NATURAL LANGUAGE PROCESSING

An Introduction

Ramaseshan Ramachandran

How's it going?

What's new?

How have you been?

How's everything?

How are you holding up?

What's up?

How are you?

What are you up to?

How are things going?

What's happening?

What's going on?

How are you doing?

How's life?

FILL IN THE BLANKS

Are you _____ in playing cricket?

Do you _____ football?

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

Do you _____ a car?

How are _____?

The subject was _____. Most students understood it

UNSTRUCTURED CONTENT

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.
Photos

LANGUAGE

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

PROGRAMMING LANGUAGE IN

Intended for Human Machine Communications

Instructions are

- Precise
- Unambiguous
- Mathematical equations

Is NLP Hard?

Creative and analytical representation of thoughts

What is added with 15 to get 45?

Juvenile court to try shooting defendant

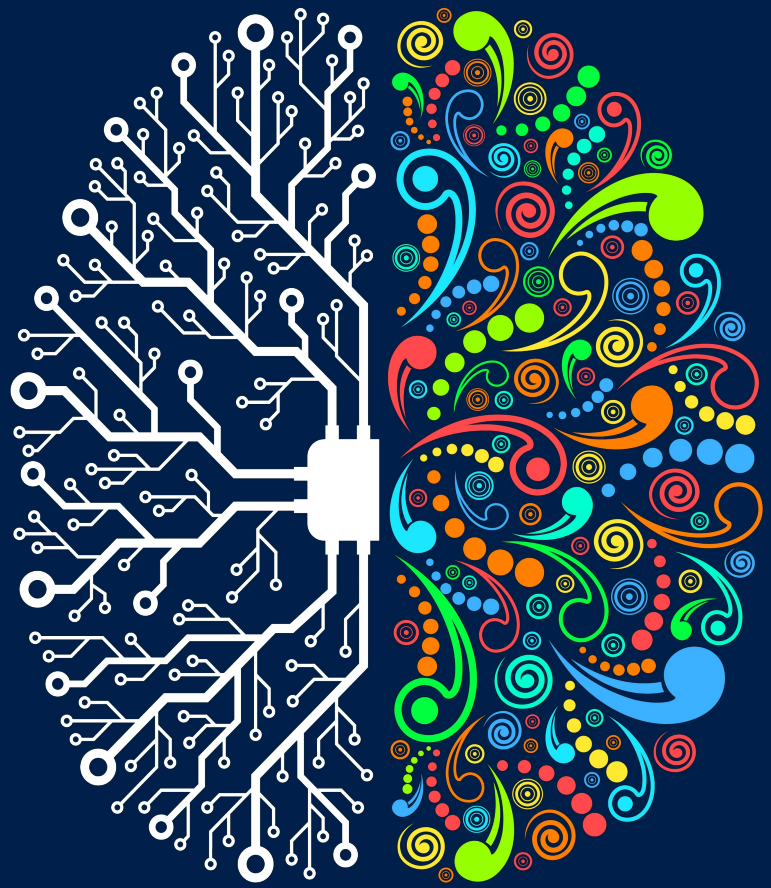
Safety experts say school bus passengers should be belted

