob * math.log2(cond_prob)
    return entropy

import math
from collections import defaultdict, Counter
def calculate_perplexity(text):
    words = text[0].split()
    bigram_counts = defaultdict(Counter)
    trigram_counts = defaultdict(Counter)
    for i in range(len(words) - 2):
        bigram = (words[i], words[i+1])
        trigram = (*bigram, words[i+2])
        bigram_counts[bigram][words[i+2]] += 1
    log_perplexity = 0
    N = len(words) - 2
    for i in range(len(words) - 2):
        bigram = (words[i], words[i+1])
        next_word = words[i+2]
        if bigram in bigram_counts and next_word in
bigram_counts[bigram]:
            prob = bigram_counts[bigram][next_word] /
sum(bigram_counts[bigram].values())
        else:
            prob = 1e-10
        log_perplexity += math.log(prob)
    perplexity = math.exp(-log_perplexity / N)
    return perplexity
```

Experimental Studies

In this section, we will individually present the fundamental corpus information, frequency statistics, as well as the information entropy and perplexity for both corpora. Due to the limitations of the computer configuration, the information entropy and perplexity were calculated only for a subset of 400 texts from the Chinese corpus. The overall statistical results for the Chinese corpus comprising 400 texts are as follows: the total number of segmented words is 45,463,758, the total number of characters is 95,181,312, and the average word length is 2.09. For the English corpus, the total number of words is 1,023,063, and the average word length is 4.7825.

Table 1: results of entropy and perplexity

	English (letter)	English (word)	Chinese (character)	Chinese (word)
Unigram entropy	3.0213	11.2329	9.9714	14.3205
Bigram entropy	2.6625	4.9053	7.3379	7.4752
Trigram entropy	3.3676	1.4173	4.3249	1.3400
Unigram perplexity	20.5185	2406.8765	1003.9292	20459.9371
Bigram perplexity	14.3324	135.0132	161.7768	177.9322

Trigram perplexity	10.4508	2.6709	20.0415	2.5315

Gutenberg Corpus

The following presents a comprehensive statistical description of the Gutenberg Corpus, including the information about the average word length, total words, the entropy and perplexity based on unigram, bigram and trigram.

text serial	average word length	total words	unigram entropy	unigram perplexity	bigram entropy	bigram perplexity	trigram entropy	trigram perplexity
1.0	5.1469	73388.0	10.0181	1036.9067	3.6771	39.5315	0.7627	1.6966
2.0	5.2196	38341.0	10.1203	1113.049	3.1743	23.9095	0.4914	1.4058
3.0	5.3136	53951.0	10.1381	1126.8345	3.4631	31.9162	0.5474	1.4615
4.0	4.4277	436965.0	9.8848	945.3997	4.4841	88.5993	1.9143	3.7692
5.0	4.6311	3801.0	9.3377	647.0321	1.6447	5.1797	0.1385	1.1008
6.0	4.7112	21784.0	9.975	1006.3964	2.7058	14.9657	0.4179	1.336
7.0	4.6837	7613.0	8.7556	432.2274	2.1476	8.5641	0.7557	1.6885
8.0	4.71	12242.0	9.3795	666.0739	2.5298	12.5506	0.48	1.3948
9.0	5.1136	39871.0	10.8126	1798.5509	2.8563	17.3968	0.3231	1.251
10.0	5.1234	35335.0	10.8601	1858.7176	2.728	15.3029	0.2805	1.2146
11.0	5.0881	28306.0	10.5892	1540.544	2.6852	14.6609	0.3001	1.2312
12.0	4.9527	78145.0	10.5177	1466.0649	3.5384	34.4113	0.5793	1.4942
13.0	5.071	110650.0	11.3861	2676.4039	3.4405	31.2024	0.3816	1.3028
14.0	5.0027	45554.0	10.9745	2012.1188	2.9687	19.4674	0.206	1.1534
15.0	4.6924	11120.0	9.7825	880.6981	2.3467	10.4508	0.2518	1.1907
16.0	4.7205	15876.0	10.2814	1244.513	2.3645	10.6389	0.2401	1.181
17.0	4.7261	10137.0	10.1709	1152.7667	2.0439	7.7207	0.1641	1.1204
18.0	5.0122	65351.0	11.2987	2519.1089	3.0747	21.6443	0.2374	1.1789

Figure 1： the comprehensive statistical description of the Gutenberg Corpus (word)

text serial	average word length	total words	unigram entropy	unigram perplexity	bigram entropy	bigram perplexity	trigram entropy	trigram perplexity
1.0	5.1469	73388.0	2.9483	19.073	2.594	13.3831	2.9895	8.8629
2.0	5.2196	38341.0	2.9307	18.7404	2.5867	13.2859	2.8546	9.1009
3.0	5.3136	53951.0	2.9237	18.6099	2.5685	13.0468	2.914	8.855
4.0	4.4277	436965.0	3.0851	21.8705	2.6476	14.1195	3.2039	9.3707
5.0	4.6311	3801.0	2.92	18.5413	2.5586	12.9173	1.8337	9.883
6.0	4.7112	21784.0	2.9395	18.9072	2.601	13.4766	2.8294	9.2572
7.0	4.6837	7613.0	2.9526	19.1559	2.5281	12.5291	1.9002	7.8776
8.0	4.71	12242.0	2.943	18.973	2.5939	13.3822	2.5412	8.9655
9.0	5.1136	39871.0	2.9463	19.0354	2.6238	13.7876	2.8616	10.0719
10.0	5.1234	35335.0	2.9352	18.8258	2.6263	13.822	3.1095	10.1303
11.0	5.0881	28306.0	2.935	18.8207	2.6146	13.662	2.8855	9.9459
12.0	4.9527	78145.0	2.9449	19.0093	2.6317	13.8972	3.0984	9.8639
13.0	5.071	110650.0	2.9495	19.0968	2.6305	13.8811	3.2332	10.3967
14.0	5.0027	45554.0	2.9224	18.5865	2.5956	13.4041	3.0141	10.0105
15.0	4.6924	11120.0	2.9077	18.3143	2.5865	13.2832	2.4465	9.5499
16.0	4.7205	15876.0	2.9157	18.4614	2.611	13.6125	2.8953	9.9199
17.0	4.7261	10137.0	2.9368	18.8553	2.6125	13.6324	2.6282	9.9553
18.0	5.0122	65351.0	2.9364	18.8483	2.63	13.8738	3.1742	10.5296

Figure 2： the comprehensive statistical description in Gutenberg Corpus (letter)

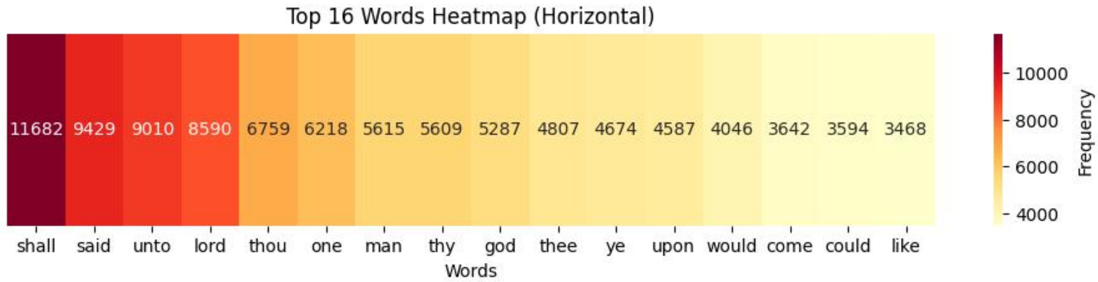

Figure 3： Top 16 frequent words in the Gutenberg Corpus

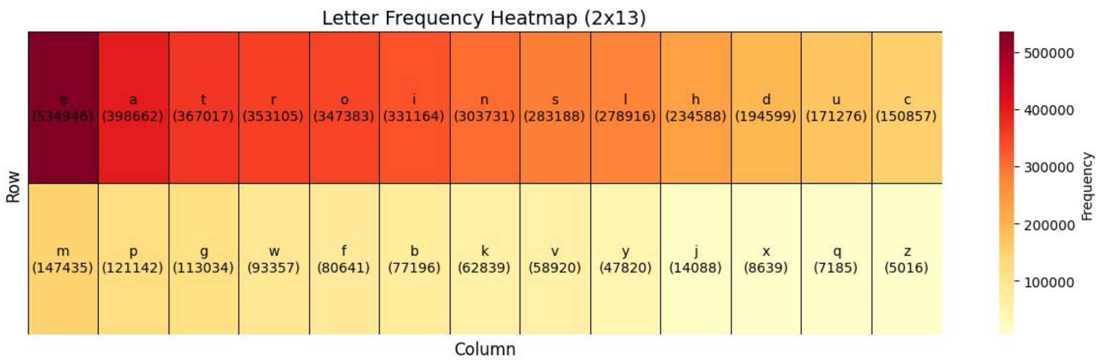

Figure 4： The frequency distribution of 26 letters in the Gutenberg Corpus

Chinese Wikipedia

This section provides a detailed statistical overview of the Chinese Wikipedia Corpus, encompassing metrics such as average word length, total word count, as well as entropy and perplexity calculated using unigram, bigram, and trigram models.

