# Spoken and Natural Language Understanding Practical Assignment 1

### Submitted by Jiahui Dai

### March 27, 2025

Accompanying material: Assignment_1.ipynb

## 1 Zipf's Law

### List of unique words sorted to their frequency in descending order

Shown in accompanying Jupyter Notebook.

### Discussion on findings

Zipf's Law states that in a large collection of words, the frequency of any word is inversely proportional to its rank:

$$f \propto \frac{1}{r}$$

where

- $f$: frequency of word

- $r$: rank of the word

This means that the most frequent word occurs twice as often as the second most frequent, three times as often as the third most frequent, and so on.

With reference to the jungle book dataset, words that are short in length occur in higher frequencies compared to words with longer lengths. For example, "the", "and", "of" occur in higher frequency compared to "produce", "subscribe", "newsletter" of lower frequency.

To verify Zipf's law on a textual corpus, using the jungle book dataset, chi-square goodness of fit test is performed.

- $H_0$: The observed frequencies are equal to the expected frequencies.

- $H_1$: The observed frequencies are not equal to the expected frequencies.

$$\chi^2_{\text{statistics}} = 19760.11496822835$$
$$\chi^2_{\text{critical}} = 5107.674219300448$$
$$\chi^2_{\text{statistics}} > \chi^2_{\text{critical}} \qquad \qquad \text{(Reject } H_0)$$

As the $H_0$ is rejected, this suggests that observed frequencies are not equal to the expected frequencies, and that Zipf's Law does not hold.

With reference to Figure 1, there is a general downward trend of frequency compared to rank. The statistical analysis shows that the observed frequency is not equal to the expected frequency, suggests that the jungle book dataset contains words that does not follow Zipf's Law significantly. If Zipf's Law holds, the observed line (blue) should form a straight line with a slope close to -1, as depicted by the expected line (red) in Log-Log curve. However, the observed line does not follow a slope close to -1 in Log-Log curve.
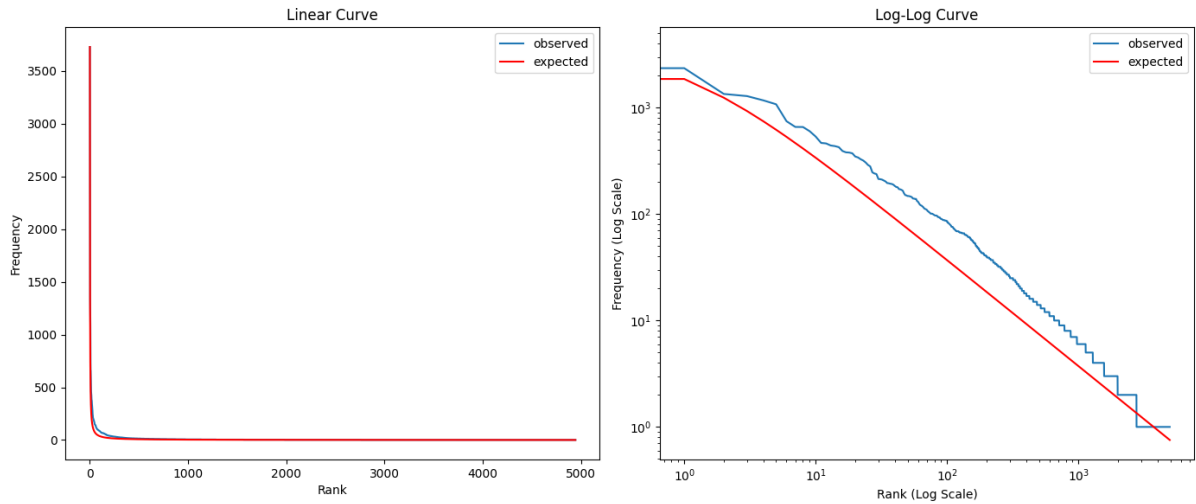
Figure 1: Linear and Log-Log Curves.
Observed line (blue) follows the frequency of words observed in the jungle book dataset. Expected line (red) follows the theoretical word frequency according to Zipf's law.

# 2 Mutual Information

# 3 Wikipedia Language Model

| w1 | w2 | pmi |
|---|---|---|
| machua | appa | 8.54334 |
| united | states | 8.30218 |
| literary | archive | 8.23318 |
| cold | lairs | 7.69419 |
| archive | foundation | 7.57926 |
| bandar | log | 7.41487 |
| petersen | sahib | 7.38589 |
| stretched | myself | 7.33937 |
| paragraph | f | 7.33937 |
| hind | legs | 7.23466 |
| fore | paws | 7.15704 |
| hind | flippers | 7.13457 |
| tree | tops | 7.02921 |
| troop | horse | 7.02804 |
| bath | room | 7.00289 |
| twenty | yoke | 6.98519 |
| paragraph | e | 6.97472 |
| electronic | works | 6.95225 |
| master | words | 6.91588 |
| whole | line | 6.90405 |
| years | ago | 6.88771 |
| within | days | 6.87221 |
| bring | news | 6.86936 |
| waingunga | river | 6.76685 |
| killing | grounds | 6.75158 |
| council | rock | 6.71941 |
| villagers | lived | 6.67154 |
| monkey | folk | 6.64622 |
| black | panther | 6.53574 |
| copyright | laws | 6.49858 |

Table 1: Top 30 Rows Table

| w1 | w2 | pmi |
|---|---|---|
| the | the | −5.50709 |
| and | and | −4.58736 |
| he | the | −4.28052 |
| the | he | −4.28052 |
| the | to | −3.7514 |
| of | of | −3.47152 |
| i | and | −3.30003 |
| they | the | −3.23617 |
| to | he | −3.21798 |
| was | and | −3.1072 |
| for | and | −2.96939 |
| is | and | −2.96079 |
| and | is | −2.96079 |
| had | the | −2.95281 |
| of | in | −2.88121 |
| little | the | −2.76773 |
| to | i | −2.69736 |
| he | in | −2.67243 |
| but | and | −2.65494 |
| that | and | −2.64759 |
| of | and | −2.64314 |
| at | and | −2.58825 |
| we | the | −2.57623 |
| not | and | −2.55316 |
| man | the | −2.47589 |
| he | and | −2.43436 |
| were | the | −2.43042 |
| and | of | −2.42 |
| of | for | −2.41147 |
| is | of | −2.40287 |

Table 2: Bottom 30 Rows Table