Zipf’s Law & Heap’s Law

Dr. Sambit Praharaj

Assistant Professor (II)

Why Study Statistical Laws of Language?

- Natural language follows strong statistical patterns
- Word usage is highly non-uniform
- These laws help us understand language structure
- They guide the design of NLP systems

Zipf’s Law – Definition

- Zipf’s Law relates word frequency to word rank
- Frequency is inversely proportional to rank
- Formula: Frequency ∝ 1 / Rank
- A small number of words dominate the corpus

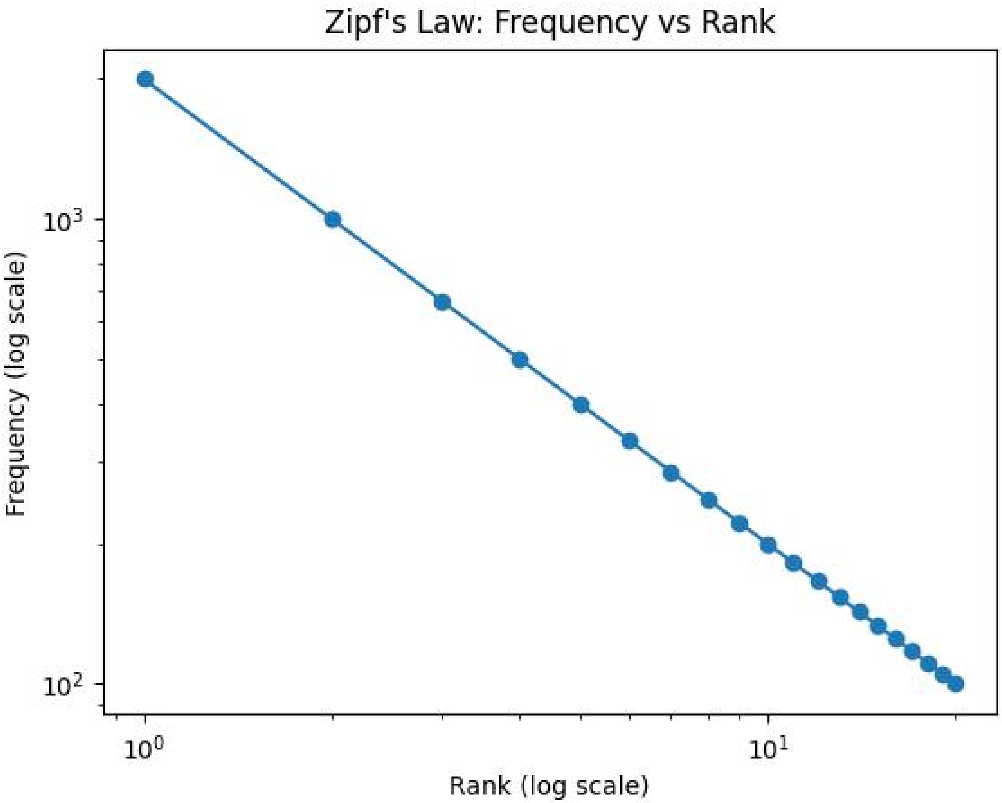
Zipf’s Law – Intuition

- Function words appear in almost every sentence
- Speakers prefer minimal effort while communicating
- Content words vary with topic
- This creates a long-tail distribution

Zipf’s Law – Numerical Example

- Rank 1 ('the'): ~2000 occurrences
- Rank 2 ('of'): ~1000 occurrences
- Rank 5: ~400 occurrences
- Rank 20: ~100 occurrences

Zipf’s Law – Log–Log Graph



Importance of Zipf’s Law in NLP

- Explains dominance of stop words
- Helps in language modeling
- Motivates smoothing techniques
- Explains rare-word problem

Heap’s Law – Definition

- Heap’s Law describes vocabulary growth with corpus size
- Vocabulary size increases as more text is processed
- Formula: $V(N) = k \times N^{\beta}$
- β usually ranges between 0.4 and 0.6

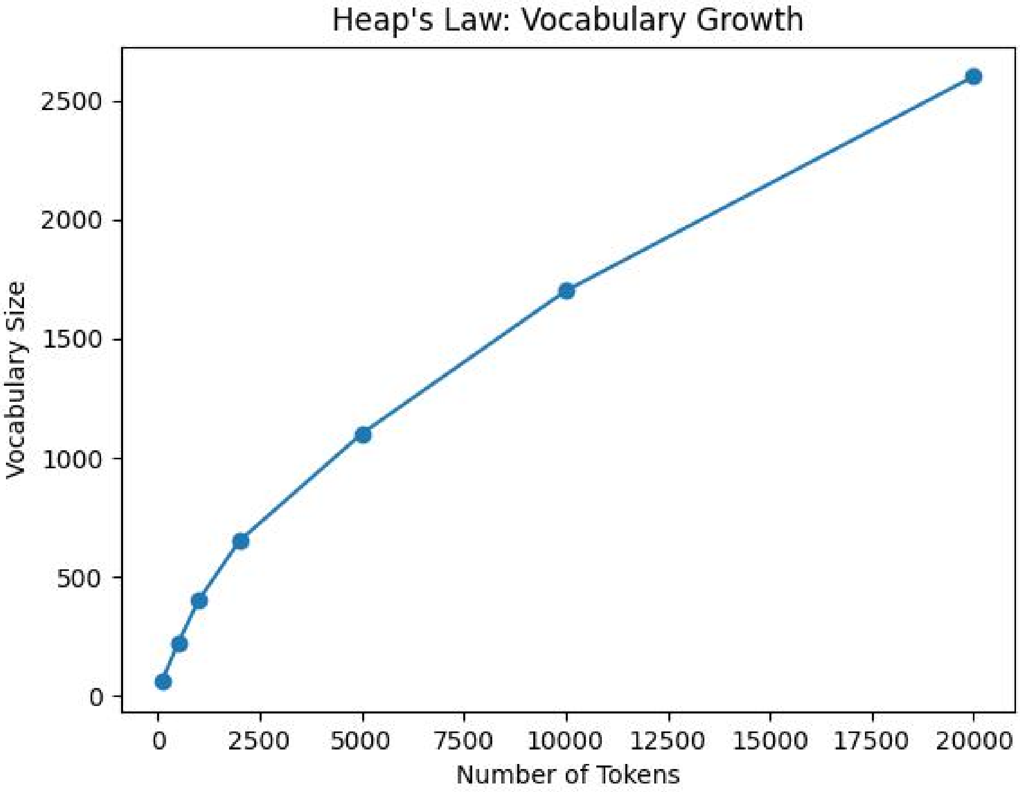
Heap’s Law – Intuition

- Frequent words repeat often
- New words appear as topics change
- Rare words continue to emerge
- Vocabulary growth slows but never stops

Heap’s Law – Numerical Example

- 1,000 tokens → ~400 unique words
- 2,000 tokens → ~650 unique words
- 10,000 tokens → ~1,700 unique words
- Growth is sub-linear

Heap’s Law – Vocabulary Growth Graph



Importance of Heap’s Law in NLP

- Helps estimate vocabulary size
- Useful for memory planning
- Motivates subword tokenization
- Explains data sparsity

Zipf’s Law vs Heap’s Law

- Zipf’s Law: Frequency vs Rank
- Explains distribution of frequent and rare words
- Heap’s Law: Vocabulary vs Corpus Size
- Explains vocabulary growth
- Both are power-law relationships