Maths + Fundamentals of AI Project
On Coldplay Dataset

Zipf's Law: Finding Hidden Patterns in Data

By Team Roomies❤️🌻

Zipf's Law is a statistical principle that describes how the frequency of words in natural language follows a specific pattern. It is commonly observed in linguistics, data science, and information theory.

ZIPF'S LAW

Zipf's Law states that the frequency of any word is inversely proportional to its rank in a frequency table. In simple terms:

- The most frequent word will appear twice as often as the second most frequent.
- The second will appear three times as often as the third.
- And so on.

ITS IMPORTANCE

Its explanations-

- When words are ranked by their frequency in a large text corpus, a small number of words appear very frequently.
- Most words appear only a few times.
- This kind of distribution is called a power law distribution.

Its real world use cases -

- Search engines: prioritize keywords based on frequency.
- Natural Language Processing (NLP): optimize vocabulary and token usage.
- Data compression: allocate shorter codes to high-frequency words.

MATHEMATICAL REPRESENTATION

This relationship can be written as:

$$f(r) \propto 1 / r^s$$

Where:

- f(r) = frequency of the word ranked r
- r = rank of the word
- s ≈ 1 (a constant for natural languages)

MATHEMATICAL BEHAVIOUR

If the highest-frequency word occurs N times:

- The second will occur approximately N/2 times.
- The third ≈ N/3, and so on.

On a log-log plot (log of frequency vs log of rank), this results in a straight line with a negative slope, which confirms the power law behavior.

Tanima Samanta

- Conducted detailed research on the concept and theoretical background of Zipf’s Law.
- Helped in implementing the Python code used for analyzing word frequencies in the dataset.
- Assisted in generating and refining visualizations, especially the rank-frequency graphs.

Koyna Arya

- Handled data preprocessing tasks such as cleaning the input text, tokenizing words, and removing stopwords.

- Took responsibility for documenting the workflow and approach using Markdown in the Colab notebook.

- Contributed to analyzing the output data and identifying trends that aligned with Zipf’s Law.

Aparajita K Singh

- Contributed to the literature review by exploring existing studies and use cases of Zipf's Law.
- Derived insights from the visualized word distributions and helped explain their meaning.
- Assisted in testing and validating the code to ensure correctness and consistency of results.

Riddhi Khera

- Focused on designing and styling visual plots such as log-log graphs and rank-frequency charts.

- Wrote explanations to describe the behavior observed in the data and how it matched Zipfian expectations.

- Verified that the results followed the mathematical model of Zipf’s Law and edited the final project output with the presentation.

Thank you!