	text serial	word segments	total character	average word length	unigram entropy	unigram perplexity	bigram entropy	bigram perplexity	trigram entropy	trigram perplexity
0	1.0	112752.0	244383.0	2.1674	12.5542	6014.3366	3.8275	14.197	0.3402	1.2659
1	2.0	113767.0	254062.0	2.2332	12.6127	6263.2817	3.679	12.8085	0.3964	1.3162
2	3.0	113349.0	247289.0	2.1817	12.7061	6681.9698	3.5848	11.9984	0.3929	1.313
3	4.0	115079.0	244439.0	2.1241	12.9277	7791.5667	3.4816	11.1705	0.3237	1.2516
4	5.0	117997.0	247895.0	2.1009	13.1692	9211.3816	3.3455	10.1645	0.2683	1.2043
5	6.0	112820.0	240696.0	2.1335	12.7051	6677.7331	3.6633	12.6696	0.3407	1.2663
6	7.0	113394.0	240832.0	2.1239	12.8883	7581.462	3.5195	11.4673	0.3122	1.2416
7	8.0	112842.0	244292.0	2.1649	12.7307	6797.2066	3.6093	12.2038	0.3439	1.2692
8	9.0	113148.0	243180.0	2.1492	12.8338	7300.6145	3.5126	11.413	0.3423	1.2678
9	10.0	110823.0	242782.0	2.1907	12.6087	6246.0666	3.6362	12.4336	0.3771	1.2988
10	11.0	115157.0	244815.0	2.1259	13.0372	8405.7912	3.3847	10.4449	0.2931	1.2253
11	12.0	117544.0	248091.0	2.1106	12.9647	7993.8036	3.4706	11.0857	0.3293	1.2564
12	13.0	114948.0	247511.0	2.1532	12.6414	6389.3192	3.653	12.5797	0.3828	1.3039
13	14.0	111647.0	238899.0	2.1398	12.8164	7213.0866	3.5262	11.521	0.3359	1.2621
14	15.0	115530.0	246067.0	2.1299	12.9779	8067.504	3.4321	10.7939	0.3154	1.2443
15	16.0	115407.0	248198.0	2.1506	12.7222	6757.3434	3.6082	12.195	0.3781	1.2996
16	17.0	114297.0	244064.0	2.1353	12.9137	7716.2874	3.4199	10.7026	0.3484	1.2732
17	18.0	114655.0	248721.0	2.1693	12.7505	6891.1973	3.588	12.0252	0.3691	1.2916
18	19.0	115566.0	245212.0	2.1218	12.9919	8146.2984	3.4123	10.6461	0.3269	1.2543
19	20.0	109198.0	230437.0	2.1103	12.8273	7267.5692	3.4923	11.2534	0.3281	1.2554
20	21.0	116922.0	247763.0	2.119	12.9388	7851.5547	3.4532	10.9524	0.3435	1.2689
21	22.0	115405.0	244233.0	2.1163	13.0534	8501.0556	3.383	10.4323	0.2977	1.2292
22	23.0	112243.0	239669.0	2.1353	12.8317	7290.0249	3.5239	11.503	0.3249	1.2526
23	24.0	117001.0	247714.0	2.1172	13.0744	8625.5193	3.3774	10.392	0.3056	1.2359
24	25.0	115657.0	248997.0	2.1529	12.849	7377.7022	3.5223	11.4897	0.3422	1.2677
25	26.0	112407.0	244261.0	2.173	12.6984	6646.7505	3.55	11.7129	0.3986	1.3182
26	27.0	116995.0	249875.0	2.1358	12.7957	7110.1581	3.5087	11.3823	0.3788	1.3002
27	28.0	110527.0	233811.0	2.1154	13.0879	8706.5548	3.2514	9.5231	0.3187	1.2472
28	29.0	116391.0	244923.0	2.1043	13.2461	9715.6727	3.2121	9.2671	0.2843	1.2178
29	30.0	114243.0	243530.0	2.1317	12.8881	7580.6715	3.4889	11.2267	0.331	1.2579
30	31.0	113631.0	242770.0	2.1365	12.8682	7476.8965	3.4559	10.9734	0.3548	1.2788
31	32.0	112469.0	241406.0	2.1464	12.9034	7661.615	3.2618	9.5916	0.4056	1.3246
32	33.0	91005.0	189373.0	2.0809	12.8468	7366.8651	2.9027	7.4784	0.2845	1.218
33	34.0	108403.0	229702.0	2.119	13.1086	8832.4283	3.2119	9.2658	0.3178	1.2464
34	35.0	113402.0	241440.0	2.1291	13.0875	8704.2443	3.3007	9.8539	0.3183	1.2469
35	36.0	114274.0	241720.0	2.1153	12.9957	8167.5768	3.3728	10.359	0.3161	1.2449
36	37.0	115425.0	243216.0	2.1071	13.1048	8808.9835	3.3066	9.8942	0.3087	1.2386
37	38.0	111609.0	244465.0	2.1904	12.7976	7119.8269	3.44	10.8529	0.3714	1.2936
38	39.0	115002.0	241784.0	2.1024	13.2551	9776.2898	3.2215	9.3274	0.2637	1.2006
39	40.0	111064.0	236751.0	2.1317	12.8436	7350.2397	3.438	10.8382	0.3646	1.2875
40	41.0	114672.0	246489.0	2.1495	12.8969	7626.9225	3.4913	11.2454	0.3328	1.2594
41	42.0	110487.0	234717.0	2.1244	12.8791	7533.7314	3.4449	10.8895	0.3275	1.2548
42	43.0	117948.0	249686.0	2.1169	12.9518	7922.6515	3.3439	10.1533	0.4034	1.3226
43	44.0	113801.0	244278.0	2.1465	12.993	8152.3928	3.373	10.36	0.3363	1.2625
44	45.0	114913.0	240845.0	2.0959	13.1411	9033.4037	3.289	9.7743	0.2948	1.2267
45	46.0	116850.0	249083.0	2.1316	12.9734	8042.0906	3.3506	10.2006	0.3827	1.3038
46	47.0	112902.0	237193.0	2.1009	12.9697	8022.0149	3.3797	10.4088	0.3311	1.258
47	48.0	112386.0	238514.0	2.1223	13.0401	8422.9025	3.3286	10.0462	0.3155	1.2445
48	49.0	113783.0	239548.0	2.1053	13.1298	8963.2611	3.2754	9.6826	0.3123	1.2416
49	50.0	113966.0	241049.0	2.1151	13.0273	8348.6703	3.3417	10.1382	0.3147	1.2437
50	51.0	113649.0	238121.0	2.0952	13.0889	8712.6499	3.2713	9.655	0.3179	1.2465
51	52.0	117545.0	246424.0	2.0964	12.8667	7468.8375	3.4472	10.9073	0.3835	1.3045
52	53.0	113647.0	241560.0	2.1255	13.0102	8250.0103	3.3119	9.9308	0.3434	1.2687
53	54.0	116263.0	243967.0	2.0984	12.9975	8177.5702	3.347	10.175	0.3646	1.2875
54	55.0	111897.0	236799.0	2.1162	12.9002	7644.4648	3.4371	10.8307	0.3308	1.2577
55	56.0	115568.0	241811.0	2.0924	13.0313	8371.4045	3.3014	9.8586	0.3539	1.278
56	57.0	110749.0	235075.0	2.1226	12.7384	6833.3777	3.4538	10.9571	0.4002	1.3197
57	58.0	113703.0	240510.0	2.1152	13.0655	8572.2887	3.3076	9.9011	0.325	1.2527
58	59.0	113891.0	238419.0	2.0934	12.7258	6774.2626	3.541	11.6402	0.4081	1.3269
59	60.0	116675.0	247253.0	2.1192	12.9853	8108.9897	3.4037	10.5832	0.3376	1.2636
60	61.0	112856.0	238881.0	2.1167	13.0195	8303.2266	3.3241	10.0153	0.3416	1.2671
61	62.0	116643.0	245266.0	2.1027	13.0764	8637.2827	3.3252	10.0226	0.3143	1.2434
62	63.0	115620.0	243500.0	2.106	12.8713	7492.7703	3.4549	10.9652	0.3775	1.2991
63	64.0	115596.0	244706.0	2.1169	12.9254	7779.3509	3.3679	10.324	0.3975	1.3172
64	65.0	114287.0	241076.0	2.1094	13.015	8277.4196	3.2773	9.6951	0.3463	1.2713
65	66.0	113829.0	239842.0	2.107	12.8601	7435.1621	3.4739	11.1108	0.3517	1.2761
66	67.0	114015.0	241545.0	2.1185	13.0269	8346.3145	3.3093	9.9127	0.3423	1.2678
67	68.0	115118.0	244621.0	2.125	12.9517	7922.3157	3.3919	10.497	0.3506	1.2751
68	69.0	115134.0	246043.0	2.137	13.0346	8391.0559	3.3107	9.9224	0.3419	1.2674
69	70.0	114563.0	246101.0	2.1482	12.9358	7835.47	3.3466	10.1726	0.3811	1.3024
70	71.0	115146.0	243685.0	2.1163	12.98	8079.0557	3.3702	10.3404	0.3531	1.2773
71	72.0	114524.0	240827.0	2.1029	12.9642	7990.9591	3.3696	10.3358	0.354	1.2781
72	73.0	115426.0	242830.0	2.1038	13.0156	8281.2151	3.2215	9.3273	0.3612	1.2845
73	74.0	111811.0	236975.0	2.1194	13.0814	8667.5013	3.2587	9.5715	0.3171	1.2458