The king saw a rabbit with his glasses

Local high school dropouts cut in half

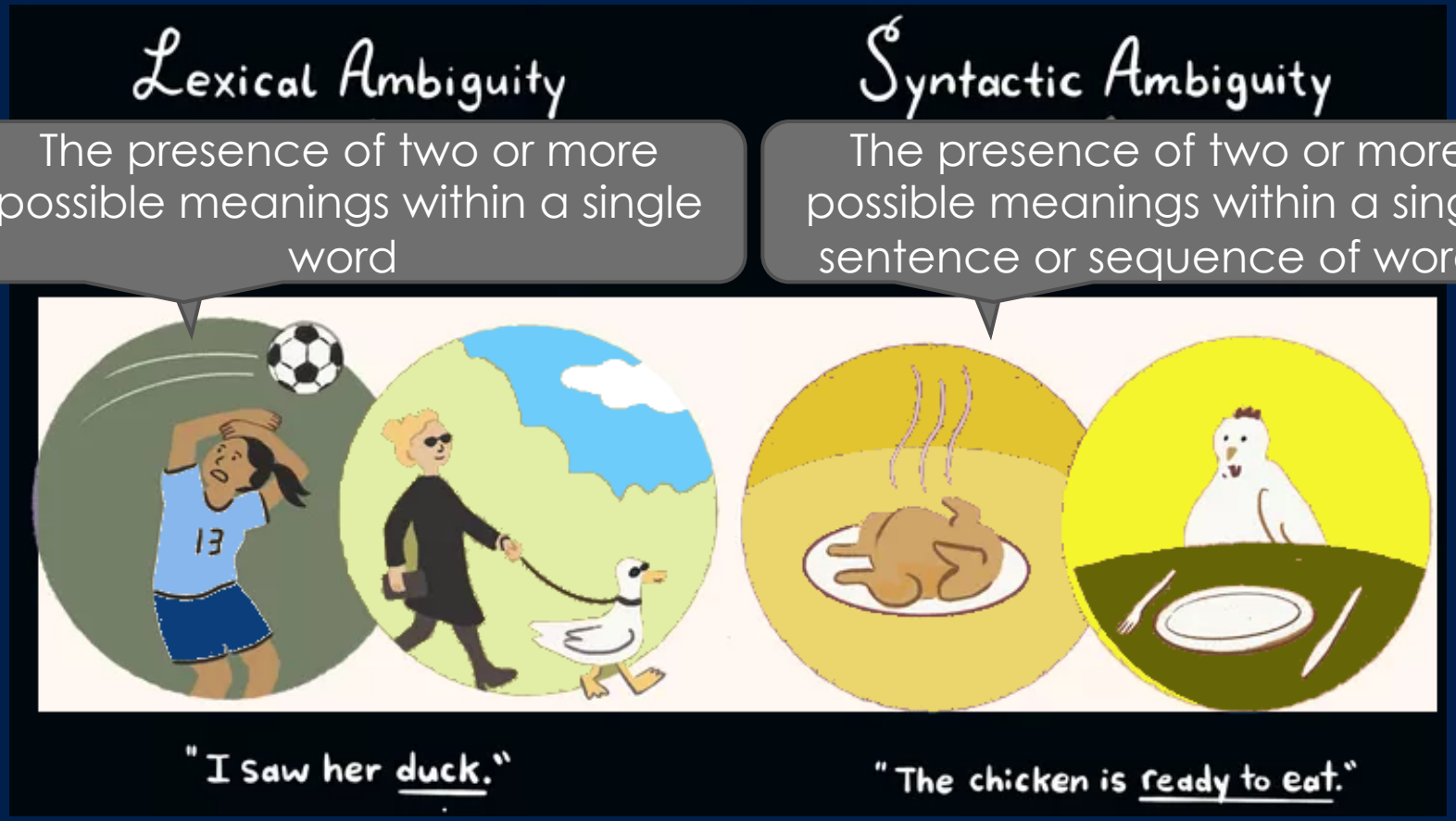
Why is NLP Hard?

Creative and Analytical Representation of ideas



WHY IS NLP HARD?

Ambiguous in nature



Source: Definition and Examples of Ambiguity in English(thoughtco.com)

WHY IS NLP HARD?

