Reg. No.										

MANIPAL INSTITUTE OF TECHNOLOGY

MANIPAL

(A constituent unit of MAHE, Manipal)

VI SEMESTER B.TECH. (INFORMATION TECHNOLOGY/COMPUTER AND COMMUNICATION ENGINEERING)

END SