



INTRODUCTION TO LANGUAGE MODELS

INTRODUCTION TO
LANGUAGE MODELS

July 17-21, 2023

Ramaseshan Ramachandran

AI Cycle
Symbolic Reasoning
Probabilistic Reasoning

Generative AI
Neural Model-based Reasoning
Models

Ramaseshan

- ▶ **Perception** - ability to receive and evaluate information about an environment
- ▶ **Learning** - ability to find common patterns, organize the knowledge
 - Represent knowledge using Propositional and first-order logic - symbolic representation
 - Parameters using statistics and probability - probabilistic representation
- ▶ **Reasoning** - Draw conclusions from the knowledge and learnings such as facts, beliefs, observations/evidence, logical rules
- ▶ Use available/partial information for problem solving
- ▶ Interpretability of reasoning
- ▶ Planning and Execution

- ▶ Knowledge is available as types, relations and their complex relationship - represented in the form of first-order and propositional logic

Who is an all rounder in cricket?

$$\exists x(\text{Bats well}(x) \wedge \text{Fields well}(x) \wedge \text{Bowls well}(x)) \Rightarrow \text{All Rounder}(x)$$

The facts are hard facts - there could be no uncertainty

- ▶ Knowledge is available in the form of parametric relationships - cooccurrences, correlations, incidence matrix, conditional relationships
- ▶ Handles uncertainty well using the observed parameters
- ▶ $\exists w_i \exists w_j (\text{Similar Contexts}(x) \wedge \text{Similar Contexts}(y)) \rightsquigarrow \text{Similar}(x, y)$
- ▶ How are related in a corpus?
- ▶ Context words may differ and not fixed

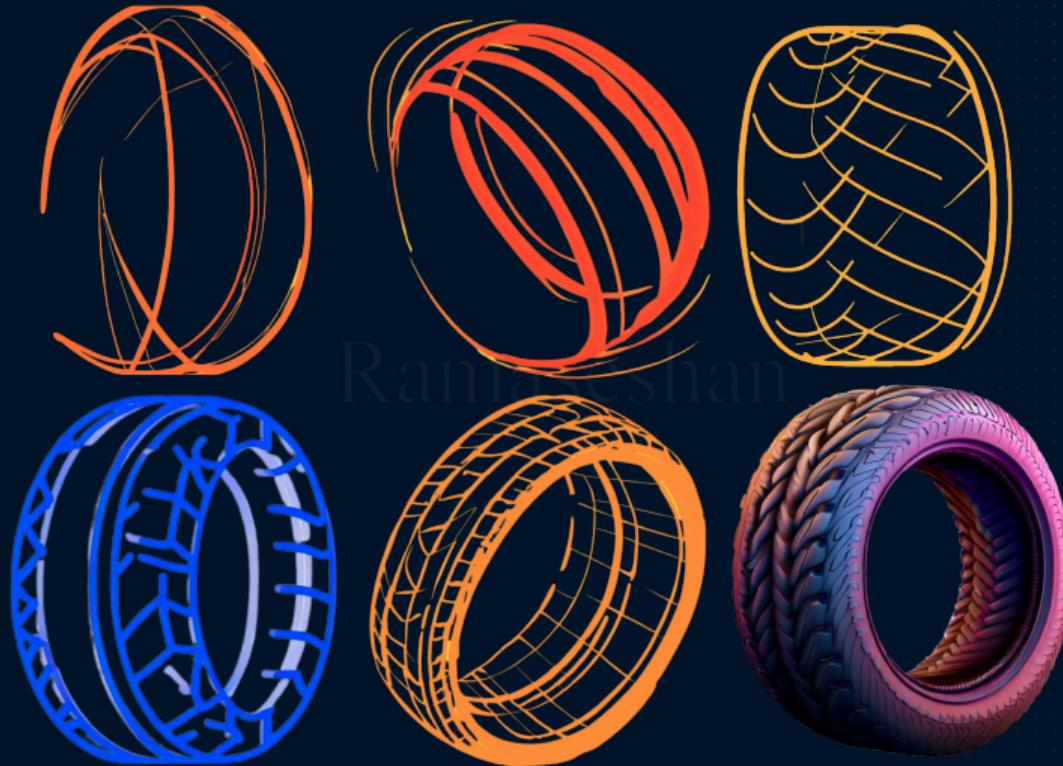
- ▶ Learns the patterns of existing content
- ▶ Creates new content such as text, images, or music using learned patterns
- ▶ Generates realistic text, images, and music
- ▶ Hallucinates on unknown areas with known keywords - Generates grammatically correct sentence but factually incorrect sentence like *The cat barks*

- ▶ Combines generalizations of symbolic and probabilistic reasoning
- ▶ Adds learning capabilities
- ▶ Backbox models

