

SESSION

Text Classification

S P E A K E R

Raghav Bali
Staff Data Scientist
DeliveryHero, Berlin



Raghav Bali



Staff Data
Scientist



Author of
Multiple Books



Inventor with
7+ patents



Speaker at
Top Events

Agenda

- What is NLP
- Applications of NLP
- Complexities
- NLP Workflow
- Context
- Deep Learning and NLP

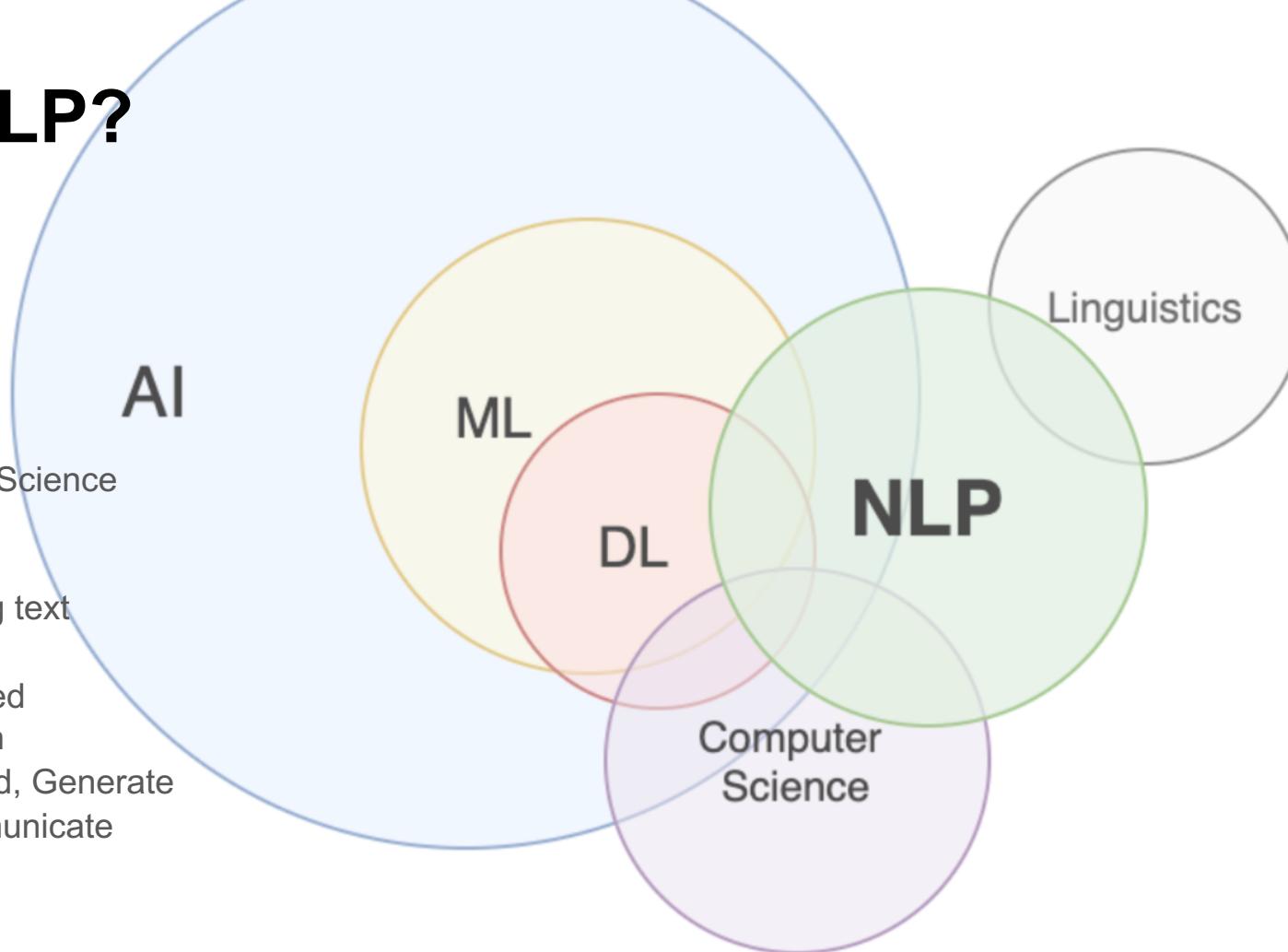


Natural Language

- Highly unstructured in Nature
- Way of Communication
- Too many variations
- Difficult to Ingest/Parse
- Available all around us

What is NLP?

