Getting Started with NLP

Raghav Bali

Agenda

Introduction

What is NLP

Why do we need NLP

Applications

Language & Its Complexities

NLP Workflows

Hands-on

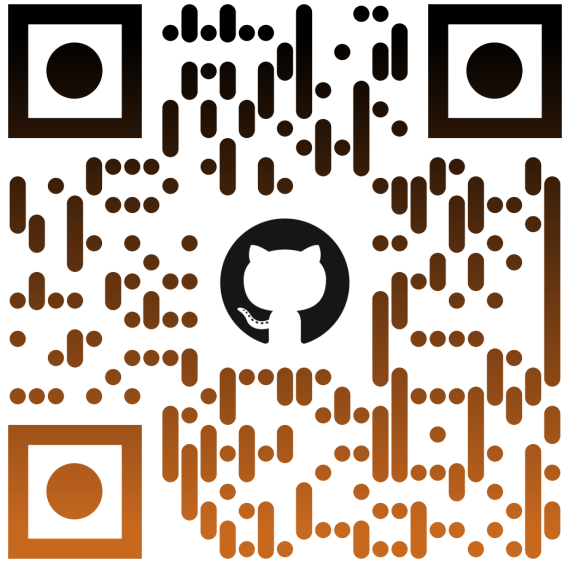
NLP & Deep Learning

Raghav Bali

- Senior Data Scientist @ Optum (UHG)
- Author of more than 5 books on ML/DL
- Springboard Mentor
- Speaker

Slides and Code

https://github.com/raghavbali/nlp_starterpack_webinar

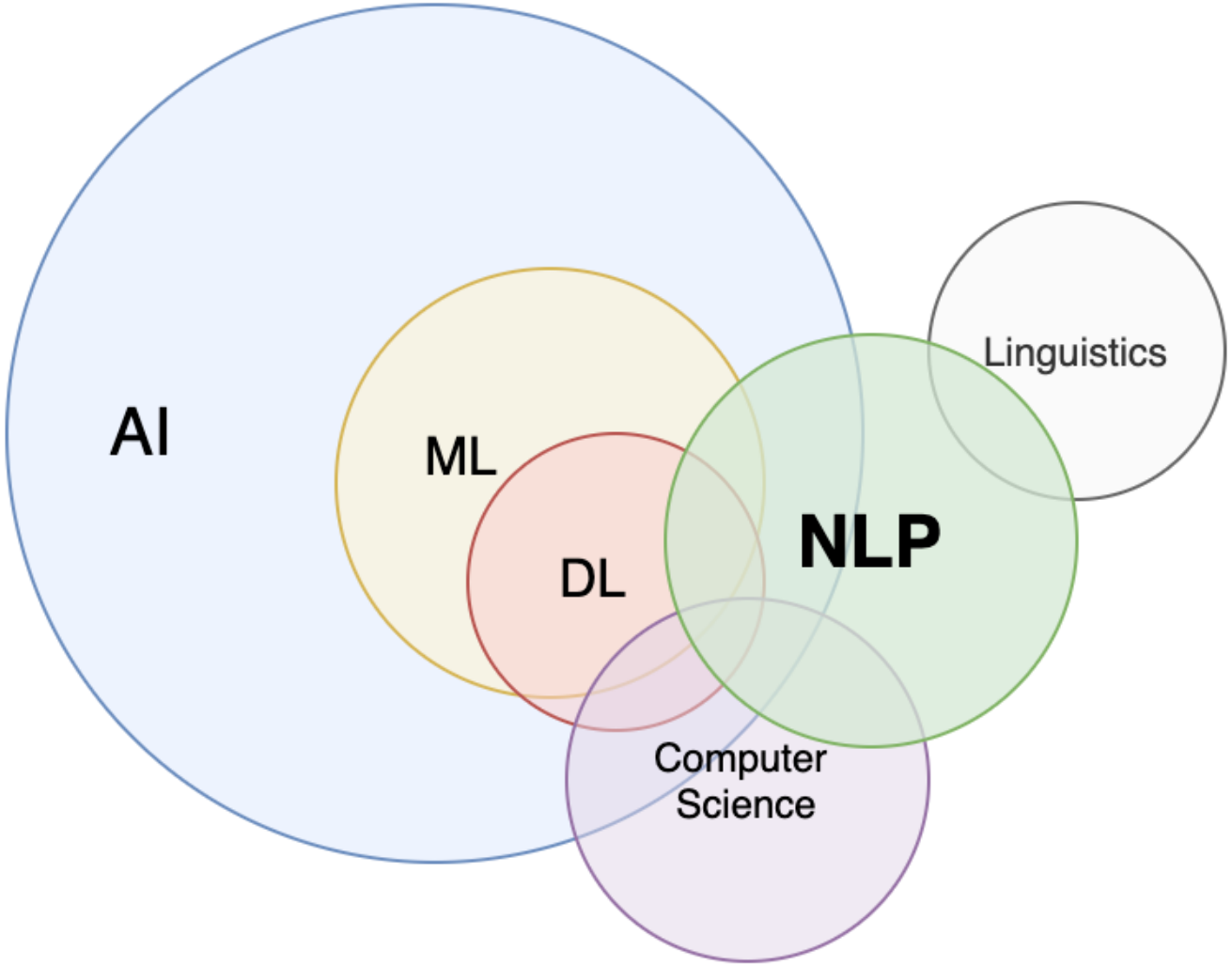


Natural Language

- Way of Communication
- Highly unstructured in nature
- Many Variations
- Difficult to parse
- Available as:
  - Text/Chat Messages
  - Documents
  - Emails and many more…