Ramaseshan

Don't present decision making evidences and are difficult to interpret

MODELS



- ▶ Represent encoded information collected from observed data
- ▶ Facilitates prediction of future events
- ▶ Possible next word in a sentence
 - Lexical Knowledge
- ▶ Validates a sentence using learned parameters
- ▶ Representation of words as vectors
- ▶ Understanding word senses
 - Semantic Knowledge
- ▶ Anaphora detection
- ▶ Unsupervised and Supervised
- ▶ No fixed procedural logic
- ▶ Learns input-output relationships

PROPERTIES OF A MODEL

- ▶ Uncovers latent patterns that is difficult to visualize using rigid procedural method
 - ▶ Information related correlations
 - ▶ n-gram frequencies, lexical and semantic relationships, word representations as vectors
 - ▶ Incidence matrix of n-grams
- ▶ Does not use any schema and no human annotation
- ▶ Helps in understanding and interpreting information for decision making/classification
- ▶ Allows open class queries
- ▶ Makes data driven decisions under uncertainty
- ▶ Presents Inferences based on the domain knowledge - conditional probability
- ▶ Store vast amount of linguistic and semantic as parameters knowledge
- ▶ Accesses the knowledge by using the context

Are you _____ in playing cricket?

Do you _____ football?

You went to a movie yesterday. How _____ the movie?

Do you _____ a car?

How are _____ ?

Ramaseshan

The subject was _____. Most students understood it

- ▶ Big Data analysis includes both structured and unstructured data
- ▶ 90% of the data in the business and in the Web Internet is unstructured
 - Text files, audio, video, web pages, pdf files, social media content, presentations, transcripts of audio, video, etc.

Allows interaction among humans to share information using a set of words and sentences constructed using a finite set of alphabets and framed using a set of grammar rules

- ▶ Arbitrary
- ▶ Structured
- ▶ Generative
- ▶ Dynamic

Intended for Human Machine Communications. Instructions are

- ▶ Precise
- ▶ Unambiguous
- ▶ Mathematical equations

Ramaseshan

IS NLP HARD?

- ▶ Creative and analytical representation of thoughts
- ▶ What is added with 15 to get 45
- ▶ Juvenile court to try shooting defendant
- ▶ Safety experts say school bus passengers should be belted
- ▶ The king saw a rabbit with his glasses
- ▶ Local high school dropouts cut in half



I saw her duck



The chicken is ready to eat

COMMON LAYERS OF NLP APPLICATIONS

- ▶ Preprocessing layer
- ▶ Data extraction layer
- ▶ Analysis of extracted information
- ▶ Semantic understanding
- ▶ Human/automatic evaluation of word meaning, sentence structure using the content obtained from the previous layers
- ▶ Pre-training
- ▶ Fine tuning

BASIC OPERATIONS ON A CORPUS

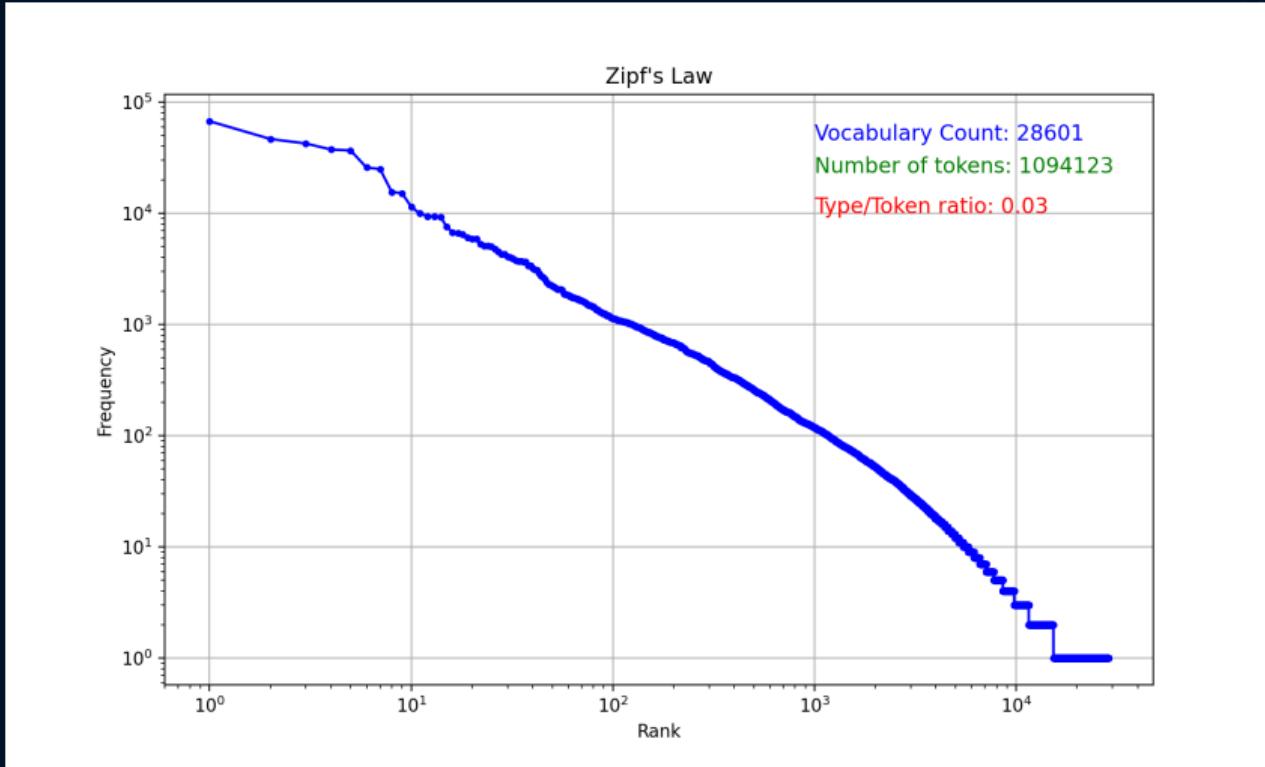
- ▶ Normalization - convert into a single canonical form
- ▶ Tokenization
- ▶ Counting