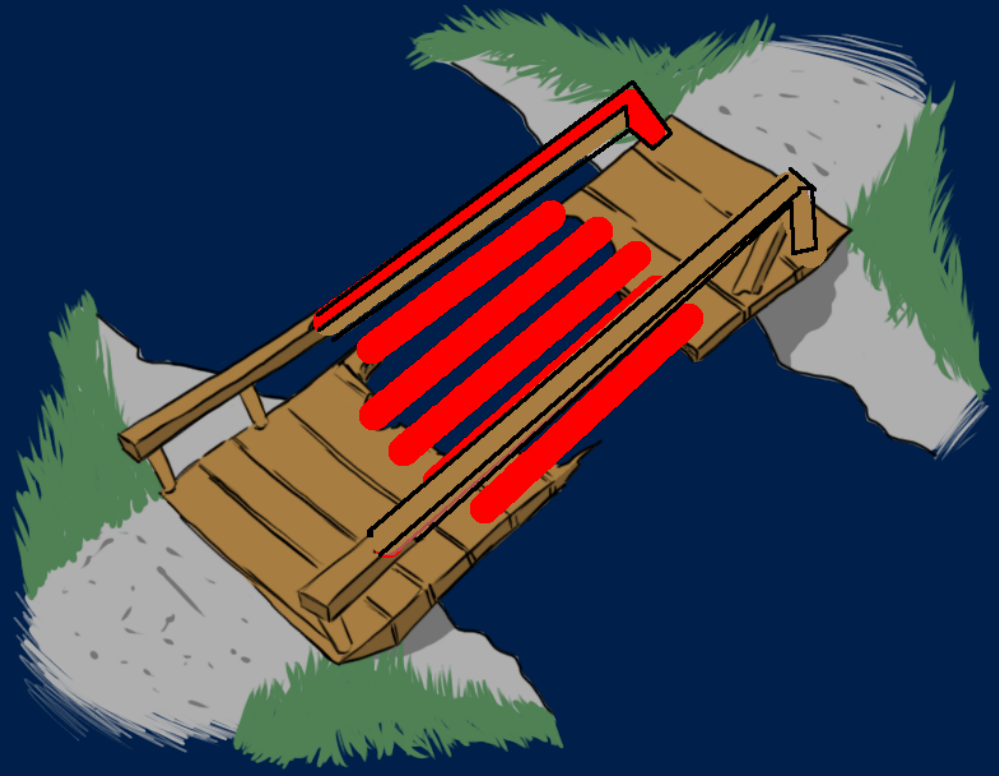
Ambiguous in nature

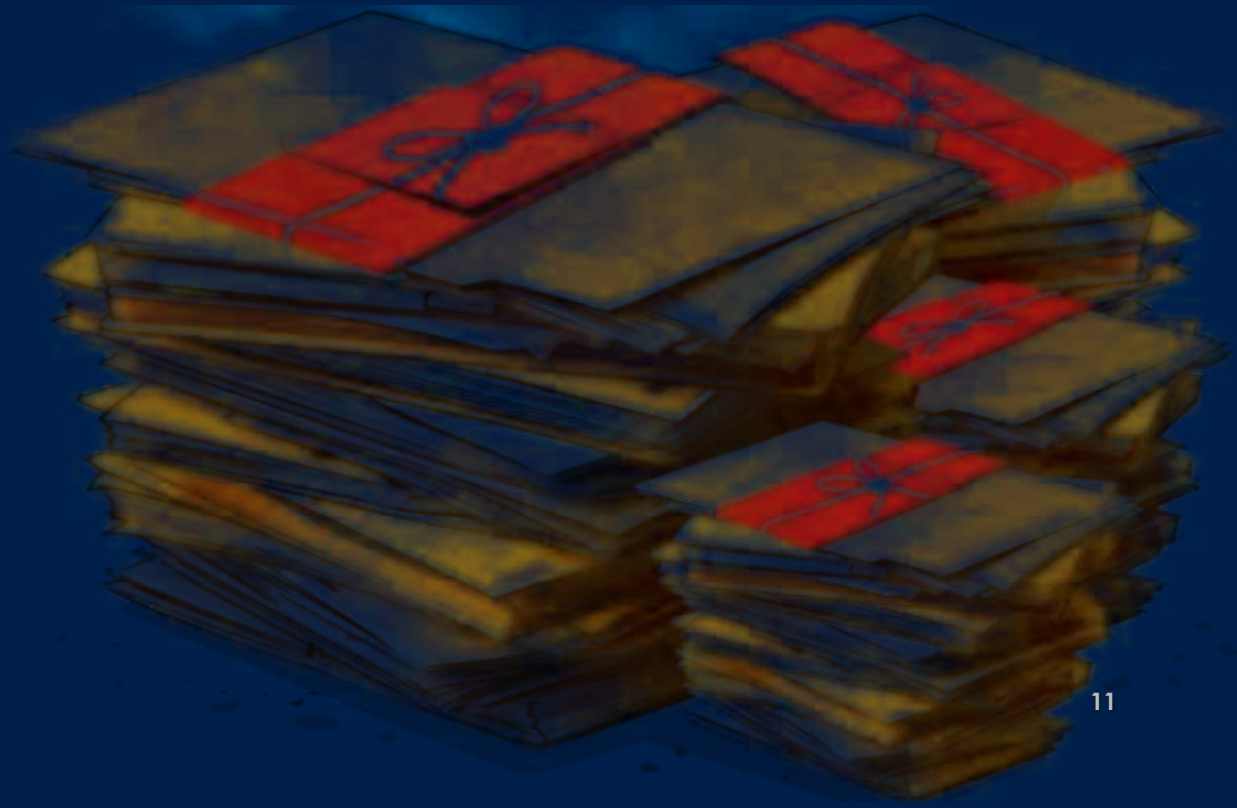
Image Source: SyntaxandApplied Lingulistics Modules - Posts | Facebook

Why is NLP Hard?

Ambiguous in nature

Red tape holds up new bridges





Why is NLP Hard?

Multiple representations of the same scenario

- The weather is extremely cold today
- It is freezing out there

WHY IS NLP HARD?

Includes common sense and contextual representation

- A set of words, phrases, paragraphs or a whole story to understand the meaning of the current situation/word

Why is NLP Hard?

