

Natural Language Processing

WORKSHOP

24-28 August, 2022

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?

UNSTRUCTURED CONTENT

3

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

4

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

5

Intended for Human
Machine Communications

Instructions are

Precise

Unambiguous

Mathematical equations



GOAL


6

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

IDEAL PROPERTIES OF A CORPUS

7

- 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*

IS NLP HARD?

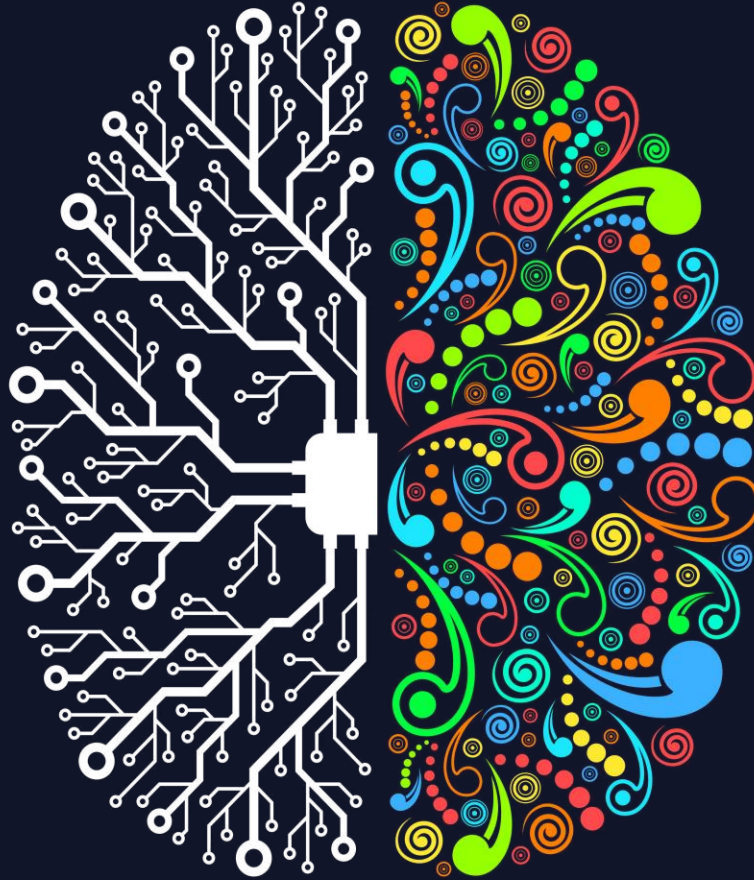
8

- 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?

9

Creative and Analytical Representation of ideas



WHY IS NLP HARD?

10

Ambiguous in nature

Lexical Ambiguity

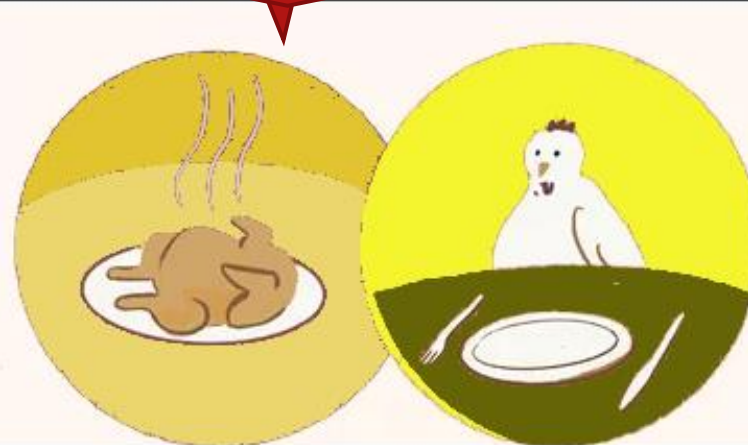
The presence of two or more possible meanings within a single word



"I saw her duck."

Syntactic Ambiguity

The presence of two or more possible meanings within a single sentence or sequence of words



"The chicken is ready to eat."

