Master of Technology
In
Data Science

SCHEME & SYLLABUS
I to IV semesters

2019
<table>
<thead>
<tr>
<th>SL. No.</th>
<th>Subject Code</th>
<th>Course Title</th>
<th>Teaching Department</th>
<th>Credits</th>
<th>Contact Hours</th>
<th>CIE</th>
<th>SEE</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>SDS110</td>
<td>Principles of Data Science</td>
<td>IS&amp;E</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>5.0</td>
<td>6</td>
</tr>
<tr>
<td>2.</td>
<td>SDS120</td>
<td>Big Data Analytics</td>
<td>IS&amp;E</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>5.0</td>
<td>6</td>
</tr>
<tr>
<td>3.</td>
<td>SDS130</td>
<td>Machine Learning</td>
<td>IS&amp;E</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>5.0</td>
<td>6</td>
</tr>
<tr>
<td>4a.</td>
<td>SDS141</td>
<td>Computational Statistical Methods</td>
<td>IS&amp;E</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>5.0</td>
<td>6</td>
</tr>
<tr>
<td>4b.</td>
<td>SDS142</td>
<td>Information Retrieval</td>
<td>IS&amp;E</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>5.0</td>
<td>6</td>
</tr>
<tr>
<td>4c.</td>
<td>SDS143</td>
<td>Image &amp; Video Analytics</td>
<td>IS&amp;E</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>5.0</td>
<td>6</td>
</tr>
<tr>
<td>5a.</td>
<td>SDS151</td>
<td>Linear Algebra and Applications</td>
<td>IS&amp;E</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>5.0</td>
<td>6</td>
</tr>
<tr>
<td>5b.</td>
<td>SDS152</td>
<td>System Security</td>
<td>IS&amp;E</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>5.0</td>
<td>6</td>
</tr>
<tr>
<td>5c.</td>
<td>SDS153</td>
<td>Exploratory Data Analysis &amp;</td>
<td>IS&amp;E</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>5.0</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Visualization</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.</td>
<td>SDS160</td>
<td>Minor Project – 1</td>
<td>IS&amp;E</td>
<td>0</td>
<td>0</td>
<td>3.0</td>
<td>3.0</td>
<td>6</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Scheme of Teaching and Examination**

**MTech in Data Science**

**First Semester MTech(DS)**
## Scheme of Teaching and Examination
### MTech in Data Science
#### Second Semester MTech (DS)

<table>
<thead>
<tr>
<th>SL. No</th>
<th>Subject Code</th>
<th>Course Title</th>
<th>Teaching Department</th>
<th>Credits</th>
<th>Contact Hours</th>
<th>Marks</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>SDS210</td>
<td>Advanced Data Mining Techniques</td>
<td>IS&amp;E</td>
<td>4 1 0</td>
<td>5.0</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CIE 50</td>
<td>50 100</td>
</tr>
<tr>
<td>2.</td>
<td>SDS220</td>
<td>Cloud Computing and Virtualization</td>
<td>IS&amp;E</td>
<td>4 0 1</td>
<td>5.0</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CIE 50</td>
<td>50 100</td>
</tr>
<tr>
<td>3.</td>
<td>SDS230</td>
<td>Deep Learning</td>
<td>IS&amp;E</td>
<td>4 0 1</td>
<td>5.0</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CIE 50</td>
<td>50 100</td>
</tr>
<tr>
<td>4a.</td>
<td>SDS241</td>
<td>Optimization Theory</td>
<td>IS&amp;E</td>
<td>4 1 0</td>
<td>5.0</td>
<td>6</td>
</tr>
<tr>
<td>4b.</td>
<td>SDS242</td>
<td>Computational Linguistics</td>
<td>IS&amp;E</td>
<td>4 1 0</td>
<td>5.0</td>
<td>6</td>
</tr>
<tr>
<td>4c.</td>
<td>SDS243</td>
<td>Bioinformatics</td>
<td>IS&amp;E</td>
<td>4 1 0</td>
<td>5.0</td>
<td>6</td>
</tr>
<tr>
<td>5a.</td>
<td>SDS251</td>
<td>Scalable systems for Data Science</td>
<td>IS&amp;E</td>
<td>4 0 1</td>
<td>5.0</td>
<td>6</td>
</tr>
<tr>
<td>5b.</td>
<td>SDS252</td>
<td>Web Databases &amp; Information Systems</td>
<td>IS&amp;E</td>
<td>4 0 1</td>
<td>5.0</td>
<td>6</td>
</tr>
<tr>
<td>5c.</td>
<td>SDS253</td>
<td>Social &amp; Information Network Analysis</td>
<td>IS&amp;E</td>
<td>4 0 1</td>
<td>5.0</td>
<td>6</td>
</tr>
<tr>
<td>6.</td>
<td>SDS260</td>
<td>Minor Project – 2</td>
<td>IS&amp;E</td>
<td>0 0 3.0</td>
<td>3.0</td>
<td>6</td>
</tr>
<tr>
<td>Total</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

|         | CIE 50        | SEE 50 100 | Total 20 2.0 6.0 28 36 350 250 600 |       |