Irony, sarcasm (Yeah! right), double negatives, etc. make it difficult for automatic processing

Me: I am 25 years old
Mirror: Yeah, right

Why is NLP Hard?

- Complex representation information (simple to hard vocabulary with uncommon usage of a sentence)

He was, in the way of most men, possessed of a rudimentary intelligence, his countenance ordinary, his bearing mild, with some weakness about the shoulders, his hair the color of ash; he spoke of the weather

The complex houses married and single soldiers and their families

Typical Tasks

Information Retrieval	Find documents based on keywords
Information Extraction	Identify and extract personal name, date, company name, city..
Language generation	Description based on a photograph Title for a photograph
Text clustering	Automatic grouping of documents
Text classification	Assigning predefined categorization to documents Identify Spam emails and move them to a Spam folder
Machine Translation	Translate any language Text to another (when one language is unknown)
Grammar checkers	Check the grammar for any language

Applications

- Sentiment Analysis
- Search Engines
- Content or News curation
- Automatic Machine Translation
- Transcription of Text from Audio/Video
- Chatbots
- …
- …

Goal

Ability to harness information from
a large corpus of text with
no human intervention

Ideal Properties of a Corpus

- Corpus is huge - Several billions of words
- Useful to verify a hypothesis about a language
- Find usage of a particular sound, word, or syntactic construction varies in different contexts
- Collection of most of the words of a language
- Even distribution of texts from all domains of language use

- *The boys play cricket on the river bank.*
- *The boys play cricket by the side of a nationalized bank*

Lexical Resources

- Brown Corpus contains a collection of written American English
- Sussane is a subset of Brown, but is freely available
- A bi-lingual parallel corpus, Canadian Hansards, contains French and English transcripts of the parliament
- Penn-Treebank contains annotated text from the Wall Street journal
- Most NLP software platforms such as NLTK, Spacy include several corpora for learning purposes
- HuggingFace and Kaggle - Several corpora text and image for machine learning applications
- Wiki dumps for various languages

Important Aspects of Any NLP Application

- Dependent
- Consistent
- Transparent

To achieve Human like performance, we require all the three aspects well defined and implemented

We cannot solely dependent on statistical analysis or machines learning

They are useful in making a good guess based on the historical information

We require theoretical insights to provide human-like performance

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

Operations on a corpus

- Text normalization
  - Converting text into a single canonical form - removal of foreign words, case folding, ...
- Tokenization
  - This is the process of dividing input text into tokens/words by identifying word boundary

Operations on a corpus

- Identification/Extraction
  - Process of identifying certain tokens, sentences and paragraphs
- Counting
  - The number of tokens/words in a corpus and its vocabulary count

Operations on a corpus

- Text normalization
  - Converting text into a single canonical form - removal of foreign words, case folding, ...
- Tokenization
  - This is the process of dividing input text into tokens/words by identifying word boundary

Operations on a corpus

- Identification/Extraction
  - Process of identifying certain tokens, sentences and paragraphs
- Counting
  - The number of tokens/words in a corpus and its vocabulary count

