B.Sc. (Data Science)
Discipline Specific Course (DSC)
Semester I
BSDB31103T: Introduction to Data Science

Total Marks: 100
External Marks: 70
Internal Marks: 30
Credits: 4
Pass Percentage: 40%

Objective

To provide strong foundation for data science and application area related to it and understand the underlying core concepts and emerging technologies in data science. Understand data analysis techniques for applications handling large data.

INSTRUCTIONS FOR THE PAPER SETTER/EXAMINER:

1. The syllabus prescribed should be strictly adhered to.
2. The question paper will consist of three sections: A, B, and C. Sections A and B will have four questions from the respective sections of the syllabus and will carry 10 marks each. The candidates will attempt two questions from each section.
3. Section C will have fifteen short answer questions covering the entire syllabus. Each question will carry 3 marks. Candidates will attempt any ten questions from this section.
4. The examiner shall give a clear instruction to the candidates to attempt questions only at one place and only once. Second or subsequent attempts, unless the earlier ones have been crossed out, shall not be evaluated.
5. The duration of each paper will be three hours.

INSTRUCTIONS FOR THE CANDIDATES

Candidates are required to attempt any two questions each from the sections A and B of the question paper and any ten short questions from Section C. They have to attempt questions only at one place and only once. Second or subsequent attempts, unless the earlier ones have been crossed out, shall not be evaluated.

Section A

Unit I: Data Science-a discipline, Landscape-Data to Data science, Data Growth-issues and challenges, data science process. foundations of data science. Messy data, Anomalies and artefacts in datasets. Cleaning data.
Unit II: Introduction data acquisition, Structured Vs Unsructured data, data preprocessing techniques including data cleaning, selection, integration, transformation and reduction, data mining, interpretation.
Unit III: Representation of data: Special types-acoustic, image, sensor and network data. Problems when handling large data – General techniques for handling large data, Distributing data storage and processing with Frameworks
Unit IV: **Data Science Ethics** – Doing good data science – Owners of the data - Valuing different aspects of privacy - Getting informed consent - The Five Cs – Diversity – Inclusion – Future Trends.

Section B

Unit V: Data Wrangling Combining and Merging Data Sets – Reshaping and Pivoting – Data Transformation – String manipulations – Regular Expressions

Unit VI: Data Aggregation and Group Operations Group By Mechanics – Data Aggregation – GroupWise Operations – Transformations – Pivot Tables – Cross Tabulations – Date and Time data types.

Unit VII: Data Modeling: Basics of Generative modeling and Predictive modeling. Charts-histograms, scatter plots, time series plots etc. Graphs, 3D Visualization and Presentation.

Unit VIII: Applications of Data Science: Business, Insurance, Energy, Health care, Biotechnology, Manufacturing, Utilities, Telecommunication, Travel, Governance, Gaming, Pharmaceuticals, Geospatial analytics and modeling

Suggested Readings

1. Sinan Ozdemir, Principles of Data Science, Packt Publishing, 2016
2. Joel Grus: Data Science from Scratch, O'Reilly, 2016
3. Foster Provost & Tom Fawcett: Data Science for Business O'Reilly, 2013
4. Roger D. Peng & Elizabeth Matsui: The Art of Data Science, Lean Publishing, 2015