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 criteria	Pre-requisite of the course (if any)
		Lecture	Tutorial	Practical/ Practice		
DSE8b: Natural Language Processing	4	3	0	1	Pass in Class XII	Machine Learning

Course Objective

The objectives of this course are:

1. To introduce foundational understanding in natural language
2. To understand the principles and methods of statistical 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 statistical 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 mathematical techniques.
3. Apply machine learning techniques used in NLP - HMM, RNN
4. Understand approaches to syntax and semantics analysis in NLP
5. Gain practical experience of using NLP toolkits

Syllabus

Unit 1 Introduction and Basic Text Processing: Knowledge in Speech and Language Processing, The problem of ambiguity, Typical NL Tasks, Tokenization, Stemming, Lemmatization, Stop-word removal

Unit 2 Formal Language Modeling: Regular Expressions, Text Normalization, and Edit Distance, Unigrams, Bigrams, N-grams, N-gram Language Models, Smoothing and Entropy

Unit 3 Sequence Labeling for Parts of Speech Tagging: Part-of-Speech Tagging, Named Entities and Named Entity Tagging/Recognition, Hidden Markov Model (Forward and Viterbi algorithms and EM training)

Unit 4 Vector Semantics and Embedding: Lexical Semantics, Vector Semantics, Words and Vectors, TF-IDF: Weighing terms in the vector and its applications, Learning Word Embeddings - Word2vec and Gensim, Vector Space Models

Unit 5 Applications of Text Mining: Text classification, Sentiment Analysis

Unit 6 Deep Learning Models for NLP: Feedforward Neural Networks, Recurrent Neural Networks, and LSTM

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, 2022.
2. Christopher D. Manning and Hinrich Schütze *Foundations of Statistical Natural Language Processing*, MIT Press, 1999.
3. Steven Bird, Ewan Klein, and Edward Loper *Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit*, 1st edition, O'Reilly Media, 2009.

Additional Reference

(i) Yoav Goldberg *A Primer on Neural Network Models for Natural Language Processing, 2022.*

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, filtration of stop words, removal of punctuation, stemming and lemmatization.
2. List the most common words (with their frequency) in a given text excluding stopwords.
3. Extract the usernames from the email addresses present in a given text. .
4. 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. Find the similarity between any two text documents
5. Perform dependency analysis of a text file and print the root word of every sentence.
6. Create the TF-IDF (Term Frequency -Inverse Document Frequency) Matrix for the given set of text documents
7. Extract all bigrams , trigrams using ngrams of nltk library
8. Identify and print the named entities using Name Entity Recognition (NER) for a collection of news headlines.
9. Find the latent topics in a document using any LDA and display top 5 terms that contribute to each topic along with their strength. Also visualize the distribution of terms contributing to the topics.
10. Classify movie reviews as positive or negative from the IMDB movie dataset of 50K movie reviews. (Link for dataset:

    https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews)

0. Build and train a text classifier for the given data (using textbob or simpletransformers library)

0. Generate text using a character-based RNN 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").