What is NLP

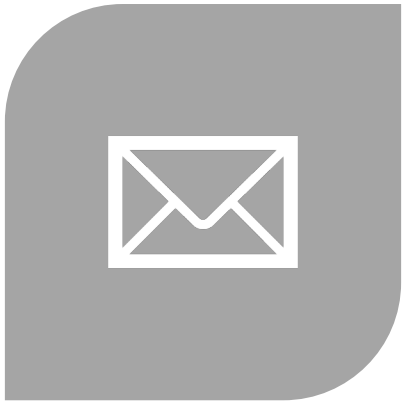
- Intersection of
  - Linguistics
  - Computer Science
  - AI
- Train Algorithms to :
  - Process
  - Analyze
  - Understand
  - Model large amounts of natural language data



Why do we need NLP

Text Is Everywhere

Information/data stored in relational datasets is easy to mine and analyze. Yet bulk of enterprise data is in unstructured form.



Real-world information flow is the form of conversations, emails, documents, forms of various types, log files and so on

Images also contain text

Applications

- **Machine Translation**
  English to French, German to Hindi…
- **Information Retrieval**
  Search, Question-Answering, Chat-bots…
- **Text Classification**
  Sentiment Analysis, Document Categorization…
- **Text Summarization**
  Topic Modeling, Book Summary, News excerpts…
- **Text Generation**
  Fake News, Creative writing, Meme Text, Captions…
- **Speech Analysis**
  Speech to Text, AI Assistants…

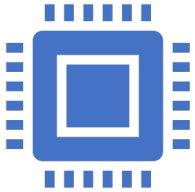
Language and Its Complexities

- Language is one of the most complex aspects of our existence. We use language to communicate our thoughts and choices
- Every language is defined with a list of characters called the alphabet, a vocabulary and a set of rules called grammar
- Languages are complex and have fuzzy grammatical rules and structures.
- Machine learning and deep learning algorithms in general work with numbers, matrices, vectors and so on

Language and Its Complexities

How can we represent text for different language related tasks?

NLP Workflow



Pre-processing

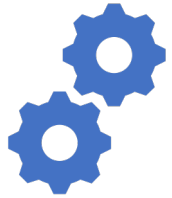
Clean-up

Standardization

Feature Engineering

Tokenization

Text Representation



Modeling

Train Models

Evaluate & Deploy

Preprocessing

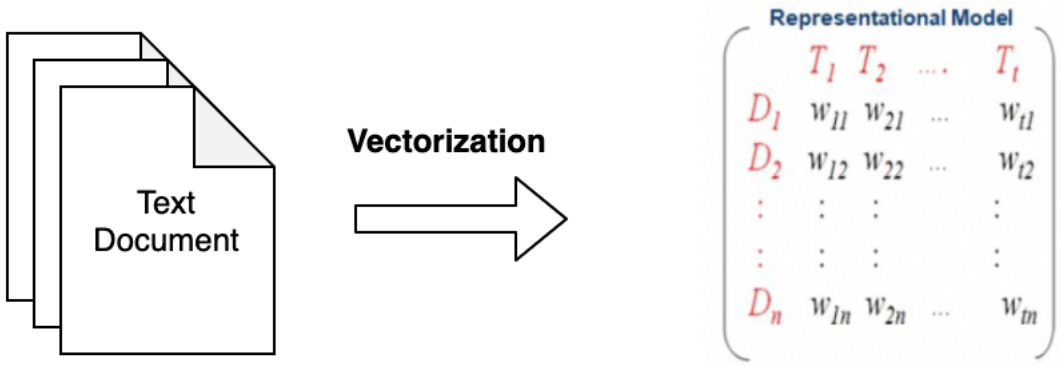
Remove unwanted symbols, special characters, newlines & whitespaces

Grammar and spell checks

Stemming & Lemmatization

Stop-word Removal

Text Representation Models



- ML/DL Models work with vectors
- Text Representation Models help in transforming textual data into vectors.
- Vector representation varies based on representation model used

Text Representation Models

Bag of Words

- Matrix with:
  - Each Word is dimension/column
  - Each document is a row
- Cell values are counts of occurrence

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

TF-IDF

- Matrix is same as BoW
- Counts are normalized with inverse document frequency

Co-Occurrence Matrices

- Matrix with:
  - Vocabulary as columns
  - Each word in the sentence as a row
- Each cell represents number of times word in given row exists in context of words in columns
- Captures word associations

Quick Hands-on

NLP and Deep Learning

- Words are **not** discrete symbols.
  Context is important
- Interactions **may-not** be local always

Context is Important

- Did you see the look on her **face** ?
- We could see the clock **face** from below
- It could be time to **face** his demons
- There are a few new **faces** in the office today

Adapted From: UCLxDeepMind-DL2020-L7

Non-Local Interactions

The man who ate pepper sneezed

The cat who bit the dog barked

Word Embeddings

Word2Vec

- Vector representations based on **local** context
- Two methods:
  - CBOW
  - Skip-gram
- Dense Representation

GloVe

- Vector representations based on **global** and local contexts
- On-par performance as Word2Vec

FastText

- Transforms every word into a **set of n-grams**
- Extends on Word2Vec setup
- Vector representation of each n-gram
- Handles **out of vocabulary** terms better than Word2Vec and GloVe

NLP and Deep Learning

- Attention is all you need
- Transformers
- BERT and its Friends

Hugging Face Transformers

Quick Hands-on

Let Us Stay Connected

github.com/raghavbali

linkedin.com/in/baliraghav

raghavbali.github.io

instagram.com/raghavbali

amazon.in/Raghav-Bali