Source: [Definition and Examples of Ambiguity in English\(thoughtco.com\)](http://Definition and Examples of Ambiguity in English(thoughtco.com))

WHY IS NLP HARD?

11

Ambiguous in nature



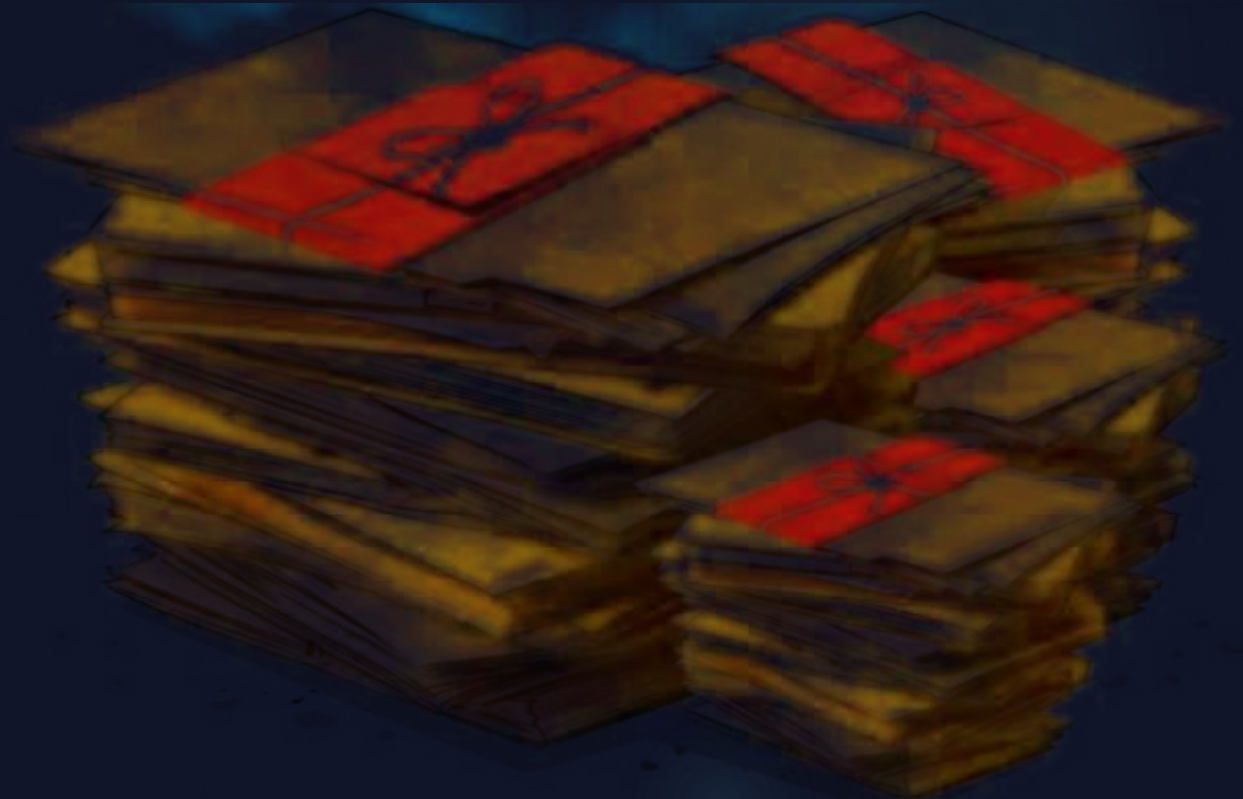
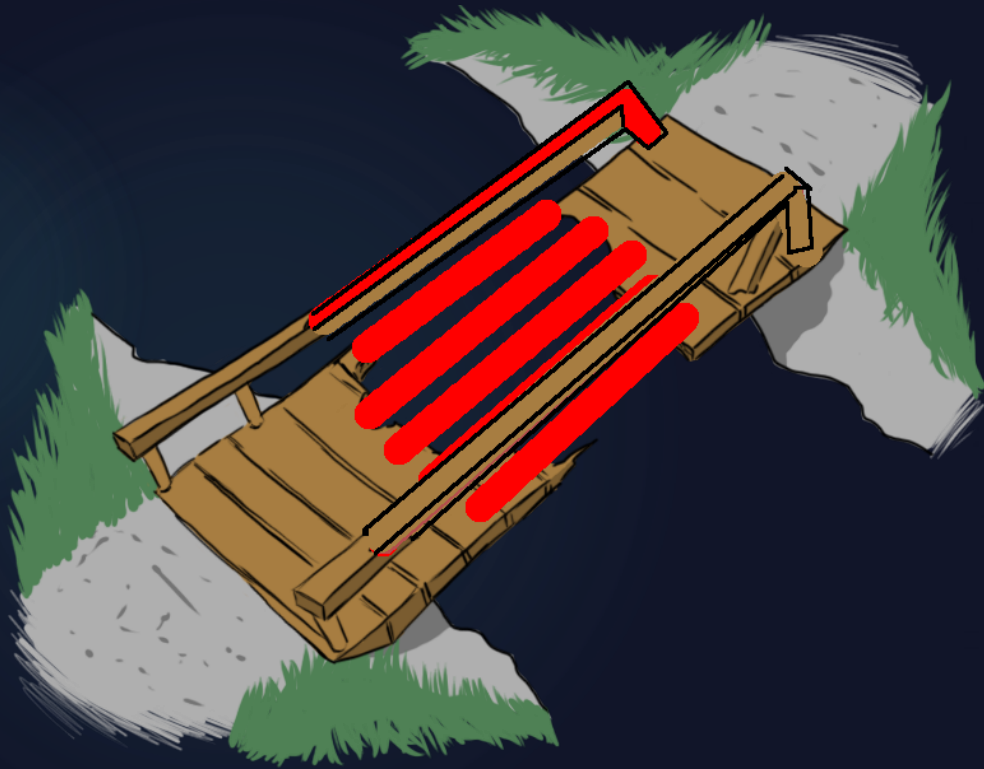
[Image Source: SyntaxandApplied Linguistics Modules - Posts | Facebook](#)