- Intersection of :
 - Linguistics
 - Computer Science
 - AI
- NLP is about:
 - Processing text
 - Analyzing unstructured information
 - Understand, Generate and Communicate



Applications

1 Text Classification

Sentiment Analysis, Spam Detection, Document Categorization

2 Information Retrieval

Search, Chat-bots, Question Answering

3 Text Summarization

News Excerpts, Document/Book Summary, Topic Modeling

4 Machine Translation

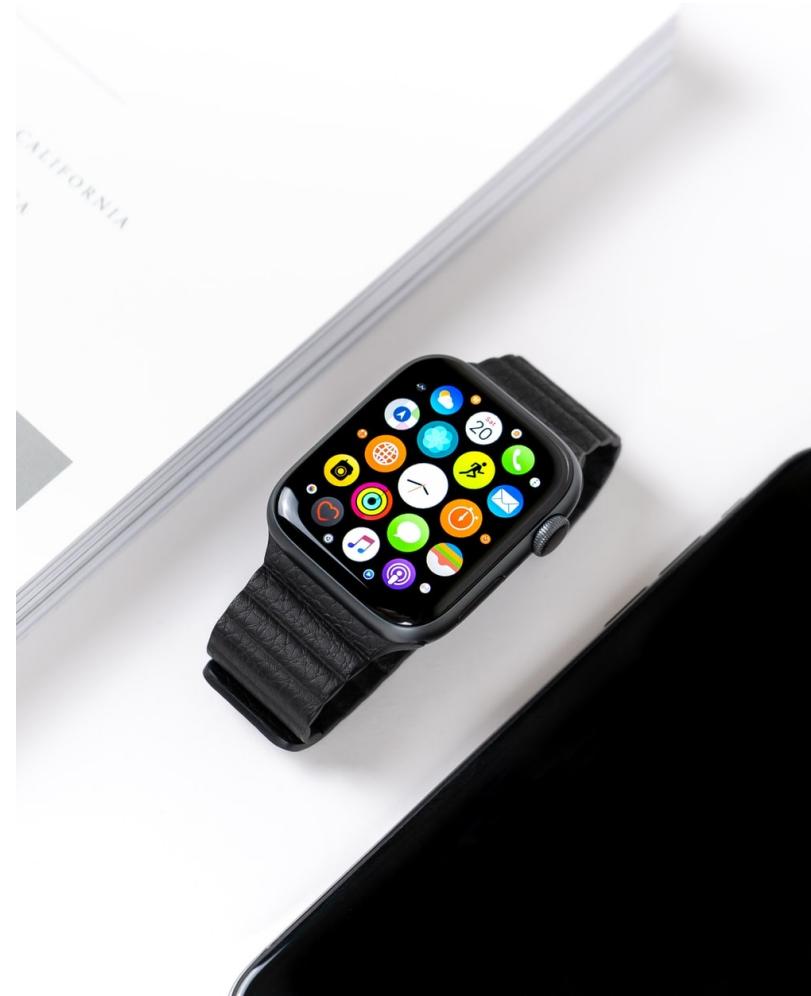
English to Hindi, German to French...

5 Text Generation

Creative writing, Headline creation, Meme text

6 Speech Analysis

AI Assistants, Speech to Text



Language and its Complexities

Every language is defined with a list of characters called the alphabet, a list of words called the vocabulary and a set of rules to use the words called grammar.

Languages are complex and have fuzzy grammatical rules and structures.

Most learning algorithms are designed to work with numbers, matrices, vectors...

