GUJARAT TECHNOLOGICAL UNIVERSITY

Program Name: Engineering

Level: PG

Branch: Artificial Intelligence and Data Science

Subject Code: ME02095061

Course/Subject Name: Natural Language Processing & Applications

WEF Academic Year	2024-25
Semester	2
Category of the Course	Professional Elective Course

Prerequisite	**Data Structures and Programming**
Rationale	There is an increasing demand of making the computing systems to be learning the natural languages. It should also able to process the different parts of natural languages and respond to the user in again natural language. The objective of this course is prepare the learners for developing the different modules of natural language processing like lexical analysis, error checking, part-of-speech tagging, syntax checking etc.

Course Outcome:

After completion of the Course, Students will be able to:

No	Course Outcomes	RBT Level*
01	Understand fundamental mathematical models and algorithms in the field of natural language processing.	UN
02	Apply lexical, syntax and semantic analysis methods on textual data.	AP
03	Apply natural language processing on real world problems of speech recognition, automated question answering, text classification.	AP
04	Analyze the lexical, syntax and semantic analysis methods.	AN
05	Evaluate the performance of natural language processing methods and compare it.	EV

*RM: Remember, UN: Understand, AP: Apply, AN: Analyze, EL: Evaluate, CR: Create

GUJARAT TECHNOLOGICAL UNIVERSITY

Program Name: Engineering

Level: PG

Branch: Artificial Intelligence and Data Science

Subject Code: ME02095061

Course/Subject Name: Natural Language Processing & Applications

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)	
03	00	02	04	70	30	20	30	150

Course Content:

Unit	Course Content	No of Hours	% of Weightage
1.	**Introduction** Concept and Use of Natural Language Processing, Ambiguity and uncertainty in language, The Turing test, Models and Algorithms	03	10%
2.	**Regular Expressions, Automata and Finite-State Transducers** Basic Regular Expression Patterns, Disjunction, Grouping, and Precedence, Advanced Operators, Regular Expression Substitution, Using FSA to Recognize Sheeptalk, Formal Languages, Nondeterministic FSAs, Using an NFSA to accept strings, Recognition as Search, Relating Deterministic and Non-deterministic Automata, Regular Languages and FSAs, Survey of (Mostly) English Morphology, Inflectional Morphology, Derivational Morphology, The Lexicon and Morphotactics, Morphological Parsing with Finite-State Transducers, Orthographic Rules and Finite-State Transducers, Human Morphological Processing, N-gram Language Models, Words, Corpora, Text Normalization, Minimum Edit Distance, N-Grams, Evaluating Language Models, Generalization and Zeros, Smoothing, Kneser-Ney Smoothing	14	30%
3.	**Word Classes and Part-of-Speech Tagging** English Word Classes, Tagsets for English, Part of Speech Tagging, Rule-based Part-of-speech Tagging, Stochastic Part-of-speech Tagging,	07	20%

	The Actual Algorithm for HMM tagging, Transformation-Based Tagging, How TBL rules are applied, How TBL Rules are Learned, Multiple tags and multiple words, Unknown words, Class-based N-grams		
4.	**Context-Free Grammars, Semantics and Pragmatics** Constituency, Context-Free Rules and Trees, Sentence-Level Constructions, The Noun Phrase, The Verb Phrase and Subcategorization, Auxiliaries, Spoken Language Syntax, Grammar Equivalence & Normal Form, Finite State & Context-Free Grammars, Grammars & Human Processing, Parsing as Search, Top-Down and Bottom-Up Parsing with comparison, A Basic Top-down Parser, Adding Bottom-up Filtering, Problems with the Basic Top-down Parser, Left Recursion, Ambiguity, Syntax-Driven Semantic Analysis, Attachments for a Fragment of English, Idioms and Compositionality, Robust Semantic Analysis, WordNet: A Database of Lexical Relations, Reference Resolution, Text Coherence, Discourse Structure, Psycholinguistic Studies of Reference and Coherence	15	30%
5.	**Recent Trends in the Domain of Natural Language Processing**	06	10%
	TOTAL	**45**	**100**

Suggested Specification Table with Marks (Theory):

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

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

Reference/Suggested Learning Resources:

(a) Books:

1. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition by Daniel Jurafsky & James H. Martin Pearson.
2. Speech and Language Processing by Daniel Jurafsky and James H. Martin Second edition, Prentice Hall.
3. Natural Language Understanding by Allen James Second edition, Benjamin/Cumming.
4. Statistical Language Learning by Charniack Eugene MIT Press.
5. Foundations of Statistical Natural Language Processing by Manning Christopher and Heinrich Schutze MIT Press.

(b) Open source software and website

- Course-related online MOOCs on NPTEL/SWAYAM platform.
- Recently Published papers/articles in reputed journals.

Suggested Course Practical List:

- The practical work will be carried out based on the content covered during the academic sessions**.**

List of Laboratory/Learning Resources Required: Programming development environment (open source is encouraged) related to the course content.

Suggested Project List: The subject teacher has to assign the relevant project work to the students in individual/team.

Suggested Activities for Students: The subject teacher has to assign the outcome based activities to the students in individual/team.

****