

24MCAE305: NATURAL LANGUAGE PROCESSING

Hours/Week: 3
Credits: 3

I.A. Marks: 30
Exam. Marks: 70

Course Learning Objectives:

- To introduce students the challenges of empirical methods for natural language processing (NLP) applications.
- To introduce basic mathematical models and methods used in NLP applications to formulate computational solutions.
- To introduce students research and development work in information retrieval, information extraction, and knowledge discovery using different natural language resources.
- Understand the principles of language resource annotation and its use in machine learning applications and apply the above principles in analysis of data and acquire intended information through the use of available tools.

Course Outcomes: After completing the course, the students will be able to,

CO1: Understand basic approaches to syntax and semantics in NLP.

CO2: Realize approaches to discourse, generation and dialogue in NLP

CO3: Familiarize the current methods for statistical approaches to machine translation.

CO4: Understand machine learning techniques used in NLP, including hidden Markov models and probabilistic context-free grammar.

CO5: Familiar with clustering and unsupervised methods, log-linear and discriminative models, and the EM algorithm as applied within NLP

CO6: Understand the design and implementation issues in various NLP applications such as information retrieval and information extraction.

CO7: Understand the principles of language resource annotation and its use in machine learning applications and apply the above principles in analysis of data and acquire intended information through the use of available tools.

UNIT-I

12 Hrs.

Introduction to NLP: History, Advantages and Disadvantages, Challenges, Components, Applications, NLP pipeline, Phases of NLP, NLP APIs, NLP Libraries, Cross-lingual and Multilingual NLP. Processing Indian Languages. Regular Expressions, Text Normalization, Edit Distance.

UNIT-II

12 Hrs.

Lexical Semantics: Vector Semantics, Words and Vectors, Cosine for measuring similarity, TF-IDF: Weighing terms in the vector, Pointwise Mutual Information (PMI), Applications of the TF-IDF or PPMI vector models. Words Representation: Bag-Of-Words, skip-gram, Continuous Bag-Of-Words, Embedding representations for words, Word2vec, fastText, Visualizing Embeddings, Semantic properties of Embeddings, Bias and Embeddings, Evaluating Vector Models.

UNIT-III

12 Hrs.

Language Modeling and Sequence Labeling: Unigram Language Model, Bigram, Trigram, N-gram, Advanced smoothing for language modeling, Empirical Comparison of Smoothing Techniques, Applications of Language Modeling, Natural Language Generation, Parts of Speech Tagging, Morphology, Named Entity Recognition.

UNIT-IV**12 Hrs.**

Machine Translation and Encoder-Decoder Models: Language Divergences and Typology, Encoder-Decoder with RNNs, Attention, Encoder-Decoder with Transformers. Transfer Learning with Contextual Embeddings and Pre-trained language models: Bidirectional Transformer Encoders, Training Bidirectional Encoders, Transfer Learning through Fine-Tuning, Prompting

REFERENCE BOOKS:

1. Daniel Jurafsky and James H. Martin. 2020. Speech and Language Processing. 3rd Edition (draft)
2. Christopher D. Manning and Hinrich Schütze. 1999. Foundations of Statistical Natural Language Processing. MIT Press.
3. Sowmya Vajjala, Bodhisattwa Majumder, Anuj Gupta, Harshit Surana. 2020. Practical Natural Language Processing. O'Reilly.
4. Hobson Lane, Cole Howard, Hannes Hapke. 2019. Natural Language Processing in Action. Live Book.

