DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
NATIONAL INSTITUTE OF TECHNOLOGY PATNA
Ashok Raj Path, PATNA 800 005 (Bihar), India
Phone No.: 0612 – 2372715, 2370419, 2370843, 2371929, 2371930, 2371715 Fax – 0612- 2670631 Website: www.nitp.ac.in

CSXX1915 ***Text Analytics*** **L-T-P-Cr: 2-0-2-3**

Pre-requisites: The students are expected to be fluent in computational and mathematical models and should have a basic knowledge of probabilities and calculus. Students are also expected basic knowledge about machine learning from courses in artificial intelligence, probabilistic graphical models: principles and techniques, machine learning or deep learning.

Objectives/Overview:

- The student has a general understanding of the current state-of-the art in Text analytics.
- The student will be able to make models to process text data to extract information
- The student with be able to think critically about work in the field.
- Provide hands-on experience on one of the most exciting areas of research in Text data mining.

Course Outcomes – After completing this course, students should be able to:

CO-1. *Developing* an advanced understanding of text data and its representation.

CO-2. *Identifying* scenarios where text data is being used and make models to do so for real-world problem-solving.

CO-3. *Building* a collection of different models and approaches to text data and understanding their various strengths and weaknesses.

CO-4. *Implement* a range of text representation models to classify and cluster texts.

CO-5. *Design* and *develop* problem-solving skills by tackling challenges and complexities in the practical implementation of text data.

Course Outcomes–Cognitive Levels–Program Outcomes Matrix –

[H: High relation (3); M: Moderate relation (2); L: Low relation (1)]

Course Outcomes	Program Outcomes											
	PO-1 (Engineering knowledge)	PO-2 (Problem analysis)	PO-3 Design/development of solutions)	PO-4 (Conduct investigations of complex problems)	PO-5 (Modern tool usage)	PO-6 (The engineer and society)	PO-7 (Environment and sustainability)	PO-8 (Ethics)	PO-9 (Individual and team work)	PO-10 (Communication)	PO-11 (Project management and finance)	PO-12 (Life-long learning)
CO-1	3	3	3	3	2	3			3	3	1	3
CO-2	3	3	3	3	2	3		1	3	3	1	3
CO-3	3	3	3	3	3	3	1	1	3	3	1	3
CO-4	3	3	3	3	2	3	1		3	3	1	3
CO-5	3	3	3	3	3	3	2	1	3	3	1	3
CO-6	3	3	2	1	3	1	1	1	3	3	2	2

UNIT 1: **Lecture: 6**

Introduction: Main Tasks of Text Data Mining, Challenges in Text Data Mining

Data Annotation and Preprocessing: Data Acquisition, Data Preprocessing, Data Annotation, Basic Tools of NLP, Tokenization and POS Tagging, , Syntactic Parser, N-gram Language Model

Text Representation: Vector Space Model, Distributed Representation of Words, Distributed Representation of Phrases, Distributed Representation of Sentences, Distributed Representation of Documents

UNIT 2 **Lecture: 6**

Text Representation with Pretraining and Fine-Tuning: ELMo: Embeddings from Language Models, GPT: Generative Pretraining, BERT: Bidirectional Encoder Representations from Transformer

UNIT 3 **Lecture: 6**

Text Classification: Feature Selection, Traditional Machine Learning Algorithms for Text Classification (Naïve Bayes, Logistic/Softmax and Maximum Entropy, Support Vector Machine, Ensemble Methods), Deep Learning Methods, Evaluation of Text Classification

Text Clustering: Text Similarity Measures, Similarity Between Documents, Similarity Between Clusters, Text Clustering Algorithms (K-Means Clustering, Single-Pass Clustering, Hierarchical Clustering, Density-Based Clustering), Evaluation of Clustering

UNIT 4 **Lecture: 6**

Topic Model: Latent Semantic Analysis, Probabilistic Latent Semantic Analysis, Latent Dirichlet Allocation

Sentiment Analysis and Opinion Mining: Sentiment Analysis Tasks, Methods for Document/Sentence-Level Sentiment Analys, Word-Level Sentiment Analysis and Sentiment Lexicon Construction, Aspect-Level Sentiment Analysis, Issues in Sentiment Analysis

UNIT 5 **Lecture: 4**

Information Extraction, Named Entity Recognition,

Automatic Text Summarization: Text Summarization, Extraction-Based Summarization, Compression-Based Automatic Summarization, Abstractive Automatic Summarization, Query-Based Automatic Summarization, Cross lingual and Multilingual Automatic Summarization, Summary Quality Evaluation Methods

Text Book:

Chengqing Zong, Rui Xia, Jiajun Zhang, Text Data Mining, Springer

Reference Book:

Charu C Aggarwal, Machine Learning for Text, Springer