NATURAL LANGUAGE PROCESSING
22AIM53

Course outcomes: At the end of the course, the student will be able to:

22AIM53.1 Understand basics of linguistics, probability and statistics associated with NLP.

22AIM53.2 Analyze the semantic of natural language.

22AIM53.3 Design an end-to-end NLP application by integrating preprocessing, feature extraction, and model-building techniques.

22AIM53.4 Evaluate the performance of advanced transformer models (e.g., BERT, GPT-3) in various NLP tasks such as text classification, summarization, and topic modeling.

22AIM53.5 Demonstrate the working of sequence models for text processing.

22AIM53.6 Implement the NLP applications on emerging trends with ethical implications.

MODULE-1	**Natural Language Processing**	22AIM53.1	**8 Hours**
Components - Basics of Linguistics and Probability and Statistics – Words-Tokenization-Morphology: Inflectional Morphology - Derivational Morphology. Finite-State Morphological Parsing - Porter Stemmer.			
Case Study	Case studies of NLP applications in various industries.		
Text Book	Text Book 1: Ch 2,3,4		
MODULE-2	**Semantic Analysis**	**2**2AIM53.2	**8 Hours**
Representing Meaning-Meaning Structure of Language-First Order Predicate Calculus Representing Linguistically Relevant Concepts -Syntax-Driven Semantic Analysis - Semantic Attachments -Syntax-Driven Analyzer. Robust Analysis - Lexemes and Their Senses - Internal Structure - Word Sense Disambiguation -Information Retrieval			
Text Book	Text Book 1: 13,14,18		
MODULE-3	**WORD REPRESENTATION AND PART OF SPEECH**	22AIM53.2, 22AIM53.3	**8 Hours**
N-grams and Language models –Smoothing- Evaluating Language Model -Text classification- Naïve Bayes classifier –- Vector Semantics – TF-IDF – Word Embeddings: Word2Vec, Glove and Fast Text-Part of Speech – Part of Speech Tagging -Named Entities –Named Entity Tagging-Conditional Random Fields(CRFs).			
Text Book	Text Book 1: Ch 4,5,10,17,19		
MODULE-4	**Transformer and Topic Models**	22AIM53.4, 22AIM53.5	**8 Hours**
Introduction to transformer architecture-BERT (Bidirectional Encoder Representations from Transformers)-GPT-3 (Generative Pre-trained Transformer 3)-Fine-tuning transformer models for NLP tasks. **Topic Modeling:** Introduction to topic modeling-Latent Dirichlet Allocation (LDA)-Non-Negative Matrix Factorization (NMF).			
Text Book	Text Book 1:16,18		
MODULE-5	**Applications and Future Directions in NLP**	22AIM53.5, 22AIM53.6	**8 Hours**
Applications and Implementation of NLP: Sentiment Analysis - Text Classification- Text Summarization- Named Entity Recognition code- Chatbots and Dialogue systems. **Future Trends in NLP-**Emerging trends and research areas-AI-driven NLP tools and services.			

Text Books:

1) **Daniel Jurafsky and James H. Martin**, “***Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition”*** (Prentice Hall Series in Artificial Intelligence), 2017. ISBN: 0133252930, 9780133252934

2) **Jacob Eisenstein**, “***Natural Language Processing***”, MIT Press, 2019. ISBN: 9780262042840 https://web.stanford.edu/~jurafsky/slp3

Introduction

Hey Google

“is it going to rain today?”

“do I need an umbrella today?”

What is NLP?

Natural language processing (NLP) is the ability of a computer program to understand **human language** as it is spoken.

NLP is a component of artificial intelligence.

Programming Languages

C++, Java, Python ...