VOCABULARY

The set of unique words used in a corpus is referred to as the vocabulary

$$V = t_1, t_2, t_3, \ldots, t_n$$

Empirical Laws

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

Empirical Laws

Heaps

- Relates the terms and the vocabulary

Zipf

- Relates frequency of a word and its rank

Mandelbrot

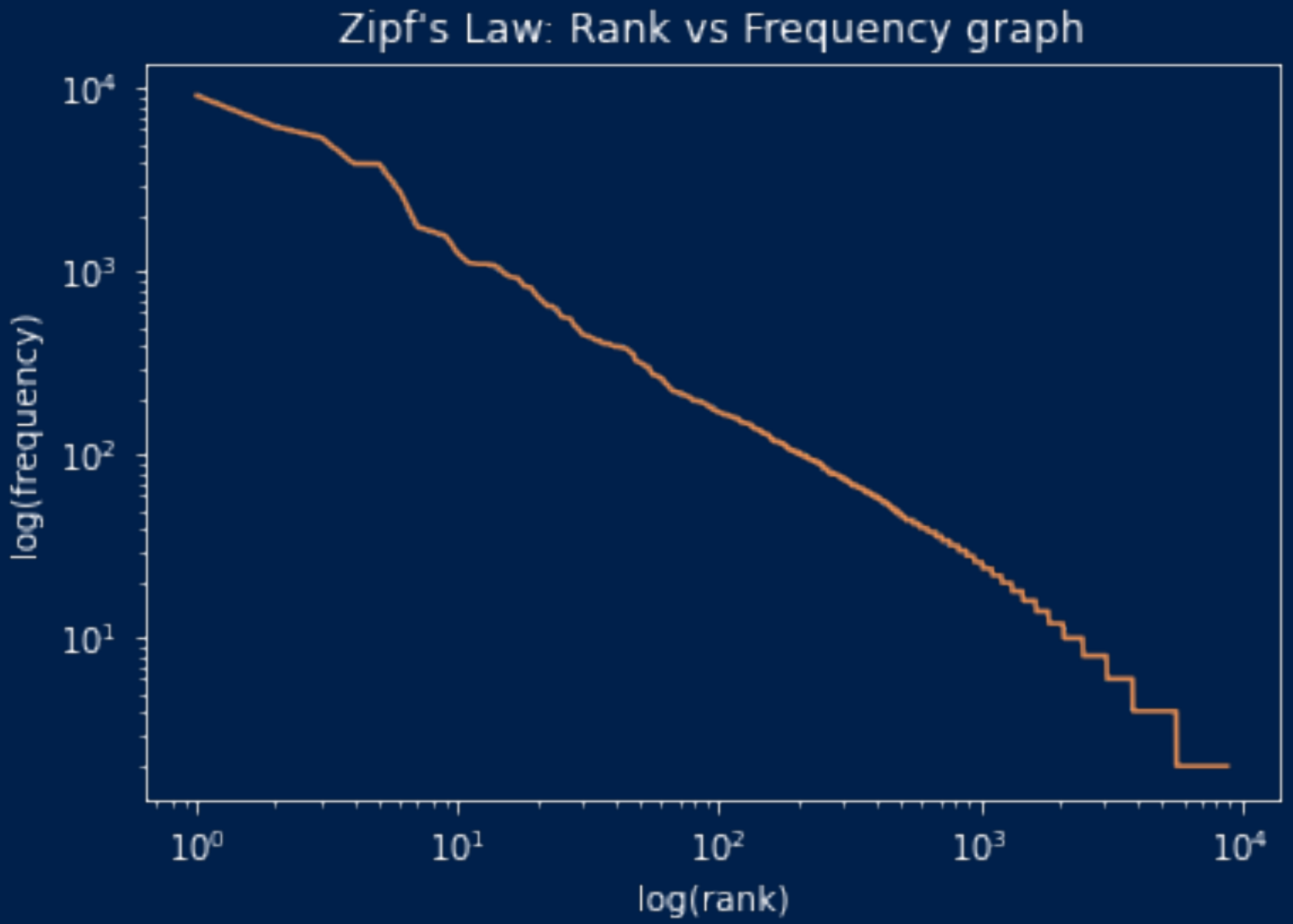
- A better approximation of Zipfs Law

Zipf's Law

The frequency of any word is inversely proportional to its rank