Typical NLP Workflow



Preprocessing

Cleanup
Standardization



Feature Engineering

Tokenization
Text Representation

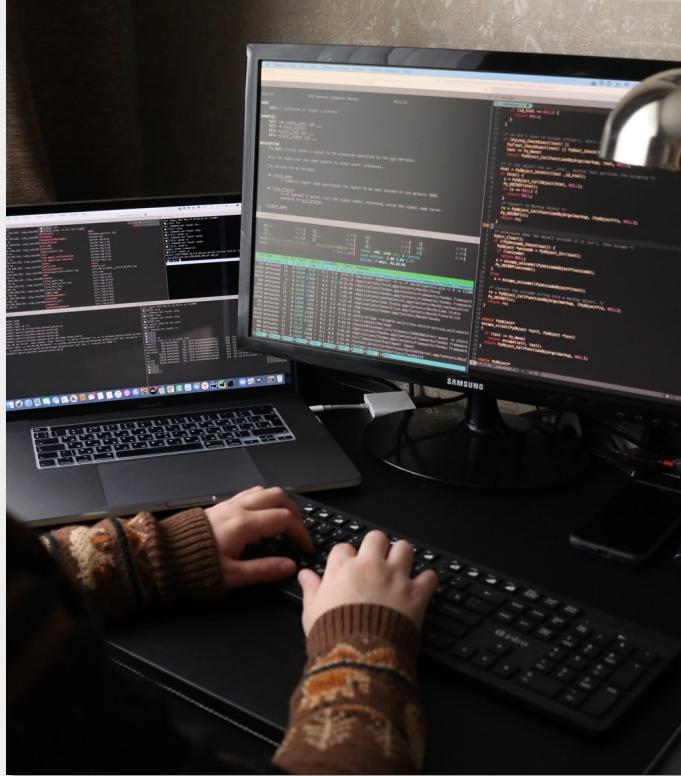


Modeling

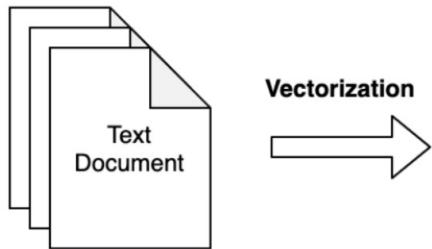
Train
Evaluate and Deploy

Time for Some Action

Hands-on - I



Text Representation Models



- Textual data is not directly ingestible
- We need methods to transform text into useful format
- Text Representation Models help in “vectorizing” textual data

Text Representation Models

Bag of Words

- Matrix Transformation of textual data
- Ordering of words is ignored
- Matrix:
 - Columns are words of the vocabulary
 - Rows are documents/sentences
 - Cell values are count of occurrences

TF-IDF

- Improvement on top of BoW
- Counts are normalized based on inverse document frequency
- More robust to noise

