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

**WORK INTEGRATED LEARNING PROGRAMMES**

Part A: Content Design

|  |  |
| --- | --- |
| **Course Title** | Advanced Data Mining |
| **Course No(s)** | SS ZG548 |
| **Credit Units** | 4 |
| **Credit Model** |  |
| **Content Authors** | Dr. Kamlesh Tiwari |

**Course Objectives**

|  |  |
| --- | --- |
| **No** |  |
| **CO1** | To learn how to mine complex data (beyond conventional record data) and complex structures such as Tree/graph, sequence data, web/text data, stream data, mining multivariate time series data, high-dimensional data etc. |
| **CO2** | To learn how to apply these techniques to specific applications such as web search, Information Retrieval, social networks etc. |
| **CO3** | To learn about distributed computing solutions for data intensive applications in data mining |

**Text Book(s)**

|  |  |
| --- | --- |
| T1 |  |
| T2 |  |

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

|  |  |
| --- | --- |
| R1 | Tan P. N., Steinbach M & Kumar V. “***Introduction to Data Mining”*** Pearson Education, 2006 |
| R2 | Yates R. B. and Neto B. R. “***Modern Information Retrieval***” Pearson Education, 2005 |
| R3 | Han J. & Kamber M., “***Data Mining: Concepts and Techniques”,*** Morgan Kaufmann Publishers, Second Edition, 2006 |
| R4 | Christopher D.M., Prabhakar R. & Hinrich S. “***Introduction to Information Retrieval***” Cambridge UP Online edition, 2009 |
| R5 | Hadzic F., Tan H. & Dillon T. S. “***Mining data with Complex Structures***” Springer, 20 |
| R6 | Agarwal Charu C. (Ed) “**Data Streams Models and Algorithms**” Springer 2007 |
| R7 | Azure HDInsight <https://docs.microsoft.com/en-in/azure/hdinsight/> |
| R8 | Azure Data Science VM <https://docs.microsoft.com/en-us/azure/machine-learning/data-science-virtual-machine/overview> |
| R9 | Azure Cosmos DB https://docs.microsoft.com/en-us/azure/cosmos-db/introduction |

**Content Structure**

1. Introduction
   1. Review of data mining
   2. Objectives
   3. Overview
2. Incremental & Stream Data Mining
   1. Incremental Algorithms for Data Mining
   2. Characteristics of Streaming Data
   3. Issues and Challenges
   4. Streaming Data Mining Algorithms
   5. Executing Streaming Data Mining on HDInsight/Data Science VM
3. Distributed computing solutions for data mining
   1. MapReduce/Hadoop
   2. Spark
   3. Setting up Hadoop and Spark cluster on Azure HDInsight to perform Data mining tasks
4. Sequence Mining
   1. Characteristics of Sequence Data
   2. Problem Modeling
   3. Sequence Pattern Discovery
   4. Timing Constraints
   5. Performing Data Mining Algorithm on Azure Data Science VM
5. Text Mining
   1. Text Classification
   2. Vector Space Model
   3. Flat and Hierarchical Clustering
   4. Streaming Data Mining Algorithms
   5. Text Classification on HDInsight
6. Web Search
   1. Crawling & Indexing
   2. Hyperlink analysis
   3. HITS and Page Rank Algorithms

**6.3** Building Web Search using built-in library

1. Mining Complex Structures
   1. Mining Trees
      1. Tree Miner
      2. Tree Model Guided Framework
      3. TMG framework for mining ordered & unordered subtrees
   2. Mining Graphs
      1. Approaches to graph mining
      2. Building and Traverse Graph Database using Azure Cosmos DB
   3. Case Study: Information Retrieval
   4. Case Study: Mining Social Networks

**Learning Outcomes:**

|  |  |
| --- | --- |
| No | Learning Outcomes |
| LO1 | To understand how to update the patterns incrementally when the data is continuously coming |
| LO2 | To understand the role of distributed computing in data intensive data mining |
| LO3 | To study how to investigate the sequence data |
| LO4 | To understand how text mining is different from data mining and how to mine it |
| LO5 | To understand what goes into the web search and to study methods of web search and their improvements |
| LO6 | To understand how to mine complex structures other than records while retaining the relations among the entities |
| LO7 | Familiarity with tool used for Data mining and advance analytics on azure |

**Part B: Learning Plan**

|  |  |
| --- | --- |
| **Academic Term** | Second Semester 2017-2018 |
| **Course Title** | Advanced Data Mining |
| **Course No** | SS ZG548 |
| **Lead Instructor** | Dr. Kamlesh Tiwari |

Contact Hour 1

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Content Ref. | Topic Title | Study/HW Resource Reference |
| Pre CH |  | Introduction  Review and Overview |  |
| During CH |
| Post CH |

