Year: M. Tech. I (Semester – II)

Subject Name: Natural Language Processing
Subject Code: MTCO14204
Type of course: Professional Elective - III
Prerequisite (if any): Python, Machine Learning

List of Courses where this course will be prerequisite: --

Rationale: Automated processing of human languages is increasingly becoming important for different types of applications including language translation, surveys, chatbots etc. This subject introduces the fundamentals of natural language processing and its applications in various problem domains.

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 course

Content:

Sr. No.	Content	Total Hrs
1	Basics of NLP and text processing: What is Natural Language Processing, Ambiguity and uncertainty in language, STeps of NLP, NLP tasks in syntax, semantics, and pragmatics. Applications such as information extraction, question answering, and machine translation. The problem of ambiguity, Discourse Analysis, Text PreProcessing -- Tokenisation, stemming, lemmatization, stop word removal, unigram, bigram, ngram, sentence segmentation	4
2	Lexical Processing/Lexical Semantics: Chomsky hierarchy, regular languages, and their limitations, Finite state Methodology for Morphology, regular expressions for finding and counting, regex tools, Distributional measures of	6

	similarity, Concept Mining using Latent Semantic Analysis, Spelling correction using edit distance, weighted edit distance	
3	Parsing for NLP: Classical Parsing (Bottom up, top down, Dynamic Programming: CYK parser), Noun Structure; Non-noun Structure, Recursion in syntactic structure and Parsing Algorithms, Grammar rule for English, Probabilistic parsing; sequence labeling, PCFG, Probabilistic parsing: Training issues, Arguments and Adjuncts, Probabilistic parsing; inside-outside probabilities, Treebank	7
4	Part Of Speech Tagging and Sequence Labeling: Rule based POS Tagging, Properties of Rule Based POS Tagging, Stochastic POS Tagging, Properties of Stochastic POS Tagging, Transformation based Tagging, Hidden Markov Model (HMM) POS Tagging	4
5	Language Model: The role of language models, Simple N-gram models. Estimating parameters and smoothing, Evaluating language models	6
6	Word Sense Disambiguation : Supervised Machine Learning Approach for WSD,Dictionary and Thesaurus Methods, Simplified Lesk Algorithm, WordNet and WordNet based similarity measures, Colocational features and Bag of word features	6
7	Information Extraction (IE) and Named Entity Recognition (NER) : Introduction to Named Entity Recognition and Relation Extraction , Approaches to NER	4
8	Machine Translation (MT) : Rule based MT, Statistical Machine Translation (SMT),	4
9.	Case studies : Dialogue and Conversational Agent, Discourse Analysis, Challenges in NLP for Indian Languages , Sentiment Analysis	4

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 :An Introduction to	Daniel Jurafsky and James H. Martin	Prentice Hall	2009	2nd edition

	Natural Language Processing, Speech Recognition, and Computational Linguistics				
2	NLP with Python	Steven Bird Ewan Klien Edward Loper	OReilly	2009	
3	Natural Language Processing in Action	Hobson Lane, Cole Howard, Hannes Hapke	Manning	2019	

Course Outcomes:

Sr.No.	CO statement	Marks % weightage
CO-1	Describe the fundamental concepts and techniques of natural language processing.	15%
CO-2	Distinguish among the various techniques, taking into account the assumptions, strengths, and weaknesses of each.	25%
CO-3	Use appropriate descriptions, visualizations, and statistics to communicate the problems and their solutions.	25%
CO-4	Analyze large volume text data generated from a range of real-world applications.	20%
CO-5	Demonstrate use of NLP in various applications	15%

List of Open learning website:

http://www.nltk.org

List of Open Source Software:

- NLTK
- Stanford Core NLP
- Spacy

FOR LAB SESSIONS:

List of Experiments:

Sr.No.	Practical
1	Perform the following tasks of NLP on brown corpus using NLTK Toolkit

	• Tokenize into sentences and words • Stopwords • Collocations • Parts of speech identification • Stemming and lemmatization • Corpus
2	Build a text classification system using TextBlob Library
3	Build a chatbot for school/college/shopping website using regex tool.
4	Preprocess the data collected from web scraping or manual data collection by applying regex tool
5	Create and run a recursive descent parser over both a syntactically ambiguous and unambiguous sentence
6	Compare the performance of the top-down, bottom-up, and left-corner parsers using the same grammar and three grammatical test sentences. Use timeit to log the amount of time each parser takes on the same sentence. Write a function that runs all three parsers on all three sentences, and prints a 3-by-3 grid of times, as well as row and column totals. Discuss your findings.
7	Identify the grammatical group (NOUN, PRONOUN, ADJECTIVE, VERB, ADVERBS) of a given sentence using POS Tagger
8	Implement Word2Vec embedding for BBC News dataset
9	Implement Named Entity recognition on News Article using Spacy Python Library
10	Case Study for Building a Conversational Agent • IBM WATSON • Slack API

Major Equipment Needed: -NIL-