Uka Tarsadia University

B. Tech.

CSE / CSE (CC) / CE (SE)

Semester VII

Program Elective - V
SPEECH AND LANGUAGE PROCESSING
AI6012

EFFECTIVE FROM July-2024

Syllabus version: 1.00

Subject Code	Subject Title
AI6012	**Speech and Language Processing**

Teaching Scheme				Examination Scheme				
Hours		Credits		Theory Marks		Practical Marks		Total Marks
Theory	Practical	Theory	Practical	Internal	External	Internal	External	
3	2	3	1	40	60	20	30	150

Objectives of the course:

- To introduce various components of natural language processing.
- To develop familiarity to linguistics and their application to part-of-speech tagging.
- To develop background to various aspects of NLP like morphology, syntax analysis and parsing, semantic analysis and discourse and pragmatics.
- To give exposure to develop applications of machine translation, text summarization and question-answering system.

Course outcomes:

Upon completion of the course, the student shall be able to,

CO1: Understand the key concepts of NLP and identify the NLP challenges and issues.

CO2: Comprehend the concepts and techniques of NLP for analyzing words based on morphology and corpus.

CO3: Get acquainted with various techniques of Part-of-Speech tagging.

CO4: Illustrate computational methods to understand language phenomena of word sense disambiguation.

CO5: Apply different techniques for anaphora and reference resolution.

CO6: Develop solutions to sub problems of applications of NLP such as machine translation, text summarization and question-answering system.

Sr. No.	Topics	Hours
	Unit – I	
1	**Introduction to NLP:** What is NLP? Components of NLP – Natural language understanding, Natural language generation, Text preprocessing for NLP, Corpus, Phases of NLP, Ambiguity in each phase. **Regular Expressions and Automata:** Regular expressions, Finite state automata.	5
	Unit – II	
2	**Word Level Analysis:**	9

	Morphology analysis – Survey of English morphology, Types of morphemes, Types of morphology – Inflectional morphology, Derivational morphology, Compounding morphology, Cliticization morphology, Finite state transducers (FST), Morphological parsing with FST, Lexicon free FST Porter stemmer, N-Grams – N-gram language model, N-gram for spelling correction, Minimum edit distance.	
	Unit – III	
3	**Syntax Analysis:** Part-Of-Speech tagging (POS) – Tag set for English (Penn Treebank), Rule based POS tagging, Stochastic POS tagging, Transformation based tagging, Issues – Multiple tags & words, Unknown words, Introduction to CFG, Sequence labeling: Hidden Markov Model (HMM), Maximum Entropy.	9
	Unit – IV	
4	**Semantic Analysis:** Lexical Semantics, Attachment for fragment of English – Sentences, Noun phrases, Verb phrases, Prepositional phrases, Relations among lexemes & their senses – Homonymy, Polysemy, Synonymy, Hyponymy, WordNet, Word Sense Disambiguation (WSD) – Knowledge based, Dictionary based, and Supervised word sense disambiguation.	9
	Unit – V	
5	**Discourse and Pragmatics:** Cohesion – Reference, Ellipsis, Lexical cohesion, Reference resolution – Constraints and preferences, Reference resolution algorithms – Resolution of anaphora procedure, Hobbs algorithm, Centering algorithm.	9
	Unit – VI	
6	**Applications:** Sentiment analysis, Voice assistants and chatbots/question answering, Machine translation, Grammar checkers, Email filtering, Autocomplete in search engines, Autocomplete in search engines, Text summarization.	4

Sr. No.	**Speech and Language Processing (Practicals)**	**Hours**
1.	Installing and introduction of NLTK.	2
2.	Write a program that pre-process a raw text using NLTK Library.	4
3.	Write a program to perform followings: a) Import spacy and load the language mod. b) Tokenize a text using the transformers package. c) Extract usernames from emails.	4

