



INSTITUTE OF AERONAUTICAL ENGINEERING

(Autonomous)

Dundigal - 500 043, Hyderabad, Telangana

COURSE CONTENT

TEXT MINING								
VII Semester: IT								
Course Code	Category	Hours / Week			Credits	Maximum Marks		
AITD25	Elective	L	T	P	C	CIA	SEE	Total
		3	-	-	3	40	60	100
Contact Classes: 48	Tutorial Classes: Nil	Practical Classes: Nil			Total Classes: 48			
Prerequisite: Probabilistic Modeling and Reasoning, Foundations of Machine Learning								

I. COURSE OVERVIEW:

Text mining, also known as text data mining or text analytics, involves deriving useful information and insights from unstructured text data. This course introduces students to the techniques and tools used to extract knowledge from textual data, with applications in areas such as information retrieval, natural language processing (NLP), sentiment analysis, and machine learning.

II. COURSE OBJECTIVES:

The students will try to learn

- Understand the basic concepts of text mining, including its definition, importance, and applications.
- Learn about various types of text data (e.g., documents, social media content, and web data).
- Explore the differences between structured and unstructured data.

III. COURSE OUTCOMES:

At the end of the course, students should be able to:

- CO1 Understand the foundational concepts of text mining, including its significance, applications, and the difference between structured and unstructured data
- CO2 Apply various text preprocessing techniques like tokenization, stemming, lemmatization, stop word removal, and part-of-speech tagging to clean and prepare raw text data for analysis.
- CO3 Apply machine learning algorithms such as Naive Bayes, Support Vector Machines (SVM), and K-means clustering to classify and group text documents based on their content.
- CO4 Understand how to perform sentiment analysis to assess opinions or emotions expressed in text data, and apply these techniques to real-world data sources like social media and product reviews.
- CO5 Evaluate various techniques such as Latent Dirichlet Allocation (LDA) to identify the underlying topics in large collections of text and interpret the results.
- CO6 Apply text mining techniques to real-world applications, such as social media analysis, customer feedback analysis, and document categorization.

IV. COURSE CONTENT:

MODULE – I: INTRODUCTION (10)

Introduction to Text Mining, Understand the basic concepts of text mining, including its definition, importance, and applications. Explore the differences between structured and unstructured data. Learn about various types of text data (e.g., documents, social media content, and web data). Pre-processing Text Data: Learn text pre-processing techniques such as tokenization, stemming, lemmatization, stop word removal, and part-of-speech tagging. Explore methods for handling noisy data, such as correcting typos and handling punctuation. Understand the importance of data cleaning in the text mining pipeline.

MODULE – II: TEXT REPRESENTATION TECHNIQUES AND NATURAL LANGUAGE PROCESSING (NLP) FUNDAMENTALS (09)

Text Representation Techniques Gain knowledge of how to convert text into numerical formats (e.g., bag-of-words, TF-IDF, word embeddings). Study various vectorization techniques, such as Count Vectorizer, TF-IDF Vectorizer, and word embeddings like Word2Vec, GloVe, or BERT. Learn how to represent documents as feature vectors to facilitate analysis. Natural Language Processing (NLP) Fundamentals Get an overview of basic NLP concepts and techniques (e.g., tokenization, named entity recognition (NER), syntactic parsing). Learn to work with NLP libraries like NLTK, SpaCy, or Hugging Face Transformers.

MODULE – III: TEXT CLASSIFICATION AND SENTIMENT ANALYSIS (10)

Understand how to build text classifiers for tasks like spam detection, topic categorization, and sentiment analysis.

Learn about machine learning algorithms (e.g., Naive Bayes, SVM, neural networks) used for text classification. Explore how sentiment analysis can be performed on textual data to extract opinions, emotions, or sentiments.

MODULE-IV CLUSTERING AND TOPIC MODELING AND TEXT MINING FOR INFORMATION RETRIEVAL (10)

Clustering and Topic Modelling : Explore unsupervised learning techniques like K-means clustering and Latent Dirichlet Allocation (LDA) for discovering hidden patterns and topics in text data. Learn how to group similar documents or texts using clustering techniques. Study methods for topic modelling and uncovering latent topics in large text corpora. Text Mining for Information Retrieval : Understand the role of information retrieval systems in text mining. Learn about techniques like inverted indexing, ranking algorithms, and relevance scoring for search engines. Study on how to evaluate the effectiveness of text-based search engines and retrieval systems.

MODULE –V: TEXT MINING WITH DEEP LEARNING AND APPLICATIONS OF TEXT MINING (09)

Text Mining with Deep Learning Explore advanced techniques like Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and Transformer-based models for text mining tasks. Understand how deep learning models can be used for text classification, machine translation, and text generation. Applications of Text Mining Explore practical applications of text mining in fields like healthcare, marketing, finance, social media analysis, and customer feedback analysis. Study real-world case studies where text mining techniques are applied to solve problems such as customer sentiment analysis, fraud detection, and legal document review.

V.TEXT BOOKS:

1. Mining Text Data. Charu C. Aggarwal and Cheng Xiang Zhai, Springer, 2012.
2. Speech & Language Processing. Dan Jurafsky and James H Martin, Pearson Education India, 2000.

VI. REFERENCE BOOKS:

1. Introduction to Information Retrieval. Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schuetze, Cambridge University Press, 2007.

VII. WEB REFERENCES:

1. <http://hadoop.apache.org>
2. <http://spark.apache.org>
3. <http://graphlab.org/projects/index.htm>

VIII. MATERIALS ONLINE

1. Course template

2. Tutorial question bank
3. Tech talk topics
4. Open-ended experiments
5. Definitions and terminology
6. Assignments
7. Model question paper – I
8. Model question paper – II
9. Lecture notes
10. PowerPoint presentation
11. E-Learning Readiness Videos (ELRV)