DISCIPLINE SPECIFIC ELECTIVES (DSE) COURSES OFFERED BY THE DEPARTMENT

DISCIPLINE SPECIFIC ELECTIVES (DSE-1)

CREDIT DISTRIBUTION, ELIGIBILITY AND PRE-REQUISITES OF THE COURSE

Course title & Code	Credits	Credit distribution of the course			Eligibility criteria	Pre-requisite of the course (if any)
		Lecture	Tutorial	Practical/ Practice		
Natural Language Processing ELDSE8A	**4**	**3**	**-**	**1**	**Class XII passed with Physics + Mathematics/Applied Mathematics + Chemistry OR Physics + Mathematics/Applied Mathematics + Computer Science/Informatics Practices**	**-**

Learning Objectives

This course introduces the student to the fundamental understanding of Natural Language Processing (NLP) which is a rapidly developing field with broad applicability throughout the hard sciences, social sciences, and the humanities. This course is intended as a theoretical and methodological introduction to the most widely used and effective current techniques, strategies and toolkits for natural language processing, with a primary focus on those available in the Python programming language.

Learning outcomes

On successful completion of this course, student will be able to:

CO1 Analyze the natural language text.

CO2 Define the importance of natural language.
CO3 Understand the concepts of Text mining.
CO4 Illustrate information retrieval techniques.
CO5 Analyze the natural language text.

SYLLABUS OF ELDSE-8A **Total Hours- Theory: 45 Hours, Practicals: 30 Hours**

Unit I: (11 Hours)
Overview and Language Modeling:
Origins and challenges of NLP-Language, Phases and components of NLP, Applications-Information Retrieval, Unigram Language Model, Bigram, Trigram, N-gram, Advanced smoothing for language modelling, Empirical comparison of smoothing techniques, Applications of Language Modelling.

Unit II: (12 Hours)
Part of Speech and Word Form:
Natural Language Generation, Parts of Speech Tagging, Morphology, Named Entity Recognition, Rule-base and Stochastic POS tagger, Markov Model, Maximum Entropy model, Bag-of-Words, skip-gram, Continuous Bag-of-Words, Embedding representations for words Lexical Semantics, Word Sense Disambiguation, Knowledge-Based and Supervised Word Sense Disambiguation.

Unit III: (11 Hours)
Text Analysis, Summarization and Extraction:
Text Summarization – Extraction and Abstraction, Information Extraction - Tokenization, Named Entity Recognition, Relation Extraction, Information Retrieval, Stop-Word, Stemming, Term weighting, Term Frequency, Document Frequency, Document Frequency Weighting (TFIDF), Text Classification (TF-IDF/Term Frequency Technique), Sentiment Mining.

Unit IV: (11 Hours)
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.

Practical component (if any) – Natural Language Processing Lab
(Python/MATLAB)

Learning outcomes

At the end of this course, Students will be able to

CO1 To experiment with the concepts introduced in the course Natural Language Processing.
CO2 Ability to program various techniques of NLP.
CO3 Design and develop applications for text or information extraction/summarization/classification

LIST OF PRACTICALS (Total Practical Hours- 30 Hours)

1. Perform sentence tokenization to break a text paragraph into individual sentences.
2. Perform word tokenization to break a text paragraph into individual words.
3. For the text selected in Practical 1, remove stop words and punctuation marks.
4. Apply the stemming technique to the text document selected in Practical 1 to obtain root words.
5. Perform different forms of lemmatization on the text document selected in Practical 1 to obtain base forms of words.
6. Extract the top 10 most common words in the selected text, excluding stop words.
7. Extract nouns and pronouns from the text and calculate similarities between any two words using a suitable method.
8. Case Study – Sentiment Analysis: Students will preprocess a text dataset (e.g., movie reviews or tweets) using tokenization, stemming, and feature extraction (TF-IDF or word embeddings). They will build and evaluate a sentiment classification model (e.g., Logistic Regression or Naive Bayes) and analyze its performance using metrics - Accuracy and F1-score.
9. Case Study - Language identification: Students will work with a multilingual dataset to preprocess text and extract features using character or word-level n-grams. They will train a language classification model (e.g., Naive Bayes or Random Forest) to identify the language of text samples and evaluate it with a confusion matrix and accuracy metrics.

Note: Students shall sincerely work towards completing all the above listed practicals for this course. In any circumstance, the completed number of practicals shall not be less than eight.

Essential/recommended readings

1. Daniel Jurafsky and James H Martin, “Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition”, 2nd Edition, Prentice Hall, 2013.
2. James Allen, “Natural Language Understanding”, 2nd edition, Benjamin/Cummings publishing company, 1995.
3. Eisenstein, J. (2019). Introduction to Natural Language Processing, MIT Press.

Suggestive readings

1. Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit, Steven Bird, Ewan Klein, and Edward Loper.

Note: Examination scheme and mode shall be as prescribed by the Examination Branch, University of Delhi, from time to time.