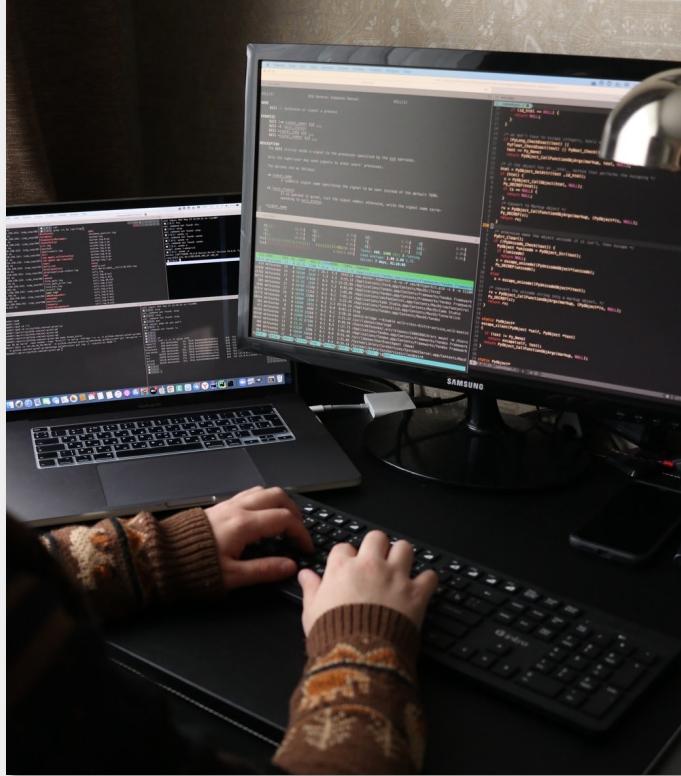
Co-Occurrence Matrices

- Matrix:
 - Column are words of the vocabulary
 - Rows are words of the sentence
 - Cell values are number of times a word in the given row exists in the context of the words in the columns
- Good at capturing word associations

	This	Is	Sentence	One	Two
This is sentence one	1	1	1	1	0
This is sentence two	1	1	1	0	1

Time for Some Action

Hands-on – II and III



Context

- Did you see the look on her **face**?
- The new clock-**face** on my watch is amazing
- It is time to **face** my demons
- How many new **faces** did you see if office today?

Non-Local Interactions

The man who ate pepper sneezed



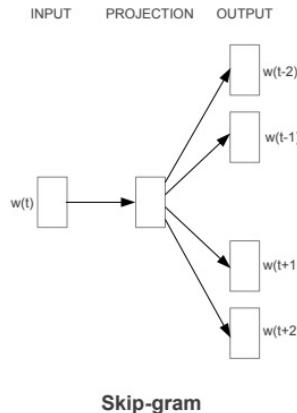
The cat who bit the dog barked



Word Embeddings

Word2Vec

- Vector representation based on **local** context
- Variations:
 - CBOW
 - Skip-Gram
- Dense representations



GloVe

- Based on **Global** and Local contexts
- Dense representation with performance on par with Word2Vec

FastText

- Extension of Word2Vec setup
- Handles **out of vocabulary** terms better than Word2Vec and GloVe

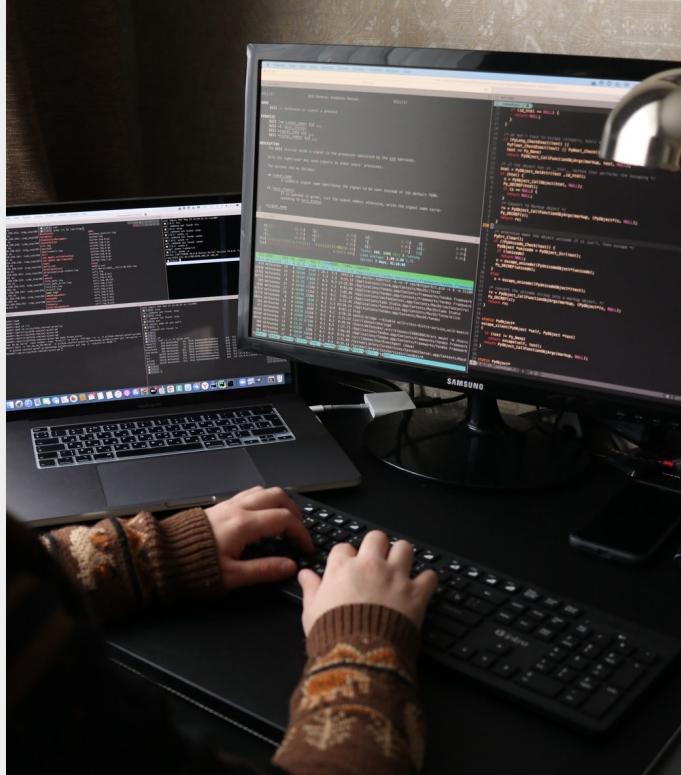
Deep Learning and NLP

- *Attention is All You Need*
- Transformers
- BERT and Friends



Time for Some Action

Hands-on - IV



End