Total 28 36 350 250 600
<table>
<thead>
<tr>
<th>Sl.N o.</th>
<th>Subject Code</th>
<th>Subject Title</th>
<th>Teaching Department</th>
<th>Credits</th>
<th>Contact Hours</th>
<th>Marks</th>
<th>Exam Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SDS31T</td>
<td>Practical Training in Industry/Exploration in Research</td>
<td>IS&amp;E</td>
<td>4</td>
<td>-</td>
<td>100</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>SDS32P</td>
<td>Project Work (Phase – I)</td>
<td>IS&amp;E</td>
<td>10</td>
<td>-</td>
<td>100</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Total</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Credits</td>
<td>14</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Total</td>
<td>Marks 200</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Note:** The table provides a summary of the teaching and examination scheme for the third semester of the MTech in Data Science program.
## Scheme of Teaching and Examination
### MTech in Data Science
#### Fourth Semester MTech (DS)

<table>
<thead>
<tr>
<th>Sl. No.</th>
<th>Subject Code</th>
<th>Course title</th>
<th>Teaching Department</th>
<th>Credits</th>
<th>Contact Hours</th>
<th>Marks</th>
<th>Exam Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SDS41P</td>
<td>Project Work (Phase –II)</td>
<td>IS&amp;E</td>
<td>L 1 T 8</td>
<td>18</td>
<td>100</td>
<td>200</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Total Credits</strong></td>
<td></td>
<td>18</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
SDS110  PRINCIPLES OF DATA SCIENCE

Total Teaching Hours:  50  No. of Credits  :  05

Course Outcomes:

On successful completion of the course students will have the ability to:

CO1: Explain a flow process for data science problems.
CO2: Obtain data to classify data science problems into standard typology using mathematical concepts.
CO3: Analyze results and Correlate them to the solution approach followed.
CO4: Design use cases to validate approach and identify modifications required and develop applications for data science solutions.
CO5: Assess the solution approach through visualization and presentation.

Introduction  10 Hours


Basic Mathematics – Vectors and Matrices  10 Hours


Probability  10 Hours

Basic definitions, Probability, Bayesian versus Frequentist, Frequentist approach, Compound events, Conditional probability, The rules of probability, Collectively exhaustive events, Bayesian ideas revisited, Bayes theorem, Random variables.

Statistics  10 Hours

Basic Statistics, What are statistics ?, How do we obtain and sample data?, Obtaining data, Sampling data, How do we measure statistics?, Point estimates, Sampling distributions, Confidence intervals, Hypothesis tests, Conducting a hypothesis test, Type I and type II errors, Hypothesis test for categorical variables.
**Visualization**

Basic principles, ideas and tools for data visualization, why does communication matter?, Identifying effective and ineffective visualizations, Scatter plots, Line graphs, Bar charts, Histograms, Box plots, When graphs and statistics lie, Correlation versus causation, Simpson's paradox, If correlation doesn't imply causation, then what does?, Verbal communication, The why/how/what strategy of presenting, Data Science ethical issues.

**Text Book:**


**Reference Books:**

2. Doing Data Science: Straight Talk from the Frontline, Cathy O'Neil, Rachel Schutt, O'Reilly Media, 2013.

**Online Courses:**

1. Swayam : https://swayam.gov.in/nd1_noc19_cs60/preview
2. NPTEL : https://nptel.ac.in/courses/106106179/
SDS120    BIG DATA ANALYTICS

Total Teaching Hours:  50                 No. of Credits  :  05

Course Outcomes (CO):

At the end of the course student will be able to:
CO1: Explain various forms of data and analytical models.
CO2: Apply big data technologies and tools to analyze the data.
CO3: Explain the various streaming techniques and recommendation system.
CO4: Apply machine learning algorithms to handle big data.
CO5: Understand the role of big data in the current industrial context.

