Assignment 1: Zipf's Law, Stemming, Lemmatization

Name: Achal Gawande
Roll Number: 22075004
Email: achal.gawande.cse22@itbhu.ac.in

January 24, 2025

Theory

Zipf's Law

Zipf's Law states that in a corpus of natural language, the frequency of any word is inversely proportional to its rank in the frequency table. Mathematically, this relationship can be expressed as:

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

where:

- $f(r)$ is the frequency of the word with rank r.
- r is the rank of the word when sorted by frequency.

This implies that the most frequent word will appear approximately twice as often as the second most frequent word, three times as often as the third most frequent word, and so on. Taking the logarithm on both sides:

$$\log f(r) = -\log r + C$$

where C is a constant. When plotted on a log-log scale, the relationship forms a straight line.

Porter Stemmer

The Porter Stemmer is an algorithm for removing common morphological and inflectional endings from words in English. It operates using a sequence of conditional rules applied iteratively to transform words into their base or root forms. For example:

- ”running” $\rightarrow$ ”run”
- ”happily” $\rightarrow$ ”happi”
- ”studies” $\rightarrow$ ”studi”

The rules focus on suffix removal and handle various cases like plurals, tenses, and derivations.

Implementation Details

Porter Stemmer for English

For the English language, we used the `nltk` library, which provides a pre-implemented version of the Porter Stemmer. This ensures accuracy and saves development time.

Stemmer for Hindi and Marathi

For Hindi and Marathi, we created custom stemmers by writing rule-based transformations. These rules consider the linguistic morphology of the respective languages, such as common suffixes for verbs and nouns. Example rules include:

- Hindi: ”करना" → "कर"
- Marathi: "चालले" → "चाल"

These rules were applied programmatically to the respective corpora.

English Rules: Porter stemmer rules include:

- Remove plural suffixes (e.g., ’cats’ → ’cat’).
- Convert past tense (e.g., ’hopping’ → ’hop’).
- Reduce words to base forms (e.g., ’running’ → ’run’).
- Remove derivational suffixes (e.g., ’national’ → ’nation’).
- Simplify endings (e.g., ’happiness’ → ’happy’).

Hindi Rules: Indian Language stemmer rules include:

- Remove common inflectional suffixes indicating gender, number, tense, and case.”
- Reduce different forms of the same word to their base or root form.
- Prevent overly aggressive stemming that might lead to extremely short or meaningless roots.”

Marathi Rules: Marathi Porter Stemmer Rules:

- Remove inflectional suffixes indicating gender, number, tense, and case.
- Examples:
  - Remove short suffixes: "◌ा", "◌ी", "◌े", "◌ु".
  - Remove common derivational suffixes: "ने", "नी", "ता", "ते".
  - Reduce suffixes for possessive forms or locative forms: "च्या", "तांचा", "तासह".
  - Simplify longer phrases ending with: "संपर्कात", "वाढवलेला".
- Process suffixes in decreasing order of length to avoid premature truncation.”
- Avoid truncating overly short words to meaningless roots.

Statistics and Results

Word Statistics

The following table summarizes the statistics for total words, unique words, and stemmed words for each language:

Language	Total Words	Unique Words	Stemmed Words
English	48262834	312016	248096
Hindi	69834603	321129	271604
Marathi	16938502	736845	608156

Table 1: Word statistics for the analyzed corpora.

Zipf's Law Graphs

The following figures illustrate the frequency vs. rank plots for the three languages, demonstrating adherence to Zipf's Law:

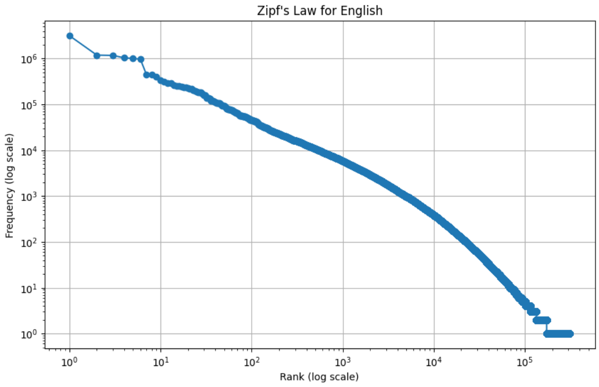

Figure 1: Log-log plot of Frequency vs. Rank for English corpus.

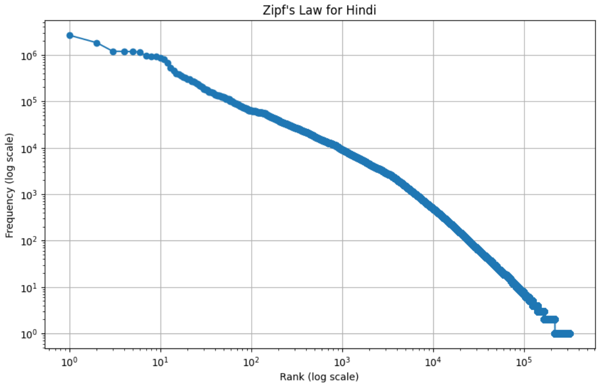

Figure 2: Log-log plot of Frequency vs. Rank for Hindi corpus.

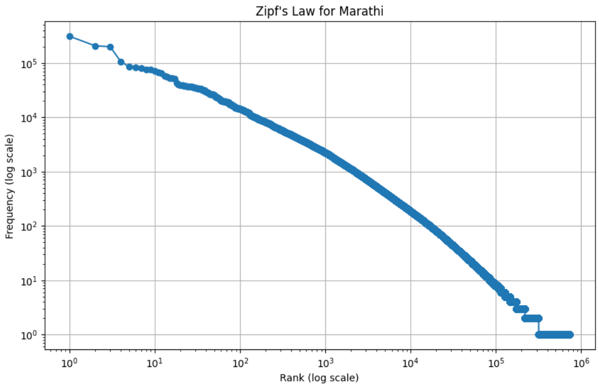

Figure 3: Log-log plot of Frequency vs. Rank for Marathi corpus.

Correlation Calculation

To quantify the adherence to Zipf's Law, we calculated the correlation between log$(frequency)$ and log$(rank)$ for each dataset. The results are as follows: These results confirm that all three languages follow Zipf's Law with high accuracy, as evidenced by the strong negative correlations.

Language	**Correlation Coefficient**
English	-0.9877
Hindi	-0.9932
Marathi	-0.9704

Table 2: Correlation coefficients indicating adherence to Zipf's Law.

Conclusion

In this assignment, we:

- Validated Zipf's Law for English, Hindi, and Marathi corpora.
- Implemented stemming for English using `nltk`'s Porter Stemmer and custom rule-based stemmers for Hindi and Marathi.
- Compared word statistics before and after stemming.

The results demonstrate the universality of Zipf's Law and the effectiveness of stemming techniques across different languages.

Resources

To access the cleaned dataset and code, follow the links below:

- **Dataset (Cleaned):** Google Drive
- **Code:** Github   Kaggle