Exploring Zipf's Law with Dua Lipa's Lyrics

A Data-Driven Dive into Word Frequency Patterns

Team Name : Cube

Presented by:

- Sneha Chepurwar
- Gayatri Jaiswal
- Ashu Choudhary
- Krish Patil

Contributions :

- Sneha Chepurwar: Data Preprocessing, Lyrics Cleaning, and Word Frequency Analysis

- Gayatri Jaiswal: Applied Zipf's Law, Created Visualizations Bar Plot, Log-Log Graph, and Interpreted Results

- Ashu Choudhary: Exploratory Data Analysis (EDA), Detected Outliers, and Prepared Graphical Insights

- Krish Patil: Created Presentation Slides, Summarized Insights, and Reported Final Observations

What is Zipf's Law?

A natural phenomenon stating that the frequency of any word is inversely proportional to its rank in the frequency table.

Dataset Used: DuaLipa.csv

Columns Present: Artist, Title, Album, Year, Date, Lyric

Focus: We're analyzing the Lyric column for Zipfian behavior.

Data Cleaning & Preprocessing:

- Combined all lyrics into one text corpus

- Converted to lowercase

- Removed punctuation and special characters

- Tokenized into individual words

- Removed common English stopwords

Observation:

The plot approximately follows a straight line in the log-log plot.

Conclusion:

Dua Lipa's lyrics exhibit Zipfian distribution, validating the law in artistic text.

Zipf's Law: Word Frequency vs Rank
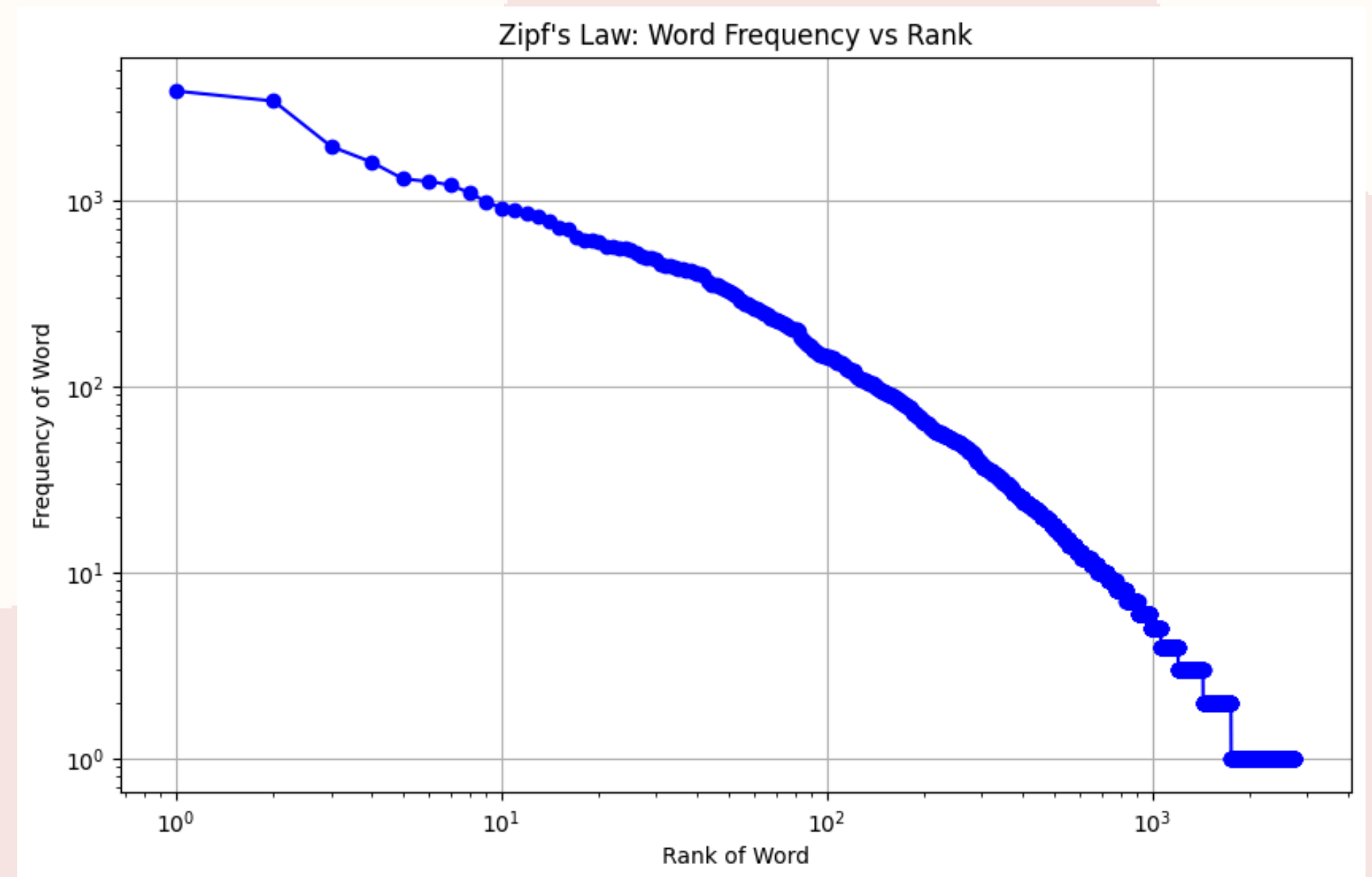
We hope our exploration of Zipf's Law using the Dua Lipa Lyrics dataset gave you a fresh perspective on the fascinating patterns hidden in language and music.

We'd love to share more of what we discovered.

— Team Cube