Introduction               10 Hours

Overview of Big Data, History, Structuring Big Data, Types of Data, Elements of Big Data, Data analytics project life cycle, Problems & challenges in understanding Data Analytics, Web page categorization, computing the frequency of stock market change. Use of Big Data in Social Networking, Use of Big Data in preventing Fraudulent activities, Use of Big Data in Retail Industry.

Big Data Technology         10 Hours

Exploring Big Data Stack, Virtualization, Virtualization Approaches, Distributed and parallel computing for Big Data, The cloud and Big Data, Cloud Deployment Models, Cloud Delivery Models, Cloud providers in Big Data Market. Introducing Hadoop, Hadoop Ecosystem, Hadoop Distributed File Systems(HDFS), Features of HDFS : Hadoop YARN, MAP Reduce, Features of Map Reduce, Working of Map Reduce, Techniques to Optimize Map Reduce Jobs, Uses of Map Reduce, HBase, Features of HBase, Role of HBase in Big Data processing, Other tools of Hadoop (Hive, Pig and Pig Latin, Sqoop, ZooKeeper, Flume, OOZie),

Mining Data Streams         08 Hours


Frequent Itemsets and Recommendation Systems        12 Hours

Large Scale Machine Learning


Text Books:


Reference Books:

SDS130  MACHINE LEARNING

Total Teaching Hours:  50  No. of Credits : 05

Course Outcomes:

On successful completion of the course students should be able to:
CO1: Understand the basic principles of Learning theories
CO2: Understand the principles of dimensionality reduction and feature selection techniques
CO3: Understand and Develop a wide variety of supervised learning algorithms
CO4: To become familiar with various clustering algorithms
CO5: To learn methodology and tools to apply machine learning algorithms to Software Engineering

Introduction & Bayesian Decision Theory  10 Hours


Dimensionality Reduction  10 Hours


Supervised Learning  10 Hours

Learning a Class from Examples, Probably Approximately Correct (PAC) Learning, Noise, Learning Multiple Classes, Regression, Model Selection and Generalization, Dimensions of a Supervised Machine Learning Algorithms, Decision Tree Induction, Nearest Neighbors, Bayesian Classifier, Model Over fitting, Performance Evaluation of classifiers.

Clustering  10 Hours


Multilayer Perceptron  10 Hours

The Perceptron, Learning Boolean Functions, MLP as a universal approximator, Back Propagation Algorithm, Training Procedures, Tuning Networks, Recurrent Networks, Radial Basis functions.
Text Books:


Reference Materials:

4. Related Research Articles
SDS141    COMPUTATIONAL STATISTICAL METHODS

Total Teaching Hours:  50          No. of Credits :  05

Course outcomes:

At the end of the course, students will have learned about the following concepts.
CO1: Explain computational statistical methods.
CO2: Illustrate unsupervised and supervised learning techniques.
CO3: Implement different learning algorithms.

Statistical Learning  10 hours
Introduction, what is Statistical Learning, Assessing Model Accuracy

Linear Regression  10 hours
Simple Linear Regression, Multiple Linear Regression, Other considerations in the Regression Model, The Market Plan, Comparison of Linear Regression with K-Nearest Neighbors

Classification  10 hours
Overview of Classification, Why not Linear Regression, Logistics Regression, LDA, Comparison of Classification Methods

Resampling Methods/ Linear Model Selection & Regularization  10 hours
Cross Validation, The Bootstrap Subset Selection, Shrinkage Methods, Dimensionality Reduction Methods, Considerations in High Dimension

Tree based Methods/SVM/Unsupervised Learning  10 hours
Basics of Decision Trees, Bagging, Random Forests, Boosting SVMs: Maximal Margin Classifier, Support Vector Classifiers, SVMs, SVMs with more than two classes, Relation to Logistic Regression Unsupervised Learning: Challenges, PCA, Cluster Methods

Text Books:

1. Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, “An Introduction to Statistical Learning with Applications in R”, 2014, Springer.

Reference Books:

SDS142 INFORMATION RETRIEVAL

Total Teaching hours: 50
No. of credits: 05

Course outcomes:

On successful completion of the course, students should be able to:
CO1: Understand the basics of Information Retrieval (IR) models
CO2: Able to understand Boolean and vector-space retrieval models
CO3: Understand the importance of efficient Indexing techniques
CO4: Will be able to understand various searching techniques
CO5: Able to understand various feedback methods and evaluation metrics

Introduction:
10Hrs

Motivation, Basic concepts, Past, present, and future, The retrieval process. **Modeling:**
Introduction, A taxonomy of information retrieval models, Retrieval: Adhoc and filtering,
A formal characterization of IR models, Classic information retrieval, Alternative set theoretic models, Alternative algebraic models, Alternative probabilistic models,
Structured text retrieval models, Models for browsing.

