# SARVAJANIK UNIVERSITY
# Sarvajanik College of Engineering and Technology
# Bachelor of Technology

## Year: B. Tech IV (Semester VII)

**Subject Name:** Natural Language Processing **Subject Code:** BTAI13702
**Type of course:** Professional Core Course
**Prerequisite (if any):** Basic course on Machine Learning

**Rationale:** Natural language processing (NLP) is a branch of Artificial Intelligence that gives computers the ability to process, comprehend, and generate natural language. It has many useful applications like NLP-enabled search engines, document classification, machine translation, etc. The purpose of this course is to give students an understanding of important tasks of NLP and perform them with Machine Learning techniques.

**Teaching and Examination Scheme:**

| Teaching Scheme | | | | Theory Marks | | | Practical Marks | | Total |
|---|---|---|---|---|---|---|---|---|---|
| L | T | P | C | TEE | CA1 | CA2 | TEP | CA3 | |
| 3 | 0 | 2 | 4 | 60 | 25 | 15 | 30 | 20 | 150 |

CA1: Continuous Assessment (assignments/projects / open book tests / closed book tests) CA2: Sincerity in attending classes/class tests / timely submissions of assignments/self-learning attitude / solving advanced problems TEE: Term End Examination TEP: Term End Practical Exam (Performance and viva on practical skills learned in course) CA3: Regular submission of Lab work / Quality of work submitted / Active participation in lab sessions/viva on practical skills learned in the course.

**Contents:**

| Sr. No. | Contents | Total Hours |
|---|---|---|
| 1. | **Introduction to Natural Language Processing (NLP):** Various stages of NLP, Why NLP is difficult? Lexical Ambiguity, Language imprecision and vagueness, Part of Speech Noun and Pronouns, Words: Determiners and adjectives | 03 |
| 2. | **Text Processing and Morphology:** Basic Regular Expression patterns, Corpus, Type/Token Ratio, Zipf's Law, Vocabulary Growth-Heaps law, Word Tokenization, Word Normalization, Lemmatization, and Stemming, Sentence Segmentation, Spelling Correction- Minimum Edit Distance, Inflectional and Derivational Morphology, Word Formation, Finite-state methods for morphology | 06 |
| 3. | **Language Modelling:** Need of n-gram probabilities-unigram, Bigram, Trigram, N-gram, Probability of words in sentences, Markov Assumption, maximum likelihood estimation, Evaluation of Language models, Perplexity, Smoothing-Laplace Smoothing, add-one smoothing, Kneser-Ney smoothing | 07 |

| | | |
|---|---|---|
| 4. | **Sequence Labelling for Parts of Speech and Named Entities**: Different Parts of Speech(POS), Penn Treebank tagset, POS Tagging, Named Entity Tagging, Hidden Markov Model POS Tagging, Solving, The Viterbi Algorithm, POS Tagging using WordNet | 06 |
| 5. | **Context-free grammar and Parsing:** Context-Free Grammar, Parse tree, Parsing, Top-down Parsing, Bottom-Up Parsing, Dynamic Programming Parsing algorithms- Cocke-Kasami-Younger (CKY) algorithm, Probabilistic Context-free grammars (PCFGs), CKY for PCFG, The Probability of a String, Inside and Outside Probabilities, Dependency Grammars, and Parsing | 06 |
| 6. | **Lexical and Vector Semantics:** Types lexical relations, Word Similarity, Semantic similarity measures counting, Information content, Computational Semantics, Word Space, Context weighting: documents as context, words and co-occurrence vectors, Term frequency- Inverse document frequency(TF-IDF), Pointwise Mutual Information (PMI), Similarity Measures for binary vectors and probability distributions, Word Embeddings-Word2vec, Continuous Bag of Words Model (CBOW) , skip-gram, Word Sense Disambiguation(WSD), Supervised Disambiguation, Dictionary-Based and Thesaurus-based disambiguation, Lesk algorithm, Latent Dirichlet Allocation | 12 |
| 7. | **Applications and Tools:** Machine Translation, Question Answering, Information Retrieval, Chatbots and dialogue Systems, APIs and Libraries: NLTK, spaCy, OpenNLP, PyTorch-NLP, RegEx, TextBlob | 05 |

**Suggested Specification table with Marks (Theory): (For B. Tech only)**

| **Distribution of Theory Marks** | | | | | |
|---|---|---|---|---|---|
| **R Level** | **U Level** | **A Level** | **N Level** | **E Level** | **C Level** |
| 20 | 20 | 10 | 10 | - | - |

Legends: R: Remembrance; U: Understanding; A: Application, N: Analyze and E: Evaluate C: Create (Revised Bloom's Taxonomy)

**Reference Books:**

| **Sr no** | **Title of book /article** | **Author(s)** | **Publisher and details like ISBN** | **Year of publication** | **Publication Edition** |
|---|---|---|---|---|---|
| 1 | Speech and Language Processing | Daniel Jurafsky, James H. Martin | Pearson Education | 2008 | 2nd Edition |

| | | | | | |
|---|---|---|---|---|---|
| 2 | Introduction to Natural Language Processing | Jacob Eisenstein | The MIT Press | 2019 | 1st Edition |
| 3 | Natural Language Understanding | James Allen | Pearson Education | 1994 | 2nd Edition |
| 4 | NLP with Python | Steven Bird Ewan Klien Edward Loper | OReilly | 2011 | 1st Edition |
| 5 | Natural Language Processing in Action | Hobson Lane, Cole Howard, Hannes Hapke | Manning | 2019 | 1st Edition |

**Course Outcomes (CO):**

| Sr. No. | CO statements | Marks % weightage |
|---|---|---|
| CO-1 | Identify different linguistic components of natural language and Demonstrate the state-of-the-art algorithms for text-based processing of natural language concerning morphology | 30% |
| CO-2 | Apply suitable language modeming and sequence modeming techniques based on the structure of the language | 25% |
| CO-3 | Analyze the syntax, semantics, and pragmatics of a statement written in a natural language. | 35% |
| CO-4 | Design NLP-based AI systems for given applications and demonstrate the NLP tools and systems. | 10% |

**List of Open learning website**

- NPTEL course on Natural Language Processing (https://onlinecourses.nptel.ac.in/noc23_cs45)

## List of Experiments:

| Sr. No. | Practicals |
|---|---|
| 1 | Perform the following tasks of NLP on given corpus using NLTK Toolkit<br>● Tokenization<br>● Stemming and lemmatization<br>● Parts of speech identification |
| 2 | Build a text classification system using TextBlob Library |
| 3 | Build a Chabot for school/college/shopping website using regex tool. |
| 4 | Pre-process the data collected from web scraping or manual data collection by applying the regex tool |
| 5 | Implement N-Grams using Python. |
| 6 | Create and run a recursive descent parser over both a syntactically ambiguous and unambiguous sentence |
| 7 | Identify the grammatical group of a given sentence using POS Tagger |
| 8 | Implement Named Entity recognition on News Article using Spacy Python Library |
| 9 | Implement text similarity algorithms in Python. |
| 10 | Implement Lesk Algorithm in Python |
| 11 | Implementation of text classification using Naïve Bayes |
| 12 | Do a detailed Study of any Conversational Agents in AI |
| 13 | Study test processing for the Indian language using iNLTK. |