$$f \propto \frac{1}{r^{\alpha}}$$

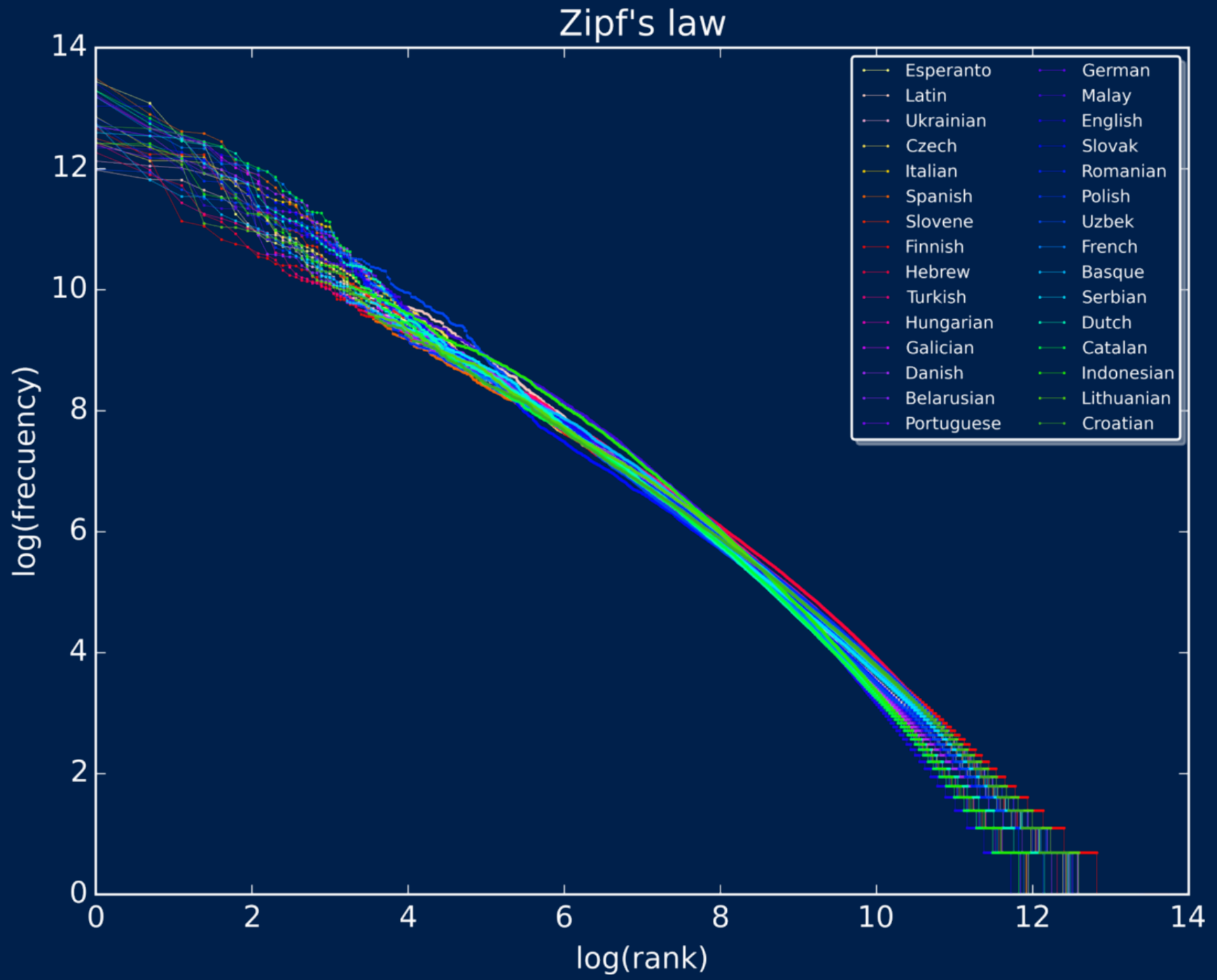
where $\alpha \approx 1$, r is the frequency rank of a word and f is the frequency of the word in the corpus.



$$f \propto \frac{1}{r^{\alpha}}$$

A plot of the rank versus frequency in a log-log scale.

DEMO



$$f \propto \frac{1}{r^{\alpha}}$$

A plot of the rank versus frequency for the first 10 million words in 30 Wikipedia (dumps from October 2015) in a log-log scale.