IR Using Boolean Model:
10 Hrs

Information retrieval using the Boolean model, An example information retrieval problem, A first take at building an inverted index, Processing Boolean queries, The
dictionary and postings lists, Determining dictionary terms, Postings lists, revisited,
Tolerant retrieval, Wildcard queries, Spelling correction, Phonetic correction.

Indexing
10 Hrs

Introduction; Inverted Files; Other indices for text; Boolean queries; Sequential
searching; Pattern matching; Structural queries; Compression. Index construction,
Construction of large indexes, Distributed indexing, Dynamic indexing, Other types of
indexes
Parallel and Distributed IR: Introduction, Parallel IR, Distributed IR.

Vector space retrieval
10 Hrs

Scoring and term weighting, Parametric and zone indexes, Weighted zone scoring,
Term frequency and weighting, Inverse document frequency, tf-idf weighting, Variants in
weighting functions, Documents as vectors, Inner products, Queries as vectors,
Heuristics for efficient scoring and ranking, Inexact top K document retrieval, Interaction
between vector space and other retrieval methods, Query parsing and composite
classification and clustering.
Retrieval Evaluation


Text Books:


Reference Books:

SDS143  IMAGE AND VIDEO ANALYTICS

Total Teaching hours: 50  No. of credits: 05

Course outcomes:

On successful completion of the course, students should be able to:
CO1: Explain the fundamental principles of image and video analysis
CO2: Design the various filters for Image enhancement.
CO3: Apply Motion Estimation techniques for Video
CO4: Apply various compression methods for video and Image
CO5: Analyze various segmentation and Compression techniques

Introduction 10 Hours

Digital Image Fundamentals, Elements of visual perception, Image sensing and acquisition, Sampling and Quantization, Relationships between pixels, Linear and Non-Linear operations.

Image Enhancement 10 Hours


Digital Video & Motion Estimation 10 Hours


Image and Video Segmentation 10 Hours

Detection of Discontinues, Edge Linking and Boundary Detection, Threshold Based Segmentation, Region Based Segmentation, Segmentation by Morphological Watersheds, The use of Motions in Segmentation, Change Detection, Motion Tracking, Image and Video Matting.

Image & Video Compression 10 Hours

Compression.

Text Books:


Reference Book:

SDS151  LINEAR ALGEBRA AND APPLICATIONS

Total Teaching Hours:  50  No. of Credits  :  05

Course Outcomes:

On successful completion of the course, students should be able to:
CO1: Solve linear equations through matrix representation.
CO2: Identify the vector space and subspace.
CO3: Test for orthogonality and apply Eigen vectors to solve differential equations.
CO4: Justify linear inequalities in the vector subspace.
CO5: Apply methods to solve different linear equations.

Introduction  10 Hours

Vector Space  10 Hours
Vector Spaces and Subspaces, Solving $Ax= 0$ and $Ax = b$, 3 Linear Independence, Basis, and Dimension, The Four Fundamental Subspaces, Graphs and Networks, Linear Transformations, Review Exercises.

Orthogonality  10 Hours

Computations with Matrices  10 Hours
Introduction, Matrix Norm and Condition Number, Computation of, Iterative Methods for $Ax = b$.

Linear Algebra and Game Theory  10 Hours
Linear Inequalities, The Simplex Method, The Dual Problem, Network, GameTheory . .

Text Book:

Reference Book:

SDS152  SYSTEM SECURITY

Total Teaching Hours:  50  No. of Credits : 05

Course Outcomes:

On successful completion of the course, students should be able to:
CO1: Identify the security, cryptography techniques and provide security for OS.
CO2: Explain ways of providing security to databases and networks.
CO3: Apply security and privacy techniques to cyber related issues.

Security / Cryptography  10 hours


Program / OS Security  10 hours

Program Security, Secure Programs, No malicious Program Errors, Viruses and Other Malicious Code, Targeted Malicious Code, Controls Against Program Threats, Summary of Program Threats and Controls
Protection in General-Purpose Operating Systems, Protected Objects and Methods of Protection, Memory and Address Protection, Control of Access to General Objects, File Protection Mechanisms, User Authentication, Summary of Security for Users

Trusted OS / Database Security  10 hours

Database and Data Mining Security, Introduction to Database, Security Requirements, Reliability and Integrity, Sensitive Data, Inference, Multilevel Databases, Proposals for Multilevel Security, Data Mining

Network / Administrative Security  10 hours

Cyber Security / Privacy

The Economics of Cyber security, making a Business Case, Quantifying Security, Modeling Cyber security, Current Research and Future Directions, Summary


Text Book:

SDS153  EXPLORATORY DATA ANALYSIS AND VISUALIZATION

Total Teaching Hours:  50  No. of Credits :  05

Course Outcomes:

On completion of the course the student should be able to
CO1: Understand the different concepts of EDA techniques
CO2: Explain the techniques for testing EDA assumptions
CO3: Apply data analysis techniques to solve real world problems.
CO4: Conduct a survey for different EDA case studies
CO5: Create visualization using basic and advanced techniques of information visualization and scientific visualization.

