Course Code:CSE3015	Course Title: Natural Language Processing	TPC	3	2	4
Version No.	**2.2**				
Course Pre-requisites/ Co-requisites	**CSE3008 - Introduction to Machine Learning**				
Anti-requisites (if any).	**None**				
Objectives:	1. To introduce students,the fundamentals of Natural Language Processing. 2. Learn the techniques in natural language processing. 3. Be familiar with the natural language generation. 4. Be exposed to Text Mining. 5. Understand the information retrieval techniques				

CO's Mapping with PO's and PEO's

Course Outcomes	Course Outcome Statement	PO's / PEO's
CO1	Understand and implement NLP Preprocessing and Embedding Techniques	PO1,PO2, PO3,PEO1,PEO3
CO2	Understand and analyse Parts of Speech Tagging	PO2, PO3
CO3	Understand and analyse Parsing Concepts	PO2, PO3
CO4	Understand and evaluate NLP using Deep Learning Techniques	PO2, PO3
CO5	Develop web crawlers	PO2, PO3
CO6	Understand various NLP applications Latest Techniques	PO2, PO3
	TOTAL HOURS OF INSTRUCTIONS : 45	

A

Module No. 1	Introduction to NLP	8 Hours

Overview: Origins and challenges of NLP-Need of NLP, Preprocessing techniques- Text Wrangling, Text cleansing, sentence splitter, tokenization, stemming, lemmatization, stop word removal, rare word removal, spell correction.Word Embeddings, Different Types : One Hot Encoding, Bag of Words (BoW), TF-IDF Static word embeddings: Word2vec, GloVe, FastText

Module No. 2	Parts of Speech Tagging	6 Hours

Parts of Speech Tagging and Named Entities –Tagging in NLP, Sequential tagger, N-gram tagger, Regex tagger,Brill tagger, NER tagger; Machine learning taggers-MEC,HMM,CRF,

Module No. 3	Parsing Structure in Text	10 Hours

Shallow vs Deep parsing, Approaches in parsing, Types of parsing- Regex parser, Dependency parser, Constituency Parsing

Meaning Representation: Logical Semantics, Semantic Role Labelling, Distributional Semantics

Discourse Processing: Anaphora and Coreference Resolution

Module No. 4	NLP Using Deep Learning	6 Hours

Types of learning techniques, Chunking, Information extraction & Relation Extraction, Recurrent neural networks, LSTMs/GRUs, Transformers, Self-attention Mechanism, Sub-word tokenization, Positional encoding,

Module No. 5	Web Crawling and Social Media Mining	6 Hours

Web crawler – Writing first crawler–Data flow in Scrapy–Scrapy shell. Social Media Mining-Data Collections, Data Extraction, Geo visualization.

Module No. 6	NLP latest Techniques and applications	9 Hours

Contextualized word embeddings: ELMo, BERT, GPT,

Pre-trained Language Models (PLMs): BERT, GPT, ELMo, Large Language Models (LLMs), Applications of NLP: Transforming text, Sentiment Analysis, Information retrieval, text summarization, Question and Answering, Automatic Summarization

Text Books

1. Daniel Jurafsky and James H. Martin. 2024. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models, 3rd edition.
2. Daniel Jurafsky, James H.Martin. Speech and Language Processing, An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition, Pearson, 2nd Edition, January 2013.
3. Nitin Hardeniya, Jacob Perkins, Deepti Chopra, Nisheeth Joshi, Iti Mathur, "Natural Language Processing: Python and NLTK", Packt publisher, 2016.

References

1. Steven Bird, Ewan Klein, Edward Loper, "Natural Language Processing with Python", O'Reilly, 1st Edition 2009.
2. Jacob Perkins, "Python 3 Text Processing with NLTK 3 Cookbook", Pearson Education, Second Edition, 2014.
3. Deepti Chopra, Nisheeth Joshi, Iti Mathur, " Mastering Natural Language Processing with Python", Packt ,2016.

LabExercises

1. Read the paragraph and obtain the frequency of words.
2. Read the content from a web page and extract the tokens /expression/word/number.
3. Read only the word content from webpage and plot their frequency?
4.Write a program to slit sentences in a document?
5. Perform tokenizing and stemming by reading the input string?
6. Remove the stopwords and rarewords in the document?
7. Identify the parts of speech in the document?
8. Implement the N-gram tagger?
9.Implement Regex tagger?
10. Implement Brill tagger?
11. Implement Maximum Entropy Classifier?
12.Implement NER tagger?
13. Write a tagger that tags Date and Money expressions?
14. Define a grammar and obtain the sentences from the grammar?
15. Implement Regex parser?
16. Implement chunking using Shallow parsing?
17.Obtain the Named entity relations in the document?
18. Choose any news article with only contents of the news dumped into a text file and obtain the top 10 sentences?
19. Implement a text classification application?
20. Implement a text clustering application?
21. Implement a web crawler?
22. write an application for social media mining?
23. Identify and visualize the Facebook influencer ?
24. Develop a text analyser from the IPL web site page for 2021
25. Build a sentiment analyzer using LSTM.
26. Text Summarization: Take an RSS feed from an online newspaper. Extract the news articles and create a Summary for each news article.

Course Type	**Embedded Theory and Lab (ETL)**

Mode of Evaluation	**Theory.** **75%** • Continuous Assessment Test-1 15% • Continuous Assessment Test-2 15% • Digital Assessment/Quizes. 30% • Final Assessment Test-3 40% **Laboratory** **25%**
Modified by	**Mr. Gundimeda Venugopal** **Dr. Beebi Naseeba.** **Dr. Visalakshi**
Recommended by the Board of Studies on	**12th BoS, 29.04.2023**
Date of Approval by the Academic Council	**10th Academic Council, 01.06.2023**