 **BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI**

**WORK INTEGRATED LEARNING PROGRAMMES**

**COURSE HANDOUT**

**Part A: Content Design**

|  |  |
| --- | --- |
| **Course Title** | Text Mining |
| **Course Author** | Aruna Malapati |
| **Lead Instructor** | Dr. Chetana Gavankar |
| **Version No** | 2.0  (Modified by Dr. Chetana Gavankar) |
| **Version No** | 2 |
| **Date** | 20/05/2020 |

**Course Description**

|  |
| --- |
| In today’s world, 80% of the data generated by enterprises is unstructured or semi structured in the form of Emails, Surveys, Feedback etc. wherein most of the data is in the form of text. This course aims to equip students with adequate knowledge in extracting the relevant text data and skills to identify patterns therein. This course covers topics like converting documents to vectors using TF-IDF, Parts of Speech Tagging, Topic modelling using LDA, sentiment analysis and recommender systems. |

**Course Objectives**

|  |  |
| --- | --- |
| **No** | Objective |
| **CO1** | Convert documents into vectors using TF-IDF and compute similarities |
| **CO2** | Implement topics modelling using LDA |
| **CO3** | Apply sentiment analysis |
| **CO3** | Implement recommender systems |

**Text Book(s)**

|  |  |
| --- | --- |
| No | Author(s), Title, Edition, Publishing House |
| T1 | Introduction to Information Retrieval, Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schutze, Cambridge University Press. 2008 |

**Reference Book(s) & other resources**

|  |  |
| --- | --- |
| No | Author(s), Title, Edition, Publishing House |
| R1 | Mining of Massive Datasets , Anand Rajaraman, Jeff Ullman, Jure Leskovec |

**Content Structure**

|  |  |  |
| --- | --- | --- |
| **No** | **Title of the Module** | **References** |
| M1 | Information Retrieval   1. Introduction to Text Mining 2. Binary term incidence matrix 3. Information Retrieval Pipeline 4. Inverted Index Construction 5. Merge Algorithm and Query Optimization 6. Tolerant Retrieval using Normalization, Query expansion, Stemming, Lemmatization, Wild card query using K-Gram index 7. Ranked Retrieval using TF-IDF and Cosine score | T1 Ch 1,2,3,6 |
| M2 | Part of Speech Tagging   1. Introduction to Part of speech tagging 2. Part of speech tagging using HMM-1 3. Implementing POS Tagging in Python | Class Notes |
| M3 | Topic modelling using LDA   1. Mathematical foundations for LDA : Multinomial and Dirichlet distributions-1 2. Mathematical foundations for LDA : Multinomial and Dirichlet distributions-2 3. Intuition behind LDA 4. LDA Generative model 5. Probabilistic Graphical Models 6. Latent Dirichlet Allocation 7. Implementing LDA in Python | Class Notes |
| M4 | Introduction to Sentiment Analysis   1. Sentiment Analysis 2. Subjectivity Analysis 3. Topic Extraction 4. Product Reviews 5. Opinion Retrieval and Spam 6. Opinion Summarization 7. Implementing Sentiment Analysis in Python | Class Notes |
| M5 | Recommender Systems   1. Introduction to Recommender Systems 2. Collaborative filtering   2.1 User based Collaborative filtering  2.2 Item based Collaborative filtering  2.3 Matrix factorization using Singular Value Decomposition  2.4 Latent Factor Models   1. Metrics used for evaluating Recommender Systems 2. Implementing Recommender System in Python 3. Industry talk on application of Recommender Systems | R1 Ch 9 |
|  |  |  |

|  |  |
| --- | --- |
| **Week** | **Content / Assignments / Exercises** |
| Week1 | Video Content: M1  Evaluative Quiz: Nil  Exercises: Nil  Assignment: Nil |
| Week 2 | Video Content: M2  Evaluative Quiz: Q1 (M1 and M2 content)  Exercises/Assignments: Nil  Assignments: Nil |
| Week 3 | Video Content: M3  Minor Project (M1/ M2/ M3) |
| Week 4 | Video Content: M4  Assignment: (based on M4 content)  Implementation of Sentiment analysis in Python on a data set. |
| Week 5 | Video Content: M5  Evaluative Quiz: Q2 (based on M4 and M5 content) |
|  |  |

Evaluation

|  |  |  |
| --- | --- | --- |
| Evaluation Component | Marks | Type |
| Comprehensive Examination | 40% | Open |
| Quizzes (2) | 24% | Open |
| Minor Project | 24% | Open |
| Assignment | 12% | Open |