Introduction 10 hours

EDA Assumptions 10 hours
Underlying Assumptions, Importance, Techniques for Testing Assumptions, Interpretation of 4-Plot, Consequences.

EDA Techniques 10 hours

EDA Case Studies 10 hours
Case Studies Introduction, Case Studies: Normal random numbers, Uniform random numbers, Random walk, Josephson Junction Cry thermometry, Beam Deflections, Filter Transmittance, Standard Resistor, Heat Flow Meter 1, Airplane Glass Failure Time, Ceramic Strength.

Data Visualization 10 hours
Introduction to R, Rstudio, and Data cleaning and aggregation, Design principles for charts and graphs, ggplot2 and Tableau tools for creating data visualizations, The process creating visualizations and selecting the appropriate visual display, Designing effective digital presentations, Visualization as exploration, Visualizing categorical data, Visualizing time series data, Visualizing multiple variables, Visualizing geospatial data, Dashboard design, Web-based visualizations, Interactive visualizations and motion.
Text Books:

1. Exploratory Data Analysis With R, Roger D.Peng
2. Interactive Data Visualization for the Web, Scott Murra

Reference Books:

2. Advanced Analytics with R and Tableau, Jen Stirrup, Packt Publications
SDS210 ADVANCED DATA MINING TECHNIQUES

No. of Credits: 05  Total Hours: 50

Course Outcome:

After completion of this course, students should be able to
CO1: Understand the Data Preparation techniques for Data Mining Process
CO2: Understand the various techniques to Mine the Data Streams
CO3: Exploring Advanced Concepts to build Association among items
CO4: Understand and Explore Data Mining methods as Tools
CO5: Prepare students for research in the area of data mining and related applications

Introduction

10 Hours

The Data Mining Process: Basic Data Types, The Major Building Blocks: A Bird's Eye View, Scalability Issues and the Streaming Scenario, A Stroll through some Application Scenarios, Data Preparation, Feature Extraction and Portability, Data Cleaning: Data Reduction and Transformation, Similarity and Distances: Multidimensional Data, Text Similarity Measures, Temporal Similarity Measures, Graph Similarity Measures, Supervised Similarity Functions

Mining Data Stream

10 Hours

Mining Time-Series Data, Mining Sequence Patterns in Transactional Databases, Mining Sequence Patterns in Biological Data, Graph Mining, Social Network Analysis, Multirelational Data Mining, Multidimensional Analysis and Descriptive Mining of Complex Data Objects, Spatial Data Mining, Multimedia Data Mining, Text Mining, Mining the World Wide Web.

Advanced Concepts in Association Analysis

8 Hours

Frequent Itemset Generation, Compact Representation of Frequent Itemsets, FP-Growth Algorithms, Handling Categorical and Continuous Attributes, Handling a Concept Hierarchy, Sequential Patterns, Subgraph Patterns, Infrequent Patterns, Counting Frequent Items in a Stream.

Data Mining Methods as Tools

10 Hours

Memory-Based Reasoning Methods, Fuzzy Sets in Data Mining, Rough Sets, Support Vector Machines, Genetic Algorithm Support to Data Mining, Performance Evaluation for Predictive Modeling.
Applications and Research Trends in Data Mining

Data Mining Applications (Financial Data Analysis, Retail Industry, Telecommunication Industry, Biological Data Analysis, Other Scientific Applications, Intrusion Detection), Data Mining System Products and Research Prototypes, Statistical Data Mining, Visual and Audio Data Mining, Data Mining and Collaborative Filtering, Data Mining, Privacy, and Data Security, Trends in Data Mining, Present Research Avenues.

Text Books:


2. *Advanced Data Mining Techniques*, David L. Olson, DursunDelen, Springer Publisher, 2008

Reference Books:


SDS220 CLOUD COMPUTING AND VIRTUALIZATION

Total Teaching Hours: 50  No. of Credits : 05

Course Outcomes:

After completing this course, the students would be able to:
CO 1: Explain the various services offered by the cloud.
CO 2: Understand the cloud computing delivery models.
CO 3: Distinguish different types of virtualization.
CO 4: Apply scheduling algorithms for cloud applications.
CO 5: Illustrate various security issues in cloud computing.