74	75.0	117564.0	247510.0	2.1053	12.8509	7387.4691	3.3901	10.4836	0.4258	1.3433
75	76.0	116496.0	243987.0	2.0944	13.1048	8809.5235	3.3007	9.8536	0.3149	1.244
76	77.0	116799.0	245096.0	2.0984	12.9208	7754.4019	3.4283	10.7649	0.363	1.2861
77	78.0	115331.0	242417.0	2.1019	13.1171	8884.4652	3.2817	9.7248	0.3207	1.249
78	79.0	114655.0	239555.0	2.0894	13.0332	8382.868	3.2912	9.7895	0.3558	1.2797
79	80.0	112941.0	240172.0	2.1265	12.9517	7922.2187	3.3365	10.1015	0.3669	1.2896
80	81.0	113630.0	237771.0	2.0925	12.8401	7332.7568	3.4652	11.0441	0.3708	1.2931
81	82.0	112533.0	237603.0	2.1114	12.9739	8045.3007	3.341	10.1331	0.3543	1.2784
82	83.0	116589.0	244027.0	2.0931	13.0234	8326.2415	3.336	10.0982	0.3518	1.2762
83	84.0	118103.0	246378.0	2.0861	13.0889	8712.4477	3.3134	9.9407	0.3354	1.2618
84	85.0	115653.0	244505.0	2.1141	13.0421	8434.5451	3.2775	9.6969	0.3599	1.2833
85	86.0	114716.0	242989.0	2.1182	13.0915	8728.4608	3.2403	9.4497	0.3599	1.2833
86	87.0	116840.0	244268.0	2.0906	13.1478	9075.4522	3.2325	9.3991	0.3461	1.2711
87	88.0	116607.0	245889.0	2.1087	13.168	9203.7495	3.1682	8.9895	0.3462	1.2712
88	89.0	112259.0	239290.0	2.1316	12.8857	7567.8118	3.449	10.9208	0.3465	1.2715
89	90.0	112886.0	238294.0	2.1109	13.0884	8709.4777	3.2978	9.8345	0.306	1.2363
90	91.0	112073.0	235309.0	2.0996	12.9427	7872.9135	3.369	10.3314	0.3458	1.2708
91	92.0	117198.0	245898.0	2.0981	13.0975	8764.8713	3.2715	9.6563	0.3534	1.2776
92	93.0	111691.0	235191.0	2.1057	13.0146	8275.4233	3.3096	9.9148	0.3344	1.2608
93	94.0	112214.0	236200.0	2.1049	13.0706	8602.6775	3.2853	9.7493	0.3323	1.2591
94	95.0	114563.0	241659.0	2.1094	13.1846	9309.9246	3.1878	9.1121	0.325	1.2527
95	96.0	117196.0	244018.0	2.0821	13.3645	10546.8825	3.0958	8.5492	0.2894	1.2221
96	97.0	114553.0	240602.0	2.1004	13.0518	8491.4598	3.3242	10.016	0.3314	1.2582
97	98.0	116409.0	245651.0	2.1102	13.092	8731.1279	3.3019	9.8618	0.3315	1.2583
98	99.0	114185.0	236025.0	2.067	13.0699	8598.9778	3.2824	9.7301	0.3341	1.2606
99	100.0	114340.0	239235.0	2.0923	13.2522	9757.1611	3.1193	8.6899	0.3184	1.2469

Figure 5： the comprehensive statistical description of 100 texts in the Chinese Wiki Corpus (word)

	text serial	total character	unigram entropy	unigram perplexity	bigram entropy	bigram perplexity	trigram entropy	trigram perplexity
0	1.0	205722.0	9.5031	725.6431	5.3785	41.6004	2.005	4.0138
1	2.0	213974.0	9.4344	691.8699	5.2829	38.9322	1.9732	3.9264
2	3.0	209287.0	9.5762	763.343	5.3328	40.3031	1.9073	3.7511
3	4.0	207651.0	9.7494	860.7381	5.4562	43.9003	1.7721	3.4156
4	5.0	210131.0	9.8437	918.8481	5.5382	46.4705	1.6849	3.2151
5	6.0	202773.0	9.6565	807.0536	5.3911	41.9649	1.8711	3.658
6	7.0	203462.0	9.7466	859.0367	5.4349	43.2582	1.7817	3.4384
7	8.0	206483.0	9.6428	799.4062	5.3719	41.4096	1.8424	3.5861
8	9.0	203944.0	9.7024	833.1224	5.3944	42.0606	1.7998	3.4818
9	10.0	204685.0	9.6296	792.1346	5.2688	38.5527	1.8687	3.6521
10	11.0	204611.0	9.8597	929.0913	5.482	44.6929	1.7061	3.2627
11	12.0	209060.0	9.8273	908.4748	5.4311	43.1456	1.7399	3.3401