Ramaseshan

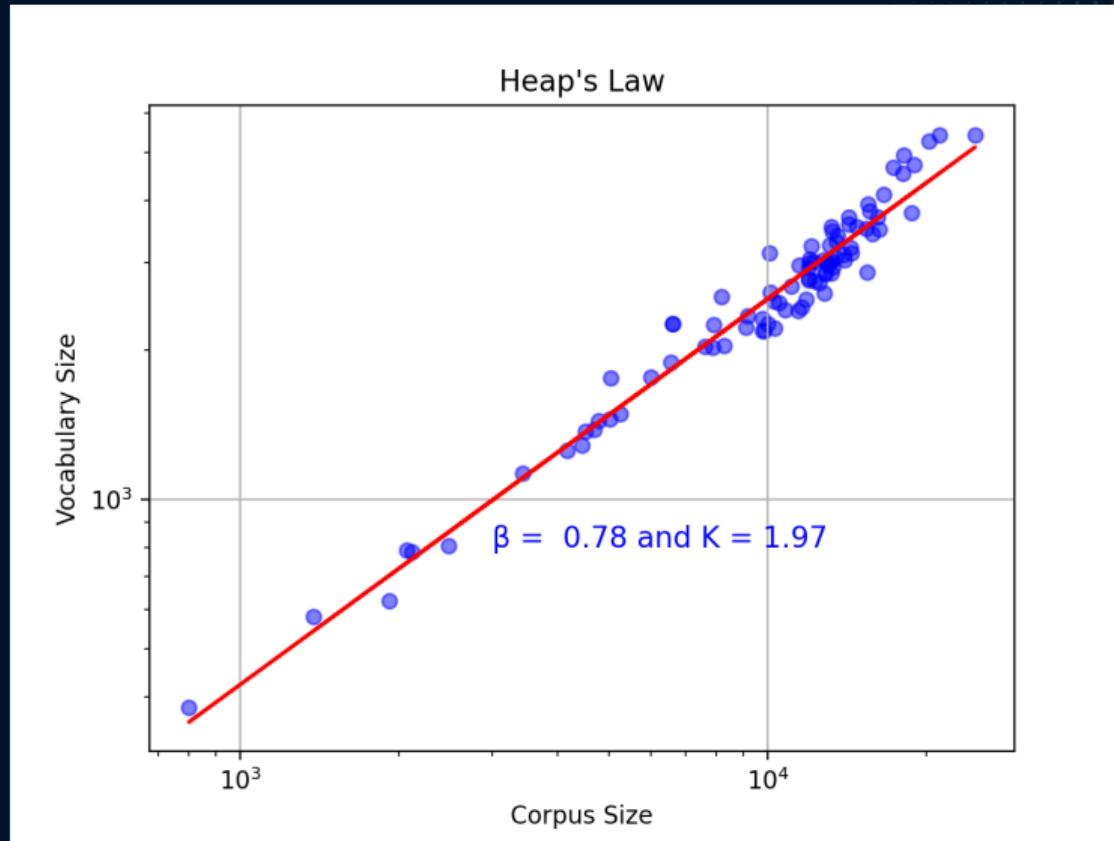
Describe word-rank and word-frequency distribution, vocabulary, and terms in a corpus

- ▶ Zipf's law - The frequency of any word is inversely proportional to its rank - $f \propto \frac{1}{r^\alpha}$
 $\alpha \approx 1$ and r is the frequency rank of the word
- ▶ Heap's law - The number of unique words in a text of n words is
- ▶ $V \propto N^\beta$
- ▶ V is the number of unique terms/vocabulary and N is the total number of terms in the corpus

ZIP'S LAW



HEAPS'S LAW



Thank you
Ramaseshan Ramachandran
Ramaseshan