Natural Languages

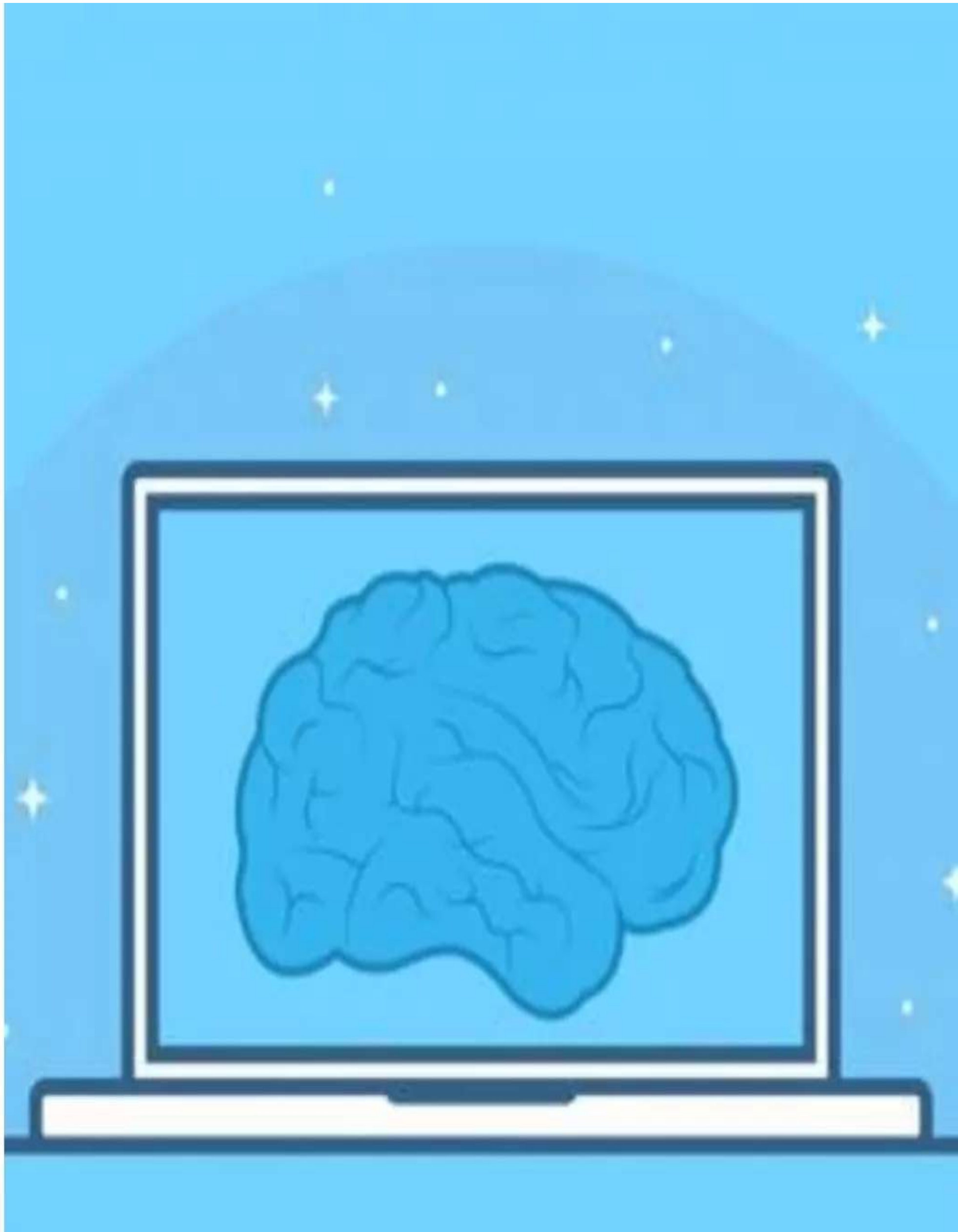
English, Hindi, Bengali ...

Computers traditionally require humans to "speak" to them in a programming language that is **precise**, **unambiguous** and **highly structured**.

Human speech, is not precise - it is often ambiguous and the linguistic structure can depend on many complex variables, like **regional dialects** and **social context**.

How does NLP work?

Current approaches to NLP are based on **machine learning algorithms**, a subset of AI that examines and uses patterns in data to improve a program's understanding.



Techniques used in NLP

1. Syntax

In NLP, syntactic analysis is used to assess how the natural language aligns with the **grammatical rules**.

2. Semantics

Semantics refers to the meaning that is conveyed by text and involves computer algorithms to interpret the same.

“CSK was on fire last Sunday, they totally destroyed KKR”

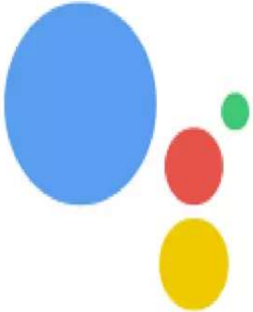
- To a computer, this may mean CSK was literally on fire.
- CSK literally destroyed KKR and it doesn’t exist anymore!

Applications of NLP

(some of the many)

1. Voice Based

Google Translate



Google Assistant

Siri

2. Text Based

SwiftKey Keyboard

Microsoft Word

Grammarly

Benefits of NLP

1. Sentiment Analysis

Enables data scientists to assess comments on social media to see how their business's brand is performing.

2. Searching Text

NLP allows analysts to sift through massive troves of text to find **relevant information**.

Enterprise Search | Text Summarization | Filtering Sensitive Keywords

Limitations of NLP

(understanding abstract rules)

1. High Level Rules

Sometimes, words can be high-level and abstract; for example, when someone uses a sarcastic remark to pass information.

2. Low Level Rules

Some of the rules can be low-levelled; for example, using the character “s” to signify the **plurality** of items.

Steps in NLP

1. Segmentation <<< break data into sentences

2. Tokenizing <<< break sentence into words

3. Stop Words <<< mark down 'Verb to be., prepositions, ...etc...‘

4. Stemming <<< same words with different prefix or suffix

5. Lemmatization <<< learning that multiple words can have the same meaning (is, am, are >>> be)

6. Speech Tagging <<< adding tags to words (noun, verb, preposition)

7. Name Entity Tagging <<< introduce machine to some group of words that may occur in documents

8. Machine Learning (ex. naive bayes calssification) <<< learning the human sentiment and speech