12	13.0	206165.0	9.6253	789.7835	5.3841	41.7614	1.8924	3.7125
13	14.0	199534.0	9.7757	876.5428	5.4004	42.2361	1.7493	3.362
14	15.0	207231.0	9.7492	860.5864	5.4854	44.7983	1.7787	3.4312
15	16.0	210810.0	9.6779	819.0969	5.3458	40.6673	1.8528	3.612
16	17.0	205760.0	9.7195	843.0376	5.426	42.992	1.7718	3.4148
17	18.0	210251.0	9.6813	821.0573	5.3493	40.7671	1.8532	3.613
18	19.0	206385.0	9.8107	898.1102	5.4688	44.2868	1.7034	3.2568
19	20.0	194481.0	9.7214	844.1924	5.4388	43.3758	1.7574	3.381
20	21.0	209630.0	9.7378	853.8239	5.4599	44.0158	1.7864	3.4496
21	22.0	204774.0	9.8094	897.272	5.5048	45.4064	1.7213	3.2973
22	23.0	202491.0	9.786	882.8225	5.3989	42.1935	1.7579	3.3822
23	24.0	209593.0	9.8586	928.3879	5.5005	45.2718	1.7094	3.2702
24	25.0	209841.0	9.7016	832.6819	5.4149	42.6631	1.8087	3.5034
25	26.0	204927.0	9.6703	814.8172	5.272	38.6399	1.7629	3.3937
26	27.0	211118.0	9.781	879.7895	5.3532	40.8756	1.7528	3.3702
27	28.0	198245.0	9.8439	918.9614	5.4552	43.8725	1.6599	3.16
28	29.0	204944.0	9.9306	975.9249	5.5521	46.9194	1.6023	3.0362
29	30.0	206334.0	9.7687	872.292	5.4234	42.9142	1.7613	3.3899
30	31.0	202983.0	9.7715	874.0241	5.4116	42.5651	1.7577	3.3815
31	32.0	204522.0	9.6387	797.1712	5.3347	40.3558	1.738	3.3357
32	33.0	160092.0	9.836	913.973	5.2283	37.4863	1.4541	2.7398
33	34.0	195769.0	9.8524	924.4292	5.4208	42.8378	1.653	3.1449
34	35.0	204517.0	9.8107	898.1008	5.4569	43.9225	1.6949	3.2375
35	36.0	204195.0	9.8258	907.5157	5.4529	43.8008	1.7211	3.2968
36	37.0	205272.0	9.883	944.2198	5.4979	45.1902	1.6635	3.1678
37	38.0	207865.0	9.6666	812.7103	5.303	39.4775	1.7634	3.395
38	39.0	203079.0	9.8995	955.0706	5.5657	47.3636	1.6251	3.0847
39	40.0	197789.0	9.7589	866.4108	5.3724	41.4235	1.7399	3.34
40	41.0	208283.0	9.7856	882.5799	5.3965	42.1225	1.7613	3.3901
41	42.0	197120.0	9.8397	916.3399	5.3527	40.8614	1.7198	3.294
42	43.0	211464.0	9.7779	877.8625	5.4119	42.5738	1.7252	3.3063
43	44.0	206622.0	9.8033	893.4603	5.4345	43.2461	1.7288	3.3144
44	45.0	203992.0	9.9026	957.1648	5.5062	45.4503	1.6423	3.1215
45	46.0	212022.0	9.719	842.7496	5.4333	43.2103	1.7489	3.361
46	47.0	201295.0	9.8361	914.0023	5.4283	43.0604	1.7074	3.2657
47	48.0	201166.0	9.8319	911.3503	5.4581	43.958	1.6843	3.2138
48	49.0	201984.0	9.9447	985.4964	5.4516	43.7615	1.6325	3.1006
49	50.0	203428.0	9.8366	914.3681	5.4692	44.2995	1.6934	3.2341
50	51.0	202982.0	9.9042	958.223	5.4205	42.8295	1.6401	3.117
51	52.0	210018.0	9.7044	834.2714	5.3903	41.9413	1.7734	3.4187
52	53.0	203444.0	9.7858	882.6889	5.4569	43.924	1.689	3.2243
53	54.0	206970.0	9.8447	919.4696	5.3947	42.0705	1.6916	3.2302
54	55.0	197996.0	9.8666	933.5499	5.424	42.9324	1.6811	3.2066
55	56.0	202274.0	9.865	932.5024	5.4564	43.9081	1.6651	3.1713
56	57.0	196406.0	9.7394	854.7923	5.2854	38.9994	1.7434	3.3483
57	58.0	201972.0	9.8442	919.2081	5.4815	44.6776	1.6812	3.207
58	59.0	199364.0	9.7264	847.1125	5.387	41.8443	1.7653	3.3994
59	60.0	208160.0	9.7715	874.0099	5.4635	44.1236	1.7499	3.3635
60	61.0	199806.0	9.8679	934.4228	5.4632	44.114	1.657	3.1537