WHY IS NLP HARD?

12

Ambiguous in nature

Red tape holds up new bridges



WHY IS NLP HARD?

13

Multiple representations of the same scenario

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



WHY IS NLP HARD?

14

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

Common sense is increasingly uncommon.



WHY IS NLP HARD?

15

- 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?

16

- 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

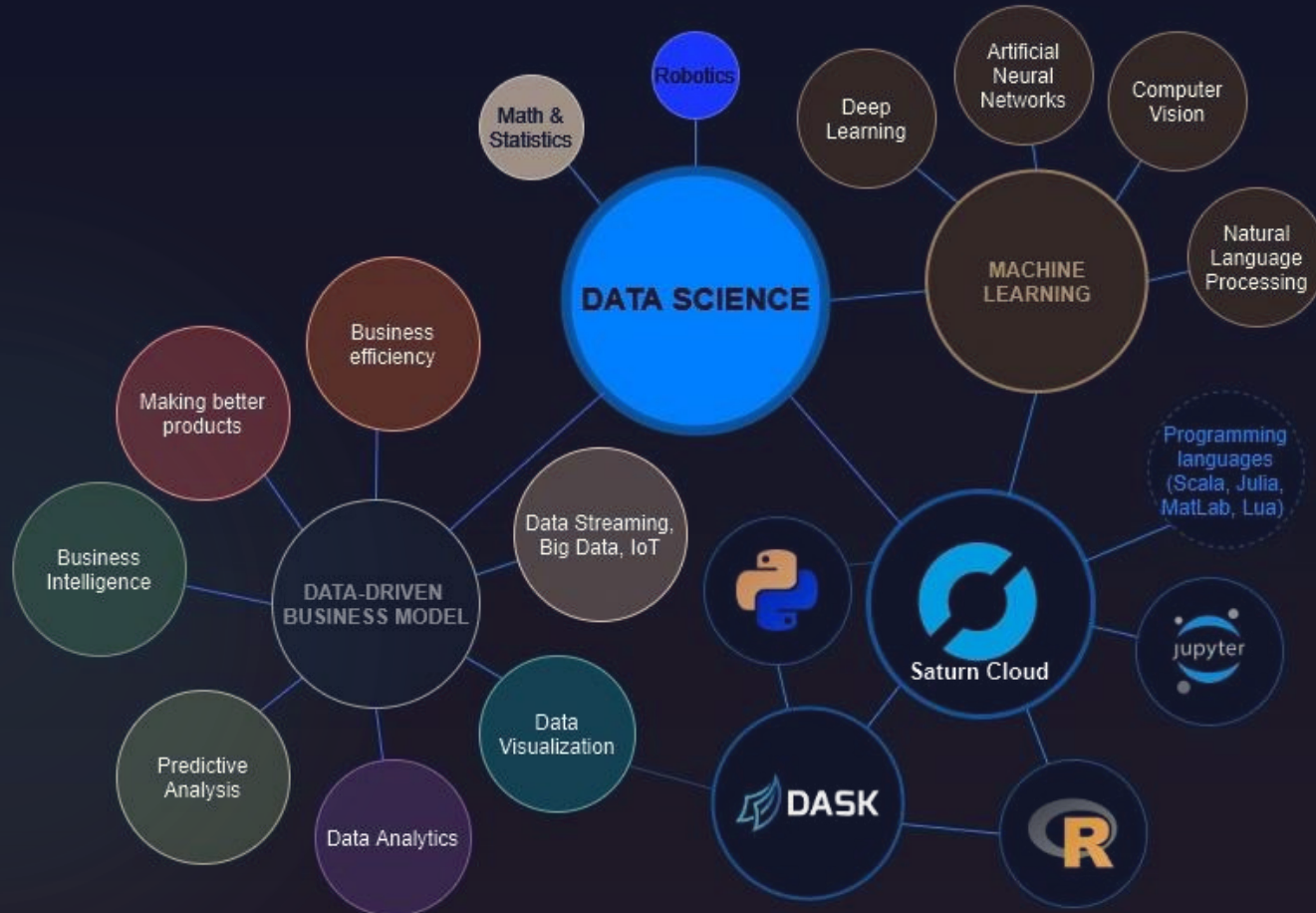
The complex houses married and single soldiers and their families



DOMAIN AND CONTEXTUAL UNDERSTANDING

17

Data Science Context



DISAMBIGUATION

18

Synset ('bank.n.01') sloping land (especially the slope beside a body of water)

Synset ('depository-financial-institution.n.01') a financial institution that accepts deposits and channels the money into lending activities

Synset ('bank.n.10') a flight maneuver; aircraft tips laterally about its longitudinal axis (especially in turning)

Synset ('trust.v.01') have confidence or faith in

TYPICAL TASKS

19

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

20

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

LEXICAL RESOURCES

21

- ❖ [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

OPERATIONS ON A CORPUS

22

❖ 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

❖ Identification/Extraction

- ❖ Process of identifying certain tokens, sentences and paragraphs

❖ Counting

- The number of tokens/words in a corpus and its vocabulary count

WORD AS ATOMIC UNIT

23

- ❖ 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

24

- ❖ 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

25

How do we represent the word in a numerical forms?

- Binary, real, complex or in a vector form?