Contact Hour 2

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Content Ref. | Topic Title | Study/HW Resource Reference |
| Pre CH |  | Incremental Data Mining  Relook traditional algorithms | See Class Slides |
| During CH |
| Post CH |

Contact Hour 3

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Content Ref. | Topic Title | Study/HW Resource Reference |
| Pre CH |  | Incremental algorithms and their design and analysis | See Class Slides |
| During CH |
| Post CH |

Contact Hour 4

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Content Ref. | Topic Title | Study/HW Resource Reference |
| Pre CH |  | Incremental algorithms and their design and analysis | See Class Slides |
| During CH |
| Post CH |

Contact Hour 5

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Content Ref. | Topic Title | Study/HW Resource Reference |
| Pre CH |  | Incremental algorithms and their design and analysis | See Class Slides |
| During CH |
| Post CH |

Contact Hour 6

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Content Ref. | Topic Title | Study/HW Resource Reference |
| Pre CH | See Class Slides | Stream Data Mining Characteristics, Issues and Challenges | R6 Ch1,4 |
| During CH |
| Post CH |

Contact Hour 7

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Content Ref. | Topic Title | Study/HW Resource Reference |
| Pre CH | See Class Slides | Stream Data Mining Algorithms and their Comparison | R6 Ch1, 4 |
| During CH |
| Post CH |

Contact Hour 8

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Content Ref. | Topic Title | Study/HW Resource Reference |
| Pre CH | See Class Slides | Stream Data Mining Algorithms and their Comparison | R6 Ch1, 4 |
| During CH |
| Post CH |

Contact Hour 9

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Content Ref. | Topic Title | Study/HW Resource Reference |
| Pre CH | See Class Slides | Distributed computing solutions for data mining | See Class Slides |
| During CH |
| Post CH |

Contact Hour 10

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Content Ref. | Topic Title | Study/HW Resource Reference |
| Pre CH | See Class Slides | Distributed computing solutions for data mining | See Class Slides |
| During CH |
| Post CH |

Contact Hour 11

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Content Ref. | Topic Title | Study/HW Resource Reference |
| Pre CH | See Class Slides | Distributed computing solutions for data mining | See Class Slides |
| During CH |
| Post CH |

Contact Hour 12

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Content Ref. | Topic Title | Study/HW Resource Reference |
| Pre CH | See Class Slides | Distributed computing solutions for data mining | See Class Slides |
| During CH |
| Post CH |

Contact Hour 13

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Content Ref. | Topic Title | Study/HW Resource Reference |
| Pre CH | R1 7.4 | Sequence Mining  Characteristics and Problem Modeling | R1 7.4 |
| During CH |
| Post CH |

Contact Hour 14

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Content Ref. | Topic Title | Study/HW Resource Reference |
| Pre CH | R1 7.4 | Sequence Pattern Discovery  Timing Constraints | R1 7.4 |
| During CH |
| Post CH |

Contact Hour 15

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Content Ref. | Topic Title | Study/HW Resource Reference |
| Pre CH | R4 Ch 1, 13 | Text Mining  Data Representation and Characteristics | R4 Ch 1, 13, R2 Ch 7 |
| During CH |
| Post CH |

Contact Hour 16

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Content Ref. | Topic Title | Study/HW Resource Reference |
| Pre CH | R4 Ch 14 | Text Classification  Feature Selection & Models | R4 Ch 14, R2 Ch 7 |
| During CH |
| Post CH |

Contact Hour 17

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Content Ref. | Topic Title | Study/HW Resource Reference |
| Pre CH | R4 Ch 14 | Text Classification  Vector Space Model | R4 Ch 14, R2 Ch 7 |
| During CH |
| Post CH |

Contact Hour 18

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Content Ref. | Topic Title | Study/HW Resource Reference |
| Pre CH | R4 Ch 13, 14 | Text Classification  Multiclass classifiers for text | R4 Ch 13,14 |
| During CH |
| Post CH |

Contact Hour 19

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Content Ref. | Topic Title | Study/HW Resource Reference |
| Pre CH | R4 Ch 16, 17 | Text Clustering  Flat and hierarchical | R4 Ch 16,17 |
| During CH |
| Post CH |

Contact Hour 20

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Content Ref. | Topic Title | Study/HW Resource Reference |
| Pre CH | R4 Ch 1, 6, 19 | Web Search | R4 Ch 1, 6, 19 |
| During CH |
| Post CH |

Contact Hour 21

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Content Ref. | Topic Title | Study/HW Resource Reference |
| Pre CH | R4 Ch 20 | Crawling & Indexing | R4 Ch 20 |
| During CH |
| Post CH |

