GUJARAT TECHNOLOGICAL UNIVERSITY

Program Name: Bachelor of Engineering
Level: UG
Subject Code : 3174808
Subject Name : Natural Language Processing

w. e. f. Academic Year:	A.Y. 2025-26
Semester:	VII
Category of the Course:	PEC

Prerequisite:	Machine Learning, Mathematics
Rationale:	To start with Natural Language Processing (NLP), you should have a good understanding of Python programming and basic data structures. A foundation in mathematics—especially linear algebra, probability, and statistics—is important for understanding machine learning models. Knowledge of machine learning concepts like supervised learning and deep learning (e.g., neural networks, transformers) is also essential. Familiarity with NLP libraries such as NLTK, spaCy, or Hugging Face Transformers will help you work on real-world projects. Some understanding of linguistics can be beneficial but is not strictly necessary.

Course Outcome:
After Completion of the Course, Student will able to:

No	Course Outcomes	RBT Level
1	Understand the foundational concepts of Natural Language Processing, including language modeling, morphology, and text preprocessing techniques such as tokenization and spelling correction.	U
2	Apply statistical methods such as n-gram modeling and part-of-speech tagging using rule-based, stochastic, and machine learning approaches for word-level analysis.	A
3	Analyze and implement syntactic structures of language using context-free grammars, parsing algorithms, and probabilistic grammar models.	An
4	Demonstrate knowledge of semantic representation techniques, including first-order logic, word sense disambiguation, and methods for calculating word similarity.	R
5	Explore fundamental concepts in speech processing and modeling, including feature extraction, dynamic time warping, and Hidden Markov Models for speech recognition tasks.	C

**Revised Bloom's Taxonomy (RBT)*

GUJARAT TECHNOLOGICAL UNIVERSITY

Program Name: Bachelor of Engineering
Level: UG
Subject Code : 3174808
Subject Name : Natural Language Processing

Teaching and Examination Scheme:

Teaching Scheme (in Hours)			Total Credits L+T+ (PR/2)	Assessment Pattern and Marks				Total Marks
				Theory		Tutorial / Practical		
L	T	PR	C	ESE (E)	PA / CA (M)	PA/CA (I)	ESE (V)	
3	0	2	4	70	30	20	30	150

Course Content:

Unit No.	Content	No. of Hours
1.	**Introduction:** Origins and challenges of nlp – language modeling: grammar-based lm, statistical lm – regular expressions, finite-state automata – english morphology, transducers for lexicon and rules, tokenization, detecting and correcting spelling errors, minimum edit distance	06
2.	**Word Level Analysis:** Unsmoothed n-grams, evaluating n-grams, smoothing, interpolation and backoff – word classes, part-of-speech tagging, rule-based, stochastic and transformation-based tagging, issues in pos tagging – hidden markov and maximum entropy models.	06
3	**Syntactic Analysis:** Context free grammars, grammar rules for english, treebanks, normal forms for grammar – dependency grammar – syntactic parsing, ambiguity, dynamic programming parsing – shallow parsing – probabilistic cfg, probabilistic probabilistic lexicalized cfgs – feature structures, unification of feature structures.	08
4	**Semantics And Pragmatics:** Requirements for representation, first-order logic, description logics – syntax-driven semantic analysis, semantic attachments – word senses, relations between senses, thematic roles, selectional restrictions – word sense disambiguation, using supervised, dictionary & thesaurus, bootstrapping methods – word similarity using thesaurus and distributional methods.	06

5.	**Basic Concepts of Speech Processing:** Speech fundamentals: articulatory phonetics – production and classification of speech sounds; acoustic phonetics – acoustics of speech production; review of digital signal processing concepts; short-time fourier transform, filter- bank and lpc methods.	04
6.	**Speech-Analysis,Speech Modeling:** Features, feature extraction and pattern comparison techniques: speech distortion measures– mathematical and perceptual – log–spectral distance, cepstral distances, weighted cepstral distances and filtering, likelihood distortions, spectral distortion using a warped frequency scale coefficients, time alignment and normalization – dynamic time warping, multiple time – alignment paths. Hidden markov models: markov processes– evaluation, optimal state sequence – viterbi search, baum-welch parameter re-estimation, implementation issues.	08

Suggested Specification Table with Marks (Theory):

Distribution of Theory Marks (in %)					
R Level	**U Level**	**A Level**	**N Level**	**E Level**	**C Level**
10	**30**	**25**	**20**	**10**	**05**

Where R: Remember; U: Understanding; A: Application, N: Analyze and E: Evaluate C: Create (as per Revised Bloom's Taxonomy)

References/Suggested Learning Resources:

(a) Books:

1. Bharati A., Sangal R., ChaitanyaV.. Natural language processing: a Paninian perspective, PHI
2. Siddiqui T., Tiwary U. S. Natural language processing and Information retrieval, OUP
3. Daniel Jurafsky And James H Martin, "Speech And Language Processing – An Introduction To Natural Language Processing, Computational Linguistics, And Speech Recognition", Pearson Education

Sample List of Experiments:

1. Write a program To understand and implement basic text preprocessing tasks including tokenization, stemming, and lemmatization using regular expressions and morphological rules.
2. Write a program To implement minimum edit distance algorithms for detecting and correcting spelling errors in natural language text.
3. Write a program to construct and evaluate n-gram language models, including techniques like smoothing, backoff, and interpolation.

4. Write a program to perform part-of-speech tagging using rule-based and statistical approaches such as Hidden Markov Models.
5. Write a program to develop a parser based on context-free grammars for syntactic analysis of English sentences.
6. Write a program to understand and implement probabilistic parsing using Probabilistic CFGs and treebank data.
7. Write a program to implement and compare different methods for disambiguating word senses, including supervised and dictionary-based techniques.
8. Write a program to compute and evaluate semantic similarity between words using thesaurus-based and distributional methods.
9. Write a program to analyze speech signals and extract relevant acoustic features using STFT, LPC, and filter bank techniques.
10. Write a program to implement speech recognition using Hidden Markov Models, including training with Baum-Welch and decoding with Viterbi algorithms.

****