	d) Find the most common words in the text excluding stop words. e) Do spell correction in a given text.	
4.	Write a program to perform followings: a) Extract all the nouns in a text. b) Extract all the pronouns in a text. c) Find similarity between two words. d) Installation and introduction of Scikit-learn library for data classification. d) Find the similarity between any two text documents. e) Replace all the pronouns in a text with their respective object names.	4
5.	Write a program to perform followings: a) Create bigrams using Genism's phraser. b) Extract all bigrams, trigrams using N-grams of NLTK library. c) Extract and print the noun phrases in given text document. d) Extract verb phrases from the text. e) Extract first name and last names present in the document.	4
6.	Implement N-gram model using financial news dataset in which sentiments are from the perspective of retail investors. Do the basic preprocessing and generate N-grams.	4
7.	Write a program to perform followings: a) Identify named entities and print all the named entities with their labels. b) Identify and extract a list of all organizations/Companies mentioned in the news article. c) Identify and replace all the person names in the news article with UNKNOWN to keep privacy. d) Display the named entities present in the given document along with their categories using spacy. e) Find and print the root word/headword of any word in the given sentence. f) Find the dependencies of all the words in the given text. g) Visualize the dependencies of various tokens of the given text using spaCy.	4
8.	Case study and project: Text summarization.	2
9.	Case study and project: Information retrieval.	2

Textbook:

1. Daniel Jurafsky and James H. Martin, "Speech and Language processing, “An introduction to Natural Language Processing, Computational linguistics, and Speech recognition", Prentice Hall.

Reference books:

1. Siddiqui and Tiwary U.S., “Natural Language Processing and Information Retrieval”, Oxford University Press.
2. Steven Bird, Ewan Klein and Edward Loper, "Natural Language Processing in Python", O’Reilly.

3. V. Manning, Christopher D. and Hinrich Schuetze, "Foundations of Statistical NaturalLanguage Processing", Cambridge, MIT Press.
4. Allen J., “Natural Language understanding”, Pearson Publication.

Course objectives and Course outcomes mapping:

- To introduce various components of natural language processing: CO1.
- To develop familiarity to linguistics and their application to part-of-speech tagging: CO3.
- To develop background to various aspects of NLP like morphology, syntax analysis and parsing, semantic analysis and discourse and pragmatics: CO2, CO3, CO4, CO5.
- To give exposure to develop applications of machine translation, text summarization and question-answering system: CO6.

Course units and Course outcomes mapping:

Unit No.	Unit Name	Course Outcomes					
		CO1	**CO2**	**CO3**	**CO4**	**CO5**	**CO6**
1	Introduction to Natural Language Processing (NLP), and Regular Expressions and Automata	✓					
2	Word Level Analysis		✓				
3	Syntax Analysis			✓			
4	Semantic Analysis				✓		
5	Discourse and Pragmatics					✓	
6	Applications						✓

Programme outcomes:

PO 1: Engineering knowledge: An ability to apply knowledge of mathematics, science, and engineering.

PO 2: Problem analysis: An ability to identify, formulates, and solves engineering problems.

PO 3: Design/development of solutions: An ability to design a system, component, or process to meet desired needs within realistic constraints.

PO 4: Conduct investigations of complex problems: An ability to use the techniques, skills, and modern engineering tools necessary for solving engineering problems.

PO 5: Modern tool usage: The broad education and understanding of new engineering techniques necessary to solve engineering problems.

PO 6: The engineer and society: Achieve professional success with an understanding and appreciation of ethical behavior, social responsibility, and diversity, both as individuals and in team environments.

PO 7: Environment and sustainability: Articulate a comprehensive world view that integrates diverse approaches to sustainability.

PO 8: Ethics: Identify and demonstrate knowledge of ethical values in non-classroom activities, such as service learning, internships, and field work.

PO 9: Individual and team work: An ability to function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.

PO 10: Communication: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give/receive clear instructions.

PO 11: Project management and finance: An ability to demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.

PO 12: Life-long learning: recognition of the need for, and an ability to engage in life-long learning.

Programme outcomes and Course outcomes mapping:

Programme	Course Outcomes					
Outcomes	CO1	CO2	CO3	CO4	CO5	CO6
PO1	✓	✓	✓	✓	✓	✓
PO2		✓	✓	✓	✓	✓
PO3	✓	✓	✓	✓	✓	✓
PO4				✓	✓	✓
PO5	✓	✓	✓	✓	✓	✓
PO6						
PO7						
PO8						
PO9						✓
PO10						
PO11						
PO12						