Zipf's Law

Presented by Hunters

Team Project

1. **Vikash Kumar**
   EMAIL Id:- vikash.kumar@adypu.edu.in

2. **Suraj Kumar Rai**
   EMAIL Id:- suraj.rai@adypu.edu.in

3. **Sahil Lakshman Khemnar**
   EMAIL Id:- Sahil.khemnar@adypu.edu.in

4. **Rahul Dhakar**
   EMAIL Id:- Rahul.dhakar@adypu.edu.in

INTRODUCTION

Zipf's Law is a principle that describes a common pattern found in natural languages and many other systems. It states that in any sufficiently large sample of text, the frequency of a word is inversely proportional to its rank in the frequency table. This means that the most frequent word in a language will appear approximately twice as often as the second most frequent word, three times as often as the third, and so on. For example, in English, the word "the" is usually the most frequent, followed by "of", "and", "to", etc., each decreasing in frequency in a predictable way.

APPLICATION

Zipf’s Law has several practical applications across different fields. In natural language processing (NLP), it helps in filtering out common stopwords and selecting meaningful vocabulary for language models. Search engines use it to rank and suggest relevant queries by analyzing word frequency patterns. In urban studies, it explains city size distributions, aiding in planning and resource allocation. On the internet, user activity, website traffic, and social media trends often follow Zipf-like distributions, which are useful for analytics and optimization. It also supports data compression techniques and can assist in cryptanalysis by predicting frequent patterns. Overall, Zipf’s Law helps in simplifying complex systems by focusing on the most impactful elements.

DATA ANALYSIS

Data analytics of Zipf’s Law involves analyzing word frequencies in a text.

Words are ranked by frequency, then plotted on a log-log scale.

A straight line indicates Zipf’s Law holds, showing a few words dominate usage.

This insight helps in keyword extraction, NLP optimization, and data compression.

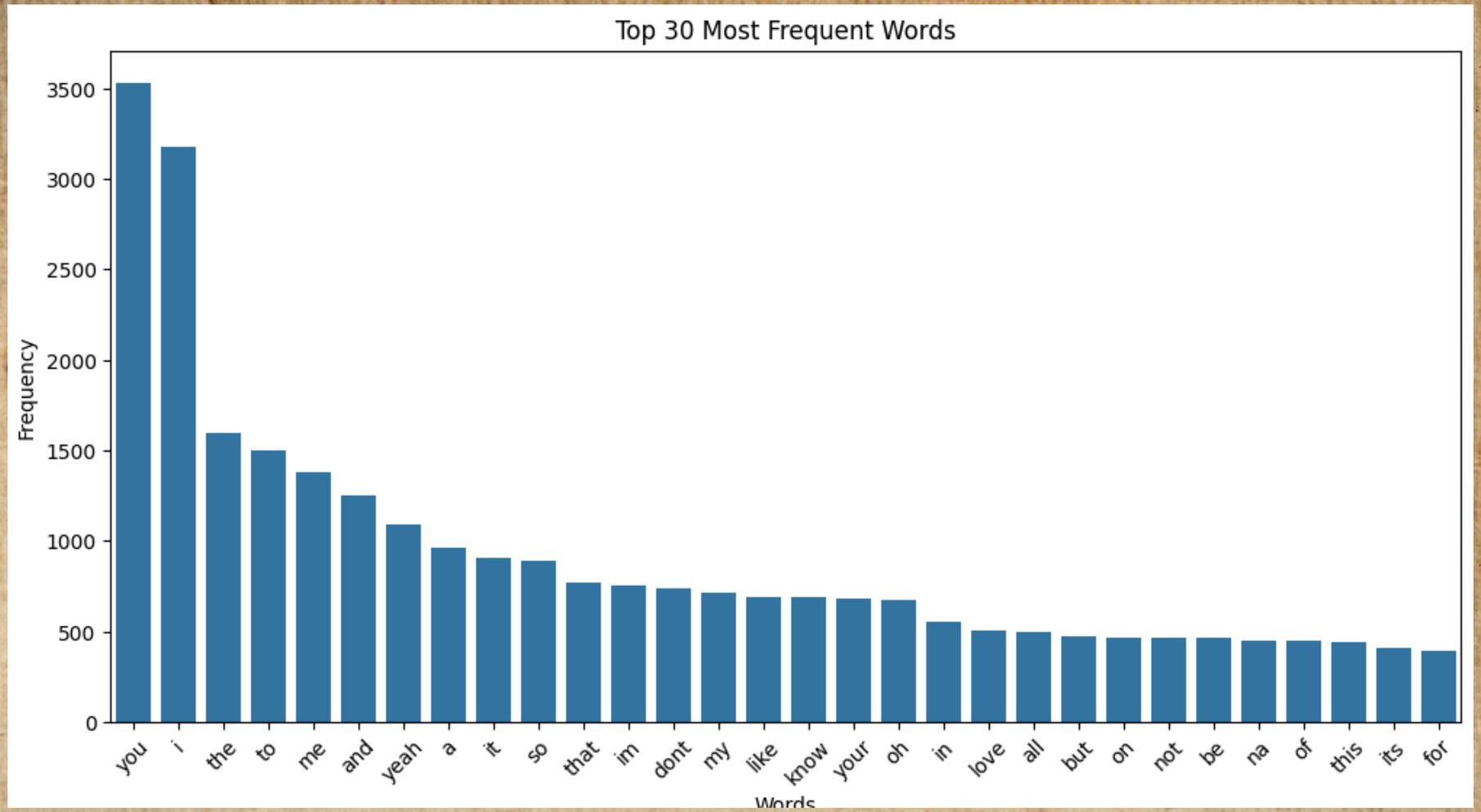
PROJECT OVERVIEW

Data Collection :- Dataset sourced from a Google Drive CSV file containing textual content (e.g., song lyrics or articles).