Source: Zipf's law -Wikipedia

DEMO

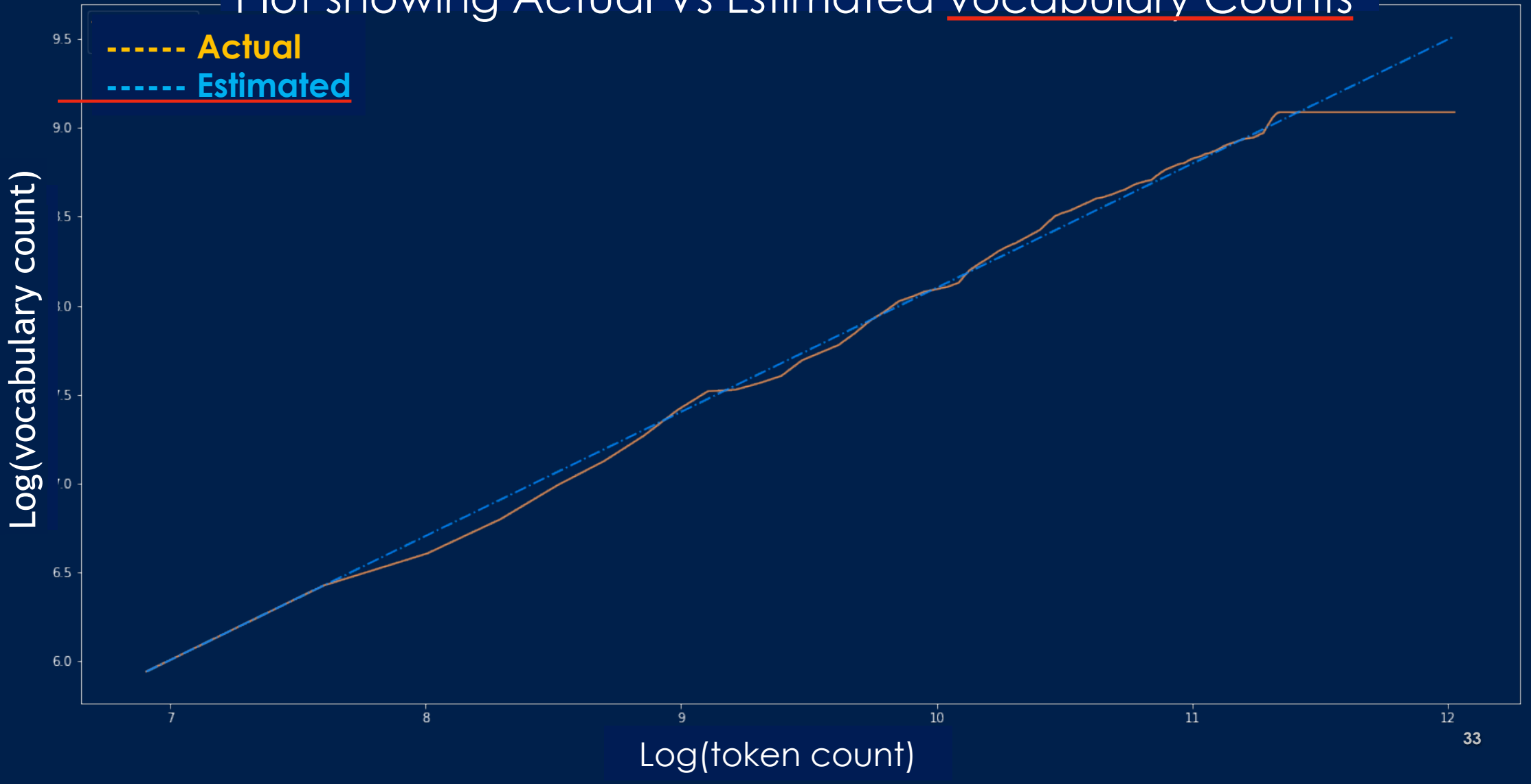
Mandelbrot's Law

The frequency of any word is inversely proportional to its rank. Mandelbrot derived a more generalized law to closely fit the frequency distribution in language by adding an offset to the rank

$$f \propto \frac{1}{(r+\beta)^{\alpha}}, \text{ where } \alpha \approx 1 \text{ and } \beta \approx 2.7$$

Heaps' Law

Plot showing Actual Vs Estimated vocabulary Counts



Word as Atomic unit

- How do we represent the words?
- How do you present it as input to the machine?
- Can the word be used as the atomic unit?

TERMS AS ATOMIC UNITS

- Term (co-located/co-occurring words) are also used as atomic units
  - I went to the post office yesterday
  - Employment is a major problem in most of the countries
  - I went to the airport to see him off

REPRESENTATION OF WORDS

How do we represent the word?

$$\text{Real} - 1.31$$

$$\text{Complex} - 13.1+17.2i$$

$$\text{One-Hot Vector} - |0\ 0\ 0\ 0\ 1|$$

$$\text{Sparse Vector} - |0.2\ 0\ 0\ 0\ 1.3|$$

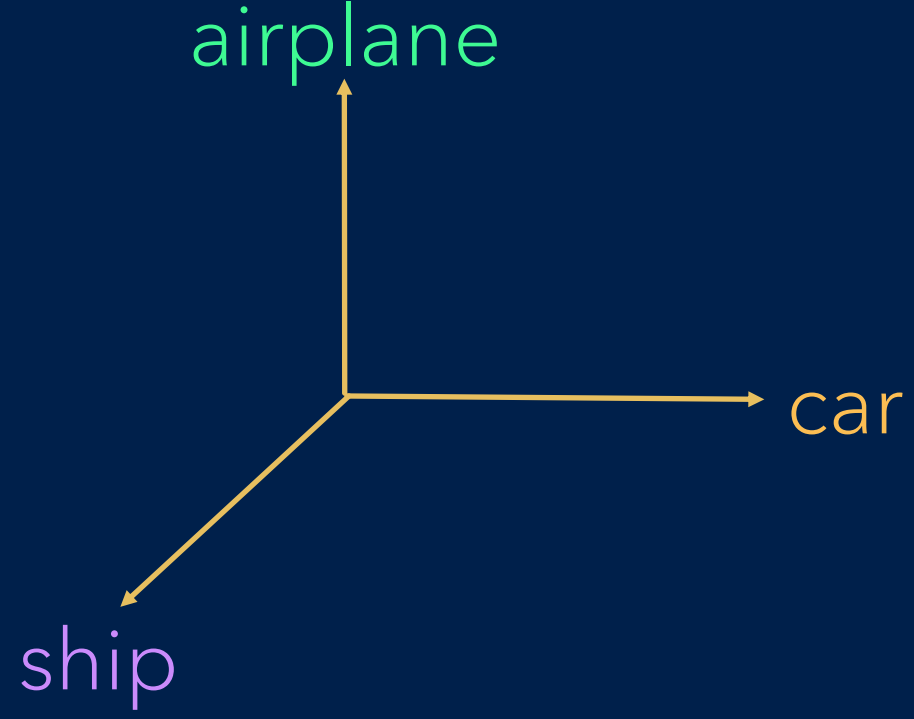
$$\text{DenseVector} - |\text{-}1.43\ \ 2.18\ \ 0.02\ \ \text{-}0.84\ \ 0.123|$$