Introduction

10 Hours

Introduction, Cloud Infrastructure Cloud computing, Cloud computing delivery models and services, Ethical issues, Cloud vulnerabilities, Cloud computing at Amazon, Cloud computing the Google perspective, Microsoft Windows Azure and online services, Open-source software platforms for private clouds, Cloud storage diversity and vendor lock-in, Energy use and ecological impact, Service level agreements, User experience and software licensing.

Computing

10 Hours


Virtualization

10 Hours


Management/Scheduling

10 Hours

Cloud Resource Management and Scheduling. Policies and mechanisms for resource management, Application of control theory to task scheduling on a cloud, Stability of a two-level resource allocation architecture, Feedback control based on dynamic thresholds, Coordination of specialized autonomic performance managers, A utility-based model for cloud-based Web services, Resourcing bundling: Combinatorial auctions for cloud resources, Scheduling algorithms for computing clouds, Fair queuing, Start-time fair queuing, Borrowed virtual time, Cloud scheduling subject to deadlines, Scheduling Map Reduce applications subject to deadlines, Resource management and
dynamic scaling.

Security


Text Book:


Reference Books:

SDS230  DEEP LEARNING

Total Teaching hours: 50  
Credits: 5

Course Outcomes:

After completion of the course student should be able to:
CO1: Explain the basic concepts in the design of neural networks
CO2: Apply the concepts of neural networks in designing deep learning applications
CO3: Illustrate parameter tuning for deep learning architecture

Introduction to shallow network  
11Hrs


Introduction to deep learning  
12Hrs


Building a simple neural network  
12Hrs

Building a neural network: matching neural network for the right problem, DL4J suit tools, modeling CSV data for MLP, modeling handwritten images using CNN, modeling sequence data using RNN, using autoencoders for anomaly detection, using variational autoencoders for reconstructing MNIST data.

Tuning the deep neural network  
15Hrs

Tuning deep neural network: matching input data and network architecture, relating model goals and network parameters, working with layer count, parameter count and memory, weight initialization strategies, apply loss function, applying optimization, controlling epochs and batch size, dealing with overfitting. Tuning specific deep network architecture: CNN, RNN, Restricted Boltzmann machine, DBN. Vectorization: an introduction to vectorization in machine learning, using dataVec for ETL, vectorizing image data, working with sequential data and vectorization, working with text data.

Text Book:


Reference Books:

SDS241   OPTIMIZATION THEORY

Total Teaching Hours:  50        No. of Credits :  05

Course outcomes:

On successful completion of the course, students should be able to:
CO1: Apply classical optimization techniques and modern methods.
CO2: Develop linear and Non-linear programming.
CO3: Identify optimized solutions to constrained systems
CO4: Contrast single variable and multivariable optimization problems.
CO5: Analyse different system and evaluate solution to optimize resources.

Introduction 10 Hours


Linear Programming 10 Hours


Nonlinear Programming 10 Hours


Geometric Programming 10 Hours

Posynomial, Unconstrained Minimization Problem, Solution of an Unconstrained Geometric Programming Program Using Differential Calculus, Solution of an Unconstrained Geometric Programming Problem Using Arithmetic–Geometric Inequality, Primal–Dual Relationship and Sufficiency Conditions in the Unconstrained Case, Constrained Minimization, Solution of a Constrained Geometric Programming Problem, Primal and Dual Programs in the Case of Less-Than Inequalities, Geometric
Programming with Mixed Inequality Constraints, Complementary Geometric Programming.

**Stochastic Programming:** Basic Concepts of Probability Theory, Stochastic Linear Programming, Stochastic Nonlinear Programming.

**Modern Methods**

10 Hours


**Text Book:**


**Reference Book:**

SSE242  COMPUTATIONAL LINGUISTICS

Total Teaching Hours:  50  No. of Credits : 05

Course outcomes

On successful completion of the course, students should be able to:

CO1: Explain finite state transducers for language processing, rules and analysis.
CO2: Demonstrate parsers in computational linguistics
CO3: Implement natural language processing applications.
CO4: Discuss clustering of text data in natural language processing applications.
CO5: Illustrate machine translation system using machine learning algorithms.

Introduction  10 Hours

What is computational linguistics? Ambiguity and uncertainty in language, regular languages, and their limitations, finite-state automata, morphology. Natural Language Toolkit.