Contact Hour 22

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Content Ref. | Topic Title | Study/HW Resource Reference |
| Pre CH | R4 Ch 20 | Crawling & Indexing | R4 Ch 20 |
| During CH |
| Post CH |

Contact Hour 23

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Content Ref. | Topic Title | Study/HW Resource Reference |
| Pre CH | R4 Ch 20 | Crawling & Indexing | R4 Ch 20 |
| During CH |
| Post CH |

Contact Hour 24

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Content Ref. | Topic Title | Study/HW Resource Reference |
| Pre CH | R4 Ch 21  See Class slides | Link Analysis | R4 Ch 21 |
| During CH |
| Post CH |

Contact Hour 25

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Content Ref. | Topic Title | Study/HW Resource Reference |
| Pre CH | R5 Ch1  See Class slides | Mining Complex Structures  Data Representation | R5 Ch1 |
| During CH |
| Post CH |

Contact Hour 26

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Content Ref. | Topic Title | Study/HW Resource Reference |
| Pre CH | R5 Ch 2, 3  See Class slides | Tree Mining problem and Tree basics | R5 Ch 2, 3 |
| During CH |
| Post CH |

Contact Hour 27

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Content Ref. | Topic Title | Study/HW Resource Reference |
| Pre CH | R5 Ch 3  See Class slides | Tree Miner | R5 Ch 3 |
| During CH |
| Post CH |

Contact Hour 28

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Content Ref. | Topic Title | Study/HW Resource Reference |
| Pre CH | R5 Ch 4, 5, 6 | TMG Model Guided Framework | R5 Ch 4, 5, 6 |
| During CH |
| Post CH |

Contact Hour 29

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Content Ref. | Topic Title | Study/HW Resource Reference |
| Pre CH | R5 Ch 11  See Class slides | Graph Mining  Introduction and applications | R5 Ch 11 |
| During CH |
| Post CH |

Contact Hour 30

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Content Ref. | Topic Title | Study/HW Resource Reference |
| Pre CH | See Class slides | Case Study: Information Retrieval |  |
| During CH |
| Post CH |

Contact Hour 31

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Content Ref. | Topic Title | Study/HW Resource Reference |
| Pre CH | See Class slides | Case Study: Social Network Mining |  |
| During CH |
| Post CH |

Contact Hour 32

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Content Ref. | Topic Title | Study/HW Resource Reference |
| Pre CH | See Class slides | Case Study: Social Network Mining |  |
| During CH |
| Post CH |

**Evaluation Scheme**:

Legend: EC = Evaluation Component; AN = After Noon Session; FN = Fore Noon Session

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| No | Name | Type | Duration | Weight | Day, Date, Session, Time |
| EC-1 | Quiz-I/ Assignment-I | Online | - | 5% | February 1 to 10, 2018 |
|  | Quiz-II |  |  | 5% | March 1 to 10, 2018 |
|  | Quiz-III/ Assignment-II |  |  | 5% | March 20 to 30, 2018 |
| EC-2 | Mid-Semester Test | Closed Book | 2 hours | 35% | 04/03/2018 (FN) 10 AM – 12 Noon |
| EC-3 | Comprehensive Exam | Open Book | 3 hours | 50% | 22/04/2018 (FN) 9 AM – 12 Noon |

Syllabus for Mid-Semester Test (Closed Book): Topics in Session Nos. 1 to 16

Syllabus for Comprehensive Exam (Open Book): All topics (Session Nos. 1 to 32)

**Important links and information:**

Elearn portal: https://elearn.bits-pilani.ac.in

Students are expected to visit the Elearn portal on a regular basis and stay up to date with the latest announcements and deadlines.

Contact sessions: Students should attend the online lectures as per the schedule provided on the Elearn portal.

Evaluation Guidelines:

1. EC-1 consists of either two Assignments or three Quizzes. Students will attempt them through the course pages on the Elearn portal. Announcements will be made on the portal, in a timely manner.
2. For Closed Book tests: No books or reference material of any kind will be permitted.
3. For Open Book exams: Use of books and any printed / written reference material (filed or bound) is permitted. However, loose sheets of paper will not be allowed. Use of calculators is permitted in all exams. Laptops/Mobiles of any kind are not allowed. Exchange of any material is not allowed.
4. If a student is unable to appear for the Regular Test/Exam due to genuine exigencies, the student should follow the procedure to apply for the Make-Up Test/Exam which will be made available on the Elearn portal. The Make-Up Test/Exam will be conducted only at selected exam centres on the dates to be announced later.

It shall be the responsibility of the individual student to be regular in maintaining the self study schedule as given in the course handout, attend the online lectures, and take all the prescribed evaluation components such as Assignment/Quiz, Mid-Semester Test and Comprehensive Exam according to the evaluation scheme provided in the handout.