61	62.0	208197.0	9.9096	961.8209	5.4458	43.5858	1.6744	3.1919
62	63.0	204749.0	9.8087	896.8353	5.3787	41.6065	1.7451	3.3522
63	64.0	205261.0	9.8188	903.1517	5.4253	42.9714	1.6846	3.2144
64	65.0	202476.0	9.8968	953.2924	5.3761	41.5311	1.6544	3.148
65	66.0	203965.0	9.8608	929.788	5.3428	40.5821	1.7159	3.285
66	67.0	204808.0	9.8292	909.6608	5.4443	43.5403	1.6883	3.2228
67	68.0	207307.0	9.7815	880.1068	5.409	42.4878	1.7452	3.3525
68	69.0	206185.0	9.8273	908.4789	5.4471	43.6259	1.6811	3.2067
69	70.0	204295.0	9.7331	851.057	5.4339	43.2282	1.7281	3.313
70	71.0	206397.0	9.8226	905.5382	5.4089	42.4863	1.693	3.2333
71	72.0	202265.0	9.8074	896.0313	5.4299	43.1095	1.7071	3.2649
72	73.0	206229.0	9.7989	890.7833	5.3888	41.8973	1.6472	3.1324
73	74.0	198571.0	9.8961	952.8667	5.4428	43.497	1.6206	3.0751
74	75.0	208463.0	9.7854	882.4618	5.3431	40.5916	1.7077	3.2664
75	76.0	206510.0	9.8422	917.8859	5.5083	45.517	1.6775	3.1988
76	77.0	206093.0	9.7716	874.094	5.4636	44.1283	1.7531	3.3709
77	78.0	205520.0	9.9012	956.1884	5.4688	44.2865	1.6509	3.1403
78	79.0	202135.0	9.8727	937.5196	5.4455	43.5759	1.6438	3.1248
79	80.0	200390.0	9.8075	896.0764	5.4246	42.9504	1.6781	3.2
80	81.0	202404.0	9.77	873.0685	5.364	41.1843	1.7736	3.4191
81	82.0	198709.0	9.8281	908.9517	5.4307	43.1333	1.6856	3.2168
82	83.0	205411.0	9.8174	902.2619	5.4958	45.1233	1.6838	3.2127
83	84.0	209075.0	9.8893	948.3503	5.4657	44.1929	1.6774	3.1985
84	85.0	207057.0	9.8392	915.9967	5.4424	43.4827	1.6812	3.207
85	86.0	204734.0	9.8194	903.5395	5.4791	44.6036	1.6616	3.1637
86	87.0	207730.0	9.9519	990.4351	5.45	43.7126	1.6359	3.1079
87	88.0	207434.0	9.9233	970.9867	5.4534	43.8169	1.6074	3.0471
88	89.0	201363.0	9.805	894.5662	5.4127	42.597	1.7035	3.257
89	90.0	200538.0	9.9404	982.5813	5.466	44.2012	1.6172	3.0678
90	91.0	198387.0	9.8361	914.0464	5.4075	42.4442	1.687	3.2199
91	92.0	207086.0	9.8925	950.4548	5.4765	44.5223	1.6476	3.1331
92	93.0	197918.0	9.846	920.352	5.4613	44.0578	1.6686	3.1791
93	94.0	198549.0	9.9095	961.7226	5.4157	42.6868	1.6315	3.0984
94	95.0	204401.0	9.9181	967.5112	5.4846	44.7733	1.6134	3.0596
95	96.0	206833.0	9.9795	1009.5784	5.5555	47.0302	1.5684	2.9658
96	97.0	202249.0	9.8937	951.2311	5.4674	44.2435	1.6461	3.1299
97	98.0	208171.0	9.8533	925.0215	5.4839	44.7528	1.6817	3.2081
98	99.0	199466.0	9.8416	917.5081	5.4971	45.1654	1.6429	3.123
99	100.0	202406.0	9.9448	985.5914	5.499	45.2248	1.5689	2.9667

Figure 6： Top 16 frequent words in the Gutenberg Corpus

Figure 7： Top 16 frequent words in the Chinese Wiki corpus

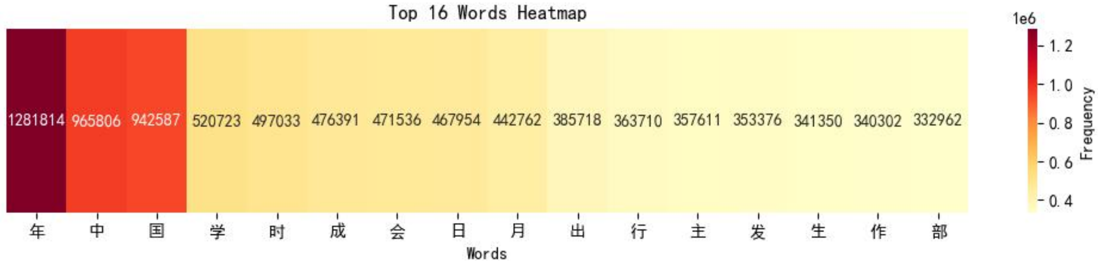

Figure 8： Top 16 frequent characters in the Chinese Wiki corpus

Conclusions

The analysis of the Gutenberg Corpus (English) and the Chinese Wiki Corpus (Chinese) using unigram, bigram, and trigram models reveals significant insights into the lexical and structural differences between the two languages, as well as the impact of model complexity on text predictability. Below are conclusions drawn based on the findings:

Firstly, the frequency heatmaps for English letters and words demonstrate a relatively dispersed distribution. High-frequency words such as “shall”, “said”, and “unto” appear with moderate frequency, indicating a diverse vocabulary usage. Similarly, the letter frequency heatmap shows a balanced distribution, with common letters like “e” and “m” appearing frequently but not overwhelmingly so. In contrast, Chinese character and word frequency heatmaps reveal a more concentrated distribution. High-frequency characters such as "年" (year) and "中" (middle/China) dominate the corpus, reflecting a higher degree of lexical repetition. This suggests that Chinese text relies more heavily on a smaller set of core characters and words compared to English.

Secondly, the entropy and perplexity results for English indicate a higher degree of unpredictability at both the character and word levels. Unigram entropy values for English words are notably higher compared to Chinese, reflecting the greater lexical diversity and lower repetition of words in English texts. Bigram and trigram entropy values further highlight the importance of context in English, as predictability increases with additional preceding words. Chinese exhibits lower unigram entropy values, suggesting a more predictable structure at the character level due to the frequent repetition of common characters. However, the higher entropy at the word level indicates that Chinese words, when considered as units, are less predictable due to the combinatorial nature of character usage.

Thirdly, across both languages, the transition from unigram to bigram and trigram models shows a consistent decrease in entropy and perplexity. This trend underscores the importance of context in language modeling. For example, in English, bigram entropy drops significantly from unigram levels, and trigram entropy further reduces this unpredictability. A similar pattern is observed in Chinese, where bigram and trigram models capture more contextual information, reducing the uncertainty in character and word sequences.

Finally, at the character level, Chinese exhibits a more predictable structure due to the high frequency of common characters. English, on the other hand, shows a more balanced distribution of letter frequencies, with no single character dominating the corpus. At the word level, English demonstrates greater lexical diversity, as evidenced by higher unigram entropy and perplexity values. Chinese, while more predictable at the character level, shows increased complexity at the word level due to the combinatorial nature of character usage.

References

[1] Brown, P.F., Pietra, S.D., Pietra, V.J., Lai, J.C., & Mercer, R.L. (1992). An Estimate of an Upper Bound for the Entropy of English. *Comput. Linguistics, 18*, 31-40.

[2] Sloane, N.J., & Wyner, A.D. (1951). Prediction and Entropy of Printed English.

Note: some code used in this study were generated with the assistance of artificial intelligence.