KESHAV MEMORIAL INSTITUTE OF TECHNOLOGY

(AN AUTONOMOUS INSTITUTE)
Accredited by NBA & NAAC, Approved by AICTE, Affiliated to JNTUH, Hyderabad

B. Tech. in COMPUTER SCIENCE AND ENGINEERING (AI & ML)
III Year I Semester Syllabus (KR23)

NATURAL LANGUAGE PROCESSING (23CM503PC)

Pre-requisites/ Co-requisites:

1. 23CM404PC – Deep Learning.

L	T	P	C
3	0	0	3

Course Objectives: The course will help to

1. To introduce fundamental concepts and applications of NLP.
2. To teach text preprocessing techniques and embedding methods.
3. To explore advanced NLP models like RNNs, LSTMs, and Seq2Seq.
4. To enable building, training, and evaluating NLP applications.
5. To impart best practices for developing and deploying NLP models.

Course Outcomes: After learning the concepts of this course, the student is able to

1. Understand and apply core NLP concepts and techniques.
2. Implement text preprocessing workflows for NLP tasks.
3. Build and train RNNs, LSTMs, and CNNs for text analysis.
4. Develop Seq2Seq models for translation and chatbot systems.
5. Deploy and optimize NLP applications using best practices.

UNIT-I:

Natural Language Processing: What is natural language processing, why NLP, How NLP is used, Building NLP applications, Development of NLP applications, Structure of NLP applications.

Your first NLP application - Introducing sentiment analysis, working with NLP datasets, Using Word Embeddings, Neural Networks, RNNs and Linear layers.

UNIT-II:

Word and document embeddings: Introducing embeddings, what are Word embeddings, why are they important, building blocks of language: Characters, words, and phrases, Tokenization, stemming, and lemmatization, Skip-gram and continuous bag of words (CBOW), Glove, fastText, Document-level embeddings, Visualizing embeddings.

A typical text processing workflow - Data Collection and labelling, collecting labelled data, Text normalization, tokenization, stop word removal, part-of-speech tagging, Vectorizing text, count-based vectorization, term frequency-Inverse document frequency, word vectors.

UNIT-III:

Sentence classification: Recurrent neural networks (RNNs) - Long short-term memory units (LSTMs) and gated recurrent units (GRUs) - Accuracy, precision, recall, and F-measure

Sequential labeling and language modeling: Introducing sequential labelling, using RNN to encode Sequences, building a part-of-speech tagger, Multilayer and bidirectional RNNs, named entity recognition, tagging spans, what is a language model, why are Language models useful, Training an RNN language model, Text Generation using RNNs, evaluating text using a language model, generating text using a language model.

UNIT-IV:

Sequence-to-sequence models: Introducing sequence-to-sequence models, machine translation, building your first translator, prepare the datasets, training the model, running the translator, how Seq2Seq models work, evaluating translation systems, Case study: Building a chatbot, introducing dialogue systems, preparing dataset, training and running a chatbot.

Convolutional neural networks: Introducing convolutional neural networks (CNNs), RNNs and their shortcomings, pattern matching for sentence classification, Convolutional Layers, Pattern matching using filters, ReLu, pooling layers, Case study: Text classification, using Cnnencoder, training and running the classifier.

UNIT-V:

Best practices in developing NLP applications: Batching instances, padding, sorting, masking, Tokenization for neural models - Avoiding overfitting, dealing with imbalanced datasets, Hyperparameter tuning.

Deploying and serving NLP applications: Architecting your NLP application - Deploying your NLP model - Case study: Serving and deploying NLP applications - Interpreting and visualizing model predictions.

Textbooks:

1. Real-World Natural Language Processing- MASATO HAGIWARA, Manning Publications Co.
2. Advanced Natural Language Processing with TensorFlow2- Ashish Bansal, Packt Publishing.