Context Free Grammars  10 Hours

Constituency, CFG definition, use and limitations. Chomsky Normal Form. Top-down parsing, bottom-up parsing, and the problems with each. The desirability of combining evidence from both directions

Computational Discourse  10 Hours

Discourse segmentation, Text coherence, reference resolution, reference phenomena, Features for pronominal anaphora resolution, Algorithms for anaphora resolution, coreference resolution, evaluation of coreference resolution.

Word Sense Disambiguation and Clustering  10 Hours


Machine Translation  10 Hours

Text Books:

SDS243 BIOINFORMATICS

Total Teaching Hours: 50 No. of Credits : 05

Course Outcome:

On successful completion of the course students will have the ability to:

CO1: Explain Bioinformatics, biological databases and data models.
CO2: Obtain sequence data from biological databases and compare different file formats.
CO3: Analyze and evaluate methods of sequence alignment, utilize similarity finding tools to compare sequence data.
CO4: Design applications using bioinformatics algorithms for gene finding, predictions and genome studies.
CO5: Engage in independent study to analyze and interpret protein structure using biostatistics & biological tools and compare different methods and strategies of structure prediction.

Introduction and NCBI Data Model 8 Hours


Bioinformatics Database 10 Hours

Importance of Databases, Characteristics and Categories of Bioinformatics Database, Navigating Databases, Biological Databases, Primary Sequence Databases, Composite Sequence Databases, Secondary Databases, Nucleic Acid Sequence Databases, Structure Databases: File Formats, Protein Structure, PDB, MMDB, CATH, Other Database Enzyme, MEROPS, BRENDA, Pathway databases, Bibliographic Databases, Specialized Genomic Resources, Analysis Packages.

Sequence Align Methods 10 Hours

Sequence Analysis of Biological Data, Significance of Sequence Alignment, Pairwise Sequence Alignment Methods, Use of Scoring Matrices and Gap Penalties in Sequence Alignments, Multiple Sequence Alignment Methods - Tools and Application of multiple sequence alignment, Gene Prediction Strategies, Protein Prediction Strategies, Phylogenetic Trees and Multiple Alignments.

Bioinformatics Algorithms 12 Hours

**Biostatistics & Tools**

10 Hours

Handling Univariate and Bivariate Data, Measures of Central Tendency, Measures of Dispersion, Skewness & Kurtosis, Correlation and Regression.

Local Alignment Search Tool (BLAST), Purpose of BLAST, BLAST Analysis, Purpose of BLAST II, Scoring Metrics, PAM, BLOSUM, Working of BLAST, Introduction to HMMER.

**Text Books:**


**Reference Books:**


**Online Courses:**

1. NPTEL: [https://nptel.ac.in/courses/102106065/](https://nptel.ac.in/courses/102106065/)
SDS251  SCALABLE SYSTEMS FOR DATA SCIENCE

Total Teaching Hours:  50  No. of Credits :  05

Course outcomes:

At the end of the course, students will have the ability to:

CO1: Explain big data platforms and distributed systems.
CO2: Obtain large scale of input data and output results using different algorithmic
techniques and software paradigms.
CO3: Analyze algorithms and modify to adapt to a distributed setting.
CO4: Design and develop scalable algorithms and systems for the common data
science tasks.
CO5: Engage in independent study and make an effective oral presentation on
applications of Scalable data science.

Big Data  10 hours

Big Data & Platform Design Goals, Big Data & other computing platforms. Programming
for Large Datasets:. Distributed systems, scalability and metrics, Degrees of
Parallelism, MapReduce- Uses, Model working, simple and advanced applications
programming. Runtime Systems: Hadoop- Open cloud server, Class cluster, Hadoop
distributed file system. Hadoop YARN, Hadoop MapReduce, Fault Tolerance.

Prediction  10 hours

Prediction over graphs – Scalable learning and inference over graphs, Semi supervised
learning (SSL)- self training and co training, Graph Based SSL. Streaming Naive Bayes
– Introduction to Naïve Bayes, Complexity of Naïve Bayes, Implementation of Naïve
Bayes Classifier, Large vocabulary counting, Sorting – Merge sort, Unix sort, Large
vocabulary Naïve Bayes, Distributed Counting, optimizations.

Learning  10 hours

Scalable Logistic Regression and SGD. Learning as optimization, Stochastic gradient
descent, SGD versus streaming, Logistic regression versus Rocchio and Naïve Bayes,
Efficient Logistic Regression with Stochastic Gradient Descent, Regularized logistic
regression, Sparse updates for Regularized logistic regression, Bounded memory
logistic regression, SGD implementation.

