GUJARAT TECHNOLOGICAL UNIVERSITY

Program Name: Bachelor of Engineering
Level : UG
Subject Code: 3174602
Branch: Computer Science and Engineering (Data Science)
Semester: VII
Subject Name: Natural Language Processing

Type of Course: Elective
Prerequisite: Probability and statistics, Programming and data structures

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			Credit	ExaminationMarks				Total Marks
L	T	P	C	TheoryMarks		PracticalMarks		
				ESE(E)	PA(M)	ESE(V)	PA(I)	
4	0	2	5	70	30	30	20	150

Course Content:

Sr.No.	Content	Total Hrs
1	**Introduction to NLP:** WhatisNLP?WhyNLPisDifficult?HistoryofNLP,AdvantagesofNLP,Disadvantages of NLP,Components of NLP, Applications of NLP,How to build an NLP pipeline? Phases of NLP, NLP APIs, NLP Libraries, Empirical Laws: Zipf's law, Heaps law	8
2	**Text Processing:** Text Processing Basics, Preprocessing: Lower case conversion, Stop words removal, Punctuation and digits removal, Stemming and lemmatization, Vector Space Models, Statistical Properties of words	4
2	**Language Modeling and Part of SpeechTagging:** Unigram Language Model, Bigram, Trigram, N-gram, Advanced smoothing for language modeling, Empirical Comparison of Smoothing Techniques, Applications of Language Modeling,NaturalLanguageGeneration,PartsofSpeechTagging,Morphology,NamedEntity Recognition	12
3	**Words and Word Forms:** Bagofwords,skip-gram,ContinuousBag-Of-Words,Embeddingrepresentationsforwords Lexical Semantics, Word Sense Disambiguation, Knowledge-Based and Supervised Word Sense Disambiguation	8

GUJARAT TECHNOLOGICAL UNIVERSITY

Program Name: Bachelor of Engineering

Level : UG

Subject Code: 3174602

Branch: Computer Science and Engineering (Data Science)

Semester: VII

Subject Name: Natural Language Processing

4	**Text Analysis, Summarization and Extraction:** Sentiment Mining, Text Classification, Text Summarization, Information Extraction, Relation Extraction,QuestionAnsweringinMultilingualSetting,NLPinInformationRetrieval,Cross-LingualIR	**12**
5	**Machine Translation:** Need of MT, Problems of Machine Translation, MT Approaches, Direct Machine Translations, Rule-Based Machine Translation, Knowledge-Based MT System, Statistical Machine Translation(SMT),Parameter learning in SMT (IBM models) using EM), Encoder-decoder architecture Neural Machine Translation	**10**
6	**Topic Models:** Introduction to topic models, Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA), Gibbs sampling for LDA	**06**

Suggested Specification table with Marks(Theory):

Distribution of Theory Marks					
RLevel	ULevel	ALevel	NLevel	ELevel	CLevel
14	14	14	14	7	7

Legends:R: Remembrance;U:Understanding;A:Application,N:AnalyzeandE:EvaluateC: Create and above Levels (Revised Bloom's Taxonomy)

Course Outcomes:

Sr. No.	**CO statement**	**Marks % weightage**
CO-1	Understand comprehend the key concepts of NLP and identify the NLP challenges, basic text preprocessing techniques, and core linguistic components involved in NLP systems	20
CO-2	Develop Language Modeling for various text corpora cross the different languages	20
CO-3	Illustrate computational methods to understand language phenomena of word sense disambiguation	15
CO-4	Design and implement NLP applications such as sentiment analysis, information extraction, summarization, classification, and multilingual question answering.	20
CO-5	Analyze and apply machine translation approaches and topic modeling techniques for multilingual text understanding.	25

GUJARAT TECHNOLOGICAL UNIVERSITY

Program Name: Bachelor of Engineering
Level : UG
Subject Code: 3174602
Branch: Computer Science and Engineering (Data Science)
Semester: VII
Subject Name: Natural Language Processing

List of Experiments: Practical work will be based on the above syllabus with minimum 10 experiments to be performed.

Reference Books Recommended: -

1. Jurafsky, D., & Martin, J. H. (2023). *Speech and Language Processing.* Pearson. [Available online: https://web.stanford.edu/~jurafsky/slp3/]
2. Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.
3. Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly.
4. Eisenstein, J. (2019). *Introduction to Natural Language Processing*. MIT Press.
5. Goyal, P., Pandey, S., & Jain, K. (2018). *Deep Learning for Natural Language Processing*. Apress.
6. Bhattacharyya, P. (2015). *Machine Translation*. CRC Press.
7. Aggarwal, C. C. (2018). *Text Mining and Analytics*. Springer.

List of e-Learning Resources:

1. https://www.kaggle.com/code/faressayah/natural-language-processing-nlp-for-beginners
2. https://nlp.stanford.edu/
3. https://www.analyticsvidhya.com/blog/2019/07/how-get-started-nlp-6-unique-ways-perform-tokenization/
4. https://www.tutorialspoint.com/natural_language_processing/index.htm
5. https://www.geeksforgeeks.org/natural-language-processing-overview/
6. https://nptel.ac.in/
7. https://www.coursera.org/

****