COURSE: NATURAL LANGUAGE PROCESSING (CS5201)

Faculty Name: Dr. Dheeraj Kodati
Credits: 3
Department/Centre/School: Ecole Centrale School of Engineering
Course Level: M.Tech
Branch: AI & DS, Year: 2025, Semester: II

Course Objectives:

CO1: Understand and apply foundational text processing techniques and linguistic structures for analyzing diverse types of textual data, including web and social media sources.

CO2: Demonstrate the ability to implement statistical and probabilistic models such as TF-IDF, Word Sense Disambiguation, and topic modeling for semantic understanding and information retrieval.

CO3: Evaluate and employ traditional and modern machine learning models—including supervised, unsupervised, and semi-supervised approaches—for text classification and clustering using appropriate feature extraction methods.

CO4: Analyze and implement deep learning models like LSTM, GRU, CNN, and sequence-to-sequence architectures to capture contextual dependencies in natural language tasks.

CO5: Design and experiment with transformer-based models and multitask learning strategies to address complex NLP problems involving attention mechanisms, pretrained language models, and task sharing techniques.

Detailed Course Syllabus:

Module 1: Foundations of Text Processing and Linguistic Structure

1.1 Introduction and overview of NLP, Data Extraction and Collection: types of datasets, sources (web, APIs, social media), web scraping (BeautifulSoup, Scrapy), handling large text data, handling noisy and imbalanced data, Preprocessing: Tokenization, stemming and lemmatization, stopword removal and punctuation handling, n-gram, Regex, POS Tagging, and NER.

Module 2: Statistical NLP

1.2 Topics in Information Retrieval: Page Rank algorithm, vector space model, vector similarity, TF-IDF, BOW, CBOW, skip-gram, Annotating Linguistic Structure: Lexical analysis- word lexicons, word net, collocations, syntactic structure, dependency parsing, probabilistic context free grammars, BPE, text summarization-extractive and abstractive and multi-document text summarization.

1.3 Semantic relations: Word Sense Disambiguation; supervised, unsupervised and semi-supervised approaches, Lesk algorithm, Zipf’s Law, WordNet and WordNet based similarity measures, perplexity, EM algorithm, distributional measures of similarity, statistical machine translation, concept mining using latent semantic analysis.
Packages: BeautifulSoup, Scrapy ,NLTK, spaCy, networkx and Rake.

Module 3: Classical Modeling and Feature Extraction

2.1 Text classification, maximum entropy modeling, and clustering, traditional machine learning methods such as logistic regression, decision trees, ensemble methods, random forests, naïve bayes, svm for NLP, Supervised, Semi-Supervised and Unsupervised methods for handling NLP tasks.

2.2 What is a word embedding, Understanding Static Embeddings: Word2Vec, GloVe, Doc2vec. Topic modeling, LDA, HMM Packages: scikit-learn, Gensim.

Module 4: Sequence Modeling

3.1 Sequential dependencies, contextual structures, convolutional networks, recurrent networks: LSTM, BiLSTM, GRU, BiGRU, autoencoders and variational autoencoders.

3.2 Sequence-to-sequence, machine translation system, optimizers, hyperparameter tuning, Packages: Keras and TensorFlow.

Module 5: Transformers and PLMs for NLP

Foundations of attention-based models, including the Transformer architecture, positional encoding, encoder-decoder models, self-attention, multi-head attention and training. pretrained language models, such as BERT and RoBERTa. Packages: Keras, PyTorch and Hugging Face Transformers.

Module 6: Multitask Learning for NLP

Hard and soft parameter sharing, task balancing, auxiliary task selection, and transfer learning connections, Some applications of MTL. Packages: Hugging Face Transformers.

Suggested Readings and References:

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. Chris Manning and Hinrich Schütze. Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA: May 1999.
3. Nitin Indurkhya and Fred J Damerau, “Handbook of natural language processing,” Chapman and Hall/CRC, 2010.
4. Research papers discussed during the lectures.

Evaluation policy of the course:

Components of Course Evaluation	Percentage Distribution
Assignment+ Viva	5
Minor I	30
End-semester examination	65
Total	100