|  |  |
| --- | --- |
| MONO | **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](http://www.nitp.ac.in/) |

***CSX448 Text Mining***

**L-T-P-Cr: 3-0-0-3**

**Pre-requisites:** None.

**Objectives/Overview:**

* To increase student awareness of the power of large amount of text data and computational methods to find patterns in large text corpora.

**Course Outcomes:**

At the end of the course, a student should:

|  |  |  |
| --- | --- | --- |
| **Sl. No.** | **Outcome** | **Mapping to POs** |
| 1 | Have knowledge on text mining | PO1 |
| 2 | Identify the required features commonly generated from text. | PO1, PO12 |
| 3 | Understand techniques for transforming text in to numerical vectors. | PO1, PO3 |
| 4 | Choose appropriate technologies for specific text analysis tasks and evaluate the benefit and challenges of the chosen technical solution. | PO1, PO2, PO3 |
| 5 | Apply machine learning approaches for information extraction. | PO1, PO3, PO12, PO6 |

**UNIT-I Lectures: 04**

Overview of Text Mining: What’s Special about Text Mining, What Types of Problems Can Be Solved, Document Classification, Information Retrieval, Clustering and Organizing, Information Extraction, Prediction and Evaluation.

**UNIT-II Lectures: 08**

From Textual Information to Numerical Vectors – Collecting Documents, Document Standardization, Tokenization; Lemmatization - Inflectional Stemming, Stemming to a Root; Vector Generation for Prediction - Multiword Features, Labels for the Right Answers, Feature Selection by Attribute Ranking; Sentence Boundary Determination, Part-of-Speech Tagging, Word Sense Disambiguation, Phrase Recognition, Named Entity Recognition, Parsing , Feature Generation.

**UNIT-III Lectures: 08**

Using Text for Prediction - Recognizing that Documents Fit a Pattern, How Many Documents Are Enough?, Document Classification; Learning to Predict from Text - Similarity and Nearest-Neighbour Methods; Document Similarity, Decision Rules, Decision Trees, Scoring by probabilities, Linear Scoring Methods; Evaluation of Performance - Estimating Current and Future Performance, Getting the Most from a Learning Method, Errors and Pitfalls in Big Data Evaluation; Applications, Graph Models for Social Networks.

**UNIT-IV Lectures: 08**

Information Retrieval and Text Mining - Is Information Retrieval a Form of Text Mining?, Key Word Search, Nearest-Neighbor Methods; Measuring Similarity- Shared Word Count, Word Count and Bonus, Cosine Similarity; Web-Based Document Search – Link Analysis; Document Matching, Inverted Lists, Evaluation of Performance.

**UNIT-V Lectures: 06**

Finding Structure in a Document Collection - Clustering Documents by Similarity; Similarity of Composite Documents – K-Means Clustering, Hierarchical Clustering, The EM Algorithm.; What Do a Cluster’s Labels Mean?, Applications , Evaluation of Performance.

**UNIT-VI Lectures: 08**

Looking for Information in Documents - Goals of Information Extraction; Finding Patterns and Entities from Text - Entity Extraction as Sequential Tagging, Tag Prediction as Classification, the Maximum Entropy Method, Linguistic Features and Encoding, Local Sequence Prediction, Global Sequence Prediction Models; Coreference and Relationship Extraction - Coreference Resolution, Relationship Extraction; Template Filling and Database Construction.

**Text/ Reference Book:**

1. Weiss, S. M., Indurkhya, N., Zhang, T. (2010). Fundamentals of Predictive Text Mining. Springer: New York. ISBN: 978-1849962254.
2. Pang, B. and Lee, L. (2008). Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval, 2(1-2): 1–135. Available online at:http://www.cs.cornell.edu/home/llee/opinion-mining-sentiment-analysis-survey.html
3. Manning, C. D., Raghavan, P., and Schutze, H. (2008). Introduction to Information Retrieval, Chapters 6 and 13-18, Cambridge University Press. Available online at: http://nlp.stanford.edu/IR-book/