Disadvantages of Term Frequency

- All terms are given equal importance
- The Common term the has no relevance to the document, but gets high relevancy score
- Repetition of frequency values possible
  - Unique representation not possible

VECTOR REPRESENTATION OF WORDS

Let us assume that the vocabulary of a corpus are considered as linearly independent basis vectors.



The vocabulary size of Google News Corpus is 3 million words. If we plot all the words in the real space R, we get 3 million axes

Co-occurrence Matrix

	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_{10}
w_1	0.1	0.0	0.4	0.1	0.2	0.0	0.1	0.9	0.9	0.3
w_2	0.1	0.0	0.4	0.1	0.2	0.0	0.1	0.9	0.9	0.3
w_3	0.0	0.9	0.0	0.2	0.3	0.1	0.7	0.0	0.2	0.7
w_4	0.0	0.9	0.3	0.9	0.5	0.1	0.9	0.3	0.8	0.4
w_5	0.4	0.0	0.3	0.2	0.5	0.9	0.3	0.7	0.4	0.6
w_6	0.6	0.0	0.4	0.7	0.3	0.3	0.9	0.1	0.9	0.0
w_7	0.0	0.8	0.5	0.6	0.6	0.6	0.0	0.1	0.4	0.9
w_8	0.4	0.0	0.6	0.5	0.5	0.1	0.7	0.1	0.5	0.3
w_9	0.3	0.0	0.7	0.9	0.8	0.7	0.7	0.8	0.6	0.6
w_{10}	0.0	0.5	0.5	0.0	0.2	0.0	0.0	0.1	0.3	0.5

Co-occurrence matrix built using the count of words w_i, w_j where $i,j = 1..10$

Similarity Measure

A similarity measure is a real-valued function that quantifies the similarity between two vectors. Some of the methods are given below.

$$\textbf{Cosine Similarity} = \frac{\vec{w}_1 \cdot \vec{w}_2}{\|\vec{w}_1\| \, \|\vec{w}_2\|} = \frac{\vec{w}_1}{\|\vec{w}_1\|} \cdot \frac{\vec{w}_2}{\|\vec{w}_2\|}$$

$$\textbf{Cosine Distance} = 1 - \frac{\vec{w}_1 \cdot \vec{w}_2}{\|\vec{w}_1\| \, \|\vec{w}_2\|}$$

$$= 1 - \frac{\vec{w}_1}{\|\vec{w}_1\|} \cdot \frac{\vec{w}_2}{\|\vec{w}_2\|}$$

Similarity Measure – Using Colors

$color1 = [255\ 255\ 255]$

$color2 = [255\ 0\ 0]$

$color3 = [255\ 144\ 0]$

$color4 = [255\ 164\ 0]$

$color5 = [0\ 0\ 255]$

$color6 = [0\ 255\ 0]$

Cosine Distance

$\mathrm{CD}(color1, color1) = 0$

$\mathrm{CD}(color1, color4) = 0.2021$

$\mathrm{CD}(color1, color2) = 0.4226$

$\mathrm{CD}(color1, color3) = 0.2134$

$\mathrm{CD}(color3, color4) = 0.0017$

$\mathrm{CD}(color2, color5) = 1.0$

$\mathrm{CD}(color2, color6) = 1.0$

Disadvantages of OHV representation

- The number of elements increases with vocabulary size
- Each vector is orthogonal to every other vector in the vocabulary
- Difficult to represent synonyms and similar words

Disadvantages of Sparse Vector representation

- The number of elements in the vector increases with vocabulary size
- The memory footprint of Co-occurrence or word-word matrix increases