Course code	Course Name	L-T-P Credits	Year of Introduction
CS466	DATA SCIENCE	3-0-0-3	2016

Course Objectives:

- To introduce fundamental algorithmic ideas to process data.
- To introduce and discuss techniques for applying hypotheses and data into actionable predictions.
- To introduce documentation and visualization techniques.

Syllabus:

Modern scientific, engineering, and business applications are increasingly dependent on data, existing traditional data analysis analysis technologies were not designed for the complexity of the modern world. Data Science has emerged as a new, exciting and fast-paced discipline that explores novel statistical, algorithmic, and implementation challenges that emerge in processing, storing, and extracting knowledge from Big Data.

Expected Outcome:

The Student will be able to :

i. explain and discuss the significance of data science and its key functionalities
ii. discuss and demonstrate various models suitable for data science
iii. perform preliminary statistical analysis using R language on simple data sets
iv. perform python-based predication and filtering on simple data sets
v. perform Hadoop and Map-Reduce for data analysis
vi. perform data visualization techniques at a basic level

References:

1. Boris Lublinsky, Kevin T. Smith. Alexcy Yakubovich, "Professional Hadoop Solutions", Wiley, 2015.
2. Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman, "Mining of Massive Datasets". Cambridge University Press, 2014.
3. Nathan Yau, "Visualize This: The Flowing Data Guide to Design, Visualization and Statistics", Wiley, 2011.
4. Nina Zumel, John Mount "Practical Data Science with R··. Manning Publications. 2014.
5. Sameer Madhavan , “Mastering Python for Data Science”, Packt Publishing Limited, 2015.
6. Tony Ojeda, Sean Patrick Murphy, Benjarnin Bengfort. Abhijit Dasgupta. "Practical Data Science Cookbook", Packt Publishing Limited, 2014.
7. W. N. Venablcs. D. M. Smith and the R Core Team, "An Introduction to R", 2013.

Course Plan

Module	Contents	Hours	End Sem. Exam Marks %
I	Data science process-roles, stages in data science project-working with data from files-working with relational databases-exploring data –managing data-cleaning and sampling for modeling and validation-introduction to NoSQL	6	15

II	Choosing and evaluating models-mapping problems to machine learning, evaluating clustering models, validating models-cluster analysis-k-means algorithm, Naive Bayes-Memorization Methods - Linear and logistic regression-unsupervised methods.	8	20
FIRST INTERNAL EXAM			
III	Reading and getting data into R- ordered and unordered factors - arrays and matrices lists and data frames - reading data from files - probability distributions - statistical models In R manipulating objects - data distribution.	8	15
IV	Python-based data visualization, predication through linear regression, collaborative filtering.	6	15
SECOND INTERNAL EXAM			
V	Introduction distributed file system mar reduce. Algorithm using Map Reduce –Matrix –Vector Multiplication by map reduce – Hadoop – Understanding Map Reduce architecture – writing Hadoop Map-Reduce programs-Loading data into HDFS Map-Reduce Programs - Loading data into HDFS - Executing the Map phase - Shuffling and sorting - Reducing phase execution.	6	20
VI	Documentation and deployment - producing effective presentations - introduction to graphical analysis – plot() function - display ing multivariate data - matrix plots multiple plots in one window - exporting graph - using graphics parameters. Case studies.	6	15
END SEMESTER EXAM			

Question Paper Pattern (End semester exam)

1. There will be ***FOUR*** parts in the question paper – **A, B, C, D**
2. **Part A**
   a. **Total marks : 40**
   b. ***TEN*** questions, each have **4 marks**, covering **all the SIX modules** (***THREE*** questions from **modules I & II**; ***THREE*** questions from **modules III & IV**; ***FOUR*** questions from **modules V & VI**).
   All the TEN questions have to be answered.
3. **Part B**
   a. **Total marks : 18**
   b. ***THREE*** questions, each having **9 marks**. One question is from **module I**; one question is from **module II**; one question ***uniformly*** covers **modules I & II**.
   c. ***Any TWO*** questions have to be answered.
   d. Each question can have ***maximum THREE*** subparts.
4. **Part C**
   a. **Total marks : 18**
   b. ***THREE*** questions, each having **9 marks**. One question is from **module III**; one question is from **module IV**; one question ***uniformly*** covers **modules III & IV**.
   c. ***Any TWO*** questions have to be answered.
   d. Each question can have ***maximum THREE*** subparts.

5. **Part D**
    a. **Total marks : 24**
    b. ***THREE*** questions, each having **12 marks**. One question is from **module V**; one question is from **module VI**; one question ***uniformly*** covers **modules V & VI**.
    c. ***Any TWO*** questions have to be answered.
    d. Each question can have ***maximum THREE*** subparts.
6. There will be ***AT LEAST* 40%** analytical/numerical questions in all possible combinations of question choices.