VOCABULARY

26

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

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

TERM FREQUENCY

27

Term frequency t_f in a corpus is defined as the number of occurrences of a term

$$\text{Raw count} = t_f$$

$$\text{Corpus length} = \frac{t_f}{n}$$

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

Empirical Laws

Heap's Law

- Relates the terms and the vocabulary

Zipf's Law

- Relates frequency of a word and its rank

Mandelbrot's Law

- A better approximation of Zipf's Law

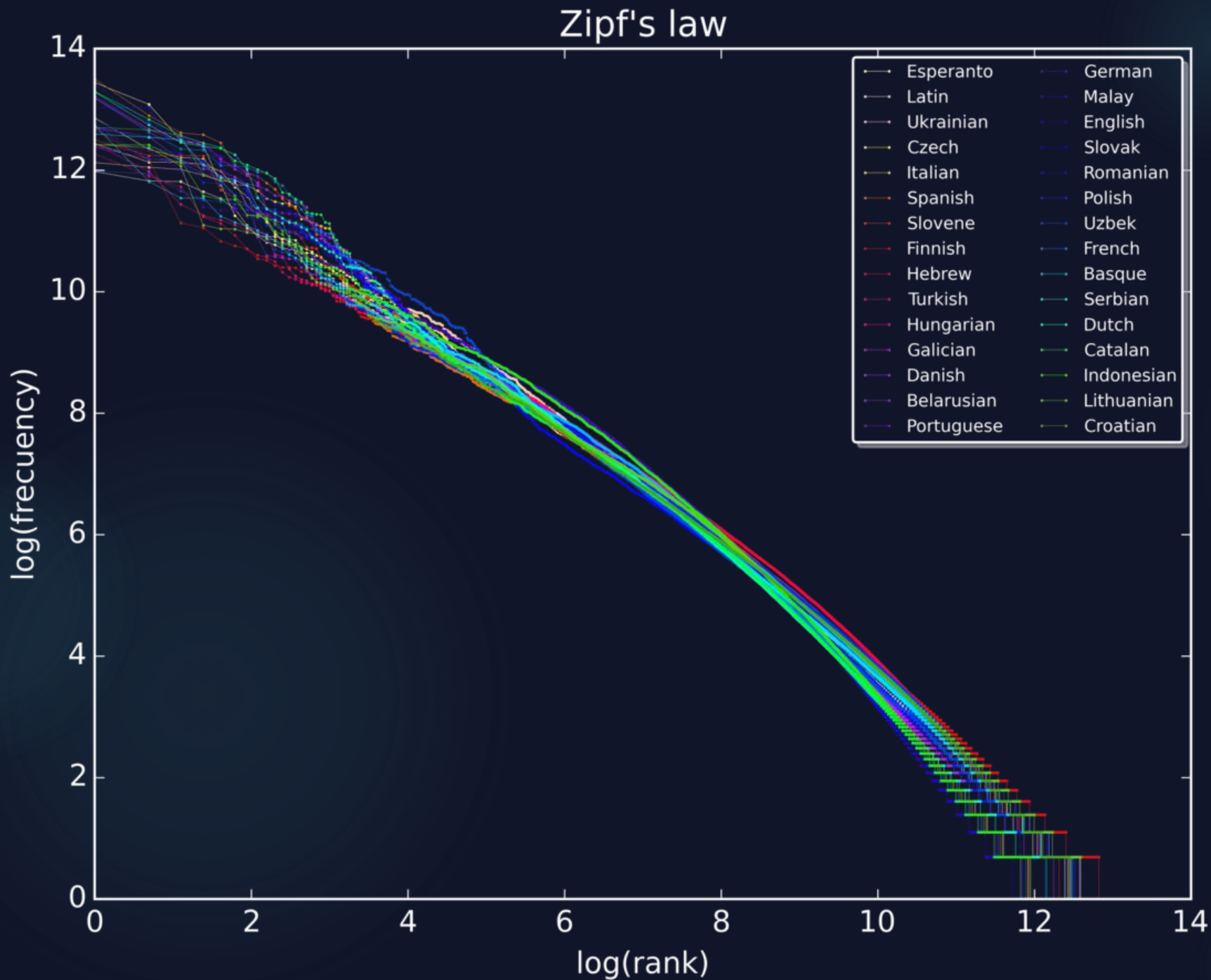
ZIPF'S LAW

29

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 for the first 10 million words in 30 Wikipedia (dumps from October 2015) in a log-log scale.

Source: [Zipf's law -Wikipedia](#)

MANDELBROT'S LAW

31

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

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

END OF SESSION – DAY 1

32

Thank you

Ramaseshan.nlp@gmail.com