Matrix Factorization  10 hours

Large-scale Matrix Factorization (MF) - Recovering Latent factor in a matrix, Matrix
factorization for collaborative filtering, MF for image and text modeling, Large scale MF
for distributed SGD, Distributed SGD for Mapreduce.MR Advanced Topics: Inverted
Index, PageRank, Dist

ributed graph processing: Apache Giraph, GoFFish.
Distributed stream processing 10 hours


Text Books:


References:

1. Research papers and articles - [MR for ML on Multicore, NIPS 06], [Hogwild!], [Bottou, 2010], [Gemulla et al., KDD 2011]

Online courses:

1. NPTEL: https://nptel.ac.in/courses/106105186/
2. SWAYAM: https://swayam.gov.in/nd1_noc19_cs61/preview
SDS252   WEB DATABASES AND INFORMATION SYSTEMS

Total Teaching Hours:  50                     No. of Credits :  05

Course Outcomes:

After completion of course students should be able to:
CO1: Illustrate the concepts of web-based information systems.
CO2: Explain n-tiered architectures to implement secure, scalable systems
CO3: Demonstrate database driven applications.
CO4: Design and develop web-based applications.
CO5: Describe XML and alternative technologies.

Web-based Information system                       10 hours

Web -Based Information Systems, Applications: electronic commerce, Variants of Web database access, Basic Web Standards, architecture. HTTP, Forms, Server-Side Programming and CGI Alternatives to CGI: Java Servlets and server APIs Browser detection, state, cookies and redirects

Using n-tiered architectures to implement secure and scalable systems 10 hours

Web App Architectures: Multi-Tier (2-Tier, 3-Tier) Model-Viewer-Controller (MVC)

Database-driven websites and applications               10 Hours

Utilize JavaScript to improve database driven websites. – Critical components of the modern Web infrastructure: DNS, CDN, etc Critical components of the modern Web infrastructure: DNS, CDN, etc

DBMS and WWW:                                                  10 hours

Introduction Off-Line access to databases Static and Dynamic Web Pages SQL embedded in HTML CGI solution to database gateway Internet database connector JDBC: databases the Java way. Solutions from database vendors Association rule mining

XML and its alternatives                           10 Hours

HTTP, XML, SQL, JavaScript, AJAX XML and its Alternatives: XML: Basics of XML, namespace, schema languages, XSLT and XPath, alternatives to XML, SQL, CSS, RSS, and others.

Text Books:

1. PROFESSIONAL WEB 2.0 PROGRAMMING: USING XHTML, CSS, JAVASCRIPT AND AJAXBy Eric Van Der, Danny Ayers, Erik Bruchez

2. Weaving a Website - Programming in HTML, Javascript, Perl, and Java

Reference Book:
Course Outcomes:

After completion of course students should be able to:
CO1: Analyze the social and information network.
CO2: Apply methods to evaluate the behavior of cascading networks.
CO3: Design a graph model representing a social network.
CO4: Apply different search algorithms over the graph model representing social network.
CO5: Interpret the diffusion over social and information network.

Introduction 10 hours

Overview: Aspects of Networks, Central Themes and Topics. Graphs: Basic Definitions, Paths and Connectivity, Distance and Breadth-First Search, Network Datasets: An Overview

The Small-World Phenomenon 10 hours

Positive and Negative Relationships: Structural Balance, Characterizing the Structure of Balanced Networks, Applications of Structural Balance, A Weaker Form of Structural Balance, Advanced Material: Generalizing the Definition of Structural Balance.

Cascading Behaviour in Networks 10 hours


Power Laws and Rich-Get-Richer Phenomena 10 hours


The structure of the Web: The World Wide Web, Information Networks, Hypertext,
Link Analysis and Web Search

Searching the Web: The Problem of Ranking, Link Analysis using Hubs and Authorities, PageRank, Applying Link Analysis in Modern Web Search, Applications beyond the Web, Advanced Material: Spectral Analysis, Random Walks, and Web Search

Strong and Weak Ties: Triadic Closure, The Strength of Weak Ties, Tie Strength and Network Structure in Large-Scale Data, Tie Strength, Social Media, and Passive Engagement Closure, Structural Holes, and Social Capital, Advanced Material: Betweenness Measures and Graph Partitioning

Text Book:

1. "Networks, Crowds, and Markets Reasoning about a Highly Connected World", David Easley, Cornell University, New York, Jon Kleinberg, Cornell University, New York, 2010

Reference Books: