PROPOSED SYLLABUS

DISCIPLINE SPECIFIC ELECTIVE COURSE: Natural Language Processing

Credit distribution, Eligibility and Pre-requisites of the Course

Course title & Code	Credits	Credit distribution of the course			Eligibility	Prerequisite of the course (if any)
		Lecture	Tutorial	Practical/ Practice		
DSE: Natural Language Processing	4	3	0	1	Pass in Class XII	Machine Learning

Course Objectives:

The objectives of this course are:

1. To introduce foundational understanding in natural language
2. To understand the principles and methods of statistical, neural and transformer based natural language processing
3. To develop an in-depth understanding of the algorithms available for the processing and analysis of natural languages
4. To perform analysis of textual data and find useful patterns from the data

Course Learning Outcomes

On successful completion of the course, students will be able to:

1. Grasp the significance of natural language processing in solving real-world problems
2. Preprocess and Analyze text using formal techniques.
3. Apply machine learning techniques used in NLP
4. Understand approaches to syntactic and semantic analysis in NLP
5. Gain practical experience of using NLP toolkits

Syllabus

Unit 1 **(12 Hours)**

Introduction, Basic Text Processing and NLP Paradigms: Knowledge in Speech and Language Processing, NLP Applications, The problem of ambiguity, Regular Expressions, Text Normalization, Tokenization, Stemming, Lemmatization, Stop-word removal, Part-of-Speech Tagging, Named Entities and Named Entity Tagging/Recognition

Unit 2 **(8 Hours)**

Formal Language Modeling: Byte Pair Encoding and Edit Distance, Unigrams, Bigrams, N-grams, Markov assumption, Maximum likelihood estimation, Types of LMs: statistical vs neural vs transformer-based, N-gram Language Models, Smoothing and Perplexity

Unit 3 **(7 Hours)**

Vector Semantics and Embedding: Words and Vectors, Bag-of-Words (BoW), TF-IDF, Learning Word Embeddings - Word2vec

Unit 4 **(18 Hours)**

Deep Learning and Large Language Models for NLP: Feedforward Neural Networks, RNN, LLMs and Transformers, Limitations of static embeddings, Contextual representation based LM vector embeddings (BERT, RoBERTa, GPT), Applications of NLP - Text classification, Sentiment Analysis, Ethical and safety considerations with Language Models

References

1. Daniel Jurafsky and James H. Martin: *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*: 3rd Edition, Pearson, 2026 (Draft version available online at https://web.stanford.edu/~jurafsky/slp3/ed3book_jan26.pdf)
2. Dipanjan Sarkar. Text Analytics with Python: A Practitioner's Guide to Natural Language Processing. Second Edition. APress. 2019
3. Steven Bird, Ewan Klein, and Edward Loper . *Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit*, 2022 (Available online at https://www.nltk.org/book/)

Additional References

1. Yoav Goldberg. *A Primer on Neural Network Models for Natural Language Processing, 2022* (Available online at https://u.cs.biu.ac.il/~yogo/nnlp.pdf)
2. Jacob Eisenstein. Introduction to natural language processing. MIT Press, 2019.
3. Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

Suggested Practical List:

Python Packages like Scikit (SKLearn), NLTK, spaCy, gensim, PyTorch, transformers (HuggingFace) etc. may be used for programming

1. Prepare/Pre-process a text corpus to make it more usable for NLP tasks using tokenization, conversion to lowercase, removal of punctuation, filtration of stop words, stemming and lemmatization. (3 hours)
2. Use regex patterns to extract the usernames from the email addresses, hashtags, dates, and phone numbers present in a given text. (2 hours)
3. List the most common words (with their frequency) in a given text excluding stopwords. (2 hours)
4. Create the TF-IDF (Term Frequency -Inverse Document Frequency) Matrix for the given set of text documents (3 hours)
5. Build a simple statistical language model: estimate unigram and bigram probabilities with add-one smoothing and compute the probability of given sentences. (4 hours)
6. Perform POS tagging in a given text file. Extract all the nouns present in the text. Create and print a dictionary with frequency of parts of speech present in the document. (2 hours)
7. Identify and print the named entities using Name Entity Recognition (NER) for a collection of news headlines. (2 hours)
8. Classify movie reviews as positive or negative from the IMDB movie dataset of 50K movie reviews. (4 hours)

   (Link for dataset:
   https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-review)

9. Build and train a text classifier for the given data (using textbob or simpletransformers or keras library), (4 hours)
10. Generate text using a character-based model using an appropriate dataset. Given a sequence of characters from a given data (eg "Shakespear"), train a model to predict the next character in the sequence ("e"). (4 hours)