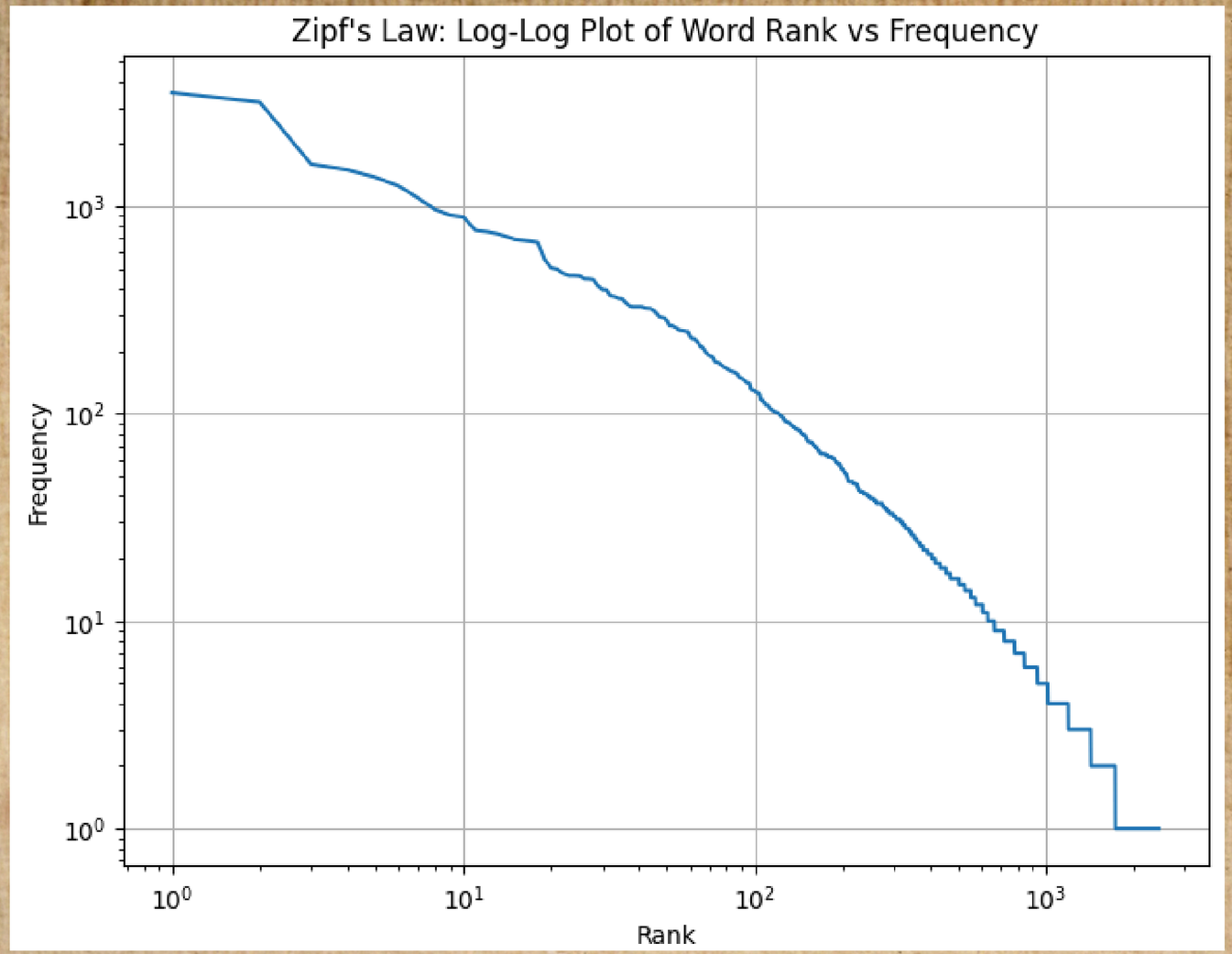
Word Frequency Analysis :- Count occurrences of each word using Counter and visualize the top frequent words with a bar chart.

Visualization :- A bar graph was created to visually show the match top 30 most frequent words and log-log plot of world rank vs frequency.

GRAPH


Top 30 Most Frequent Words
Frequency
0
500
1000
1500
2000
2500
3000
3500
you
i
the
to
me
and
yeah
a
it
so
that
im
dont
my
like
know
your
oh
in
love
all
but
on
not
be
na
of
this
its
for
Words

GRAPH


Zipf's Law: Log-Log Plot of Word Rank vs Frequency
Frequency
Rank
10^0
10^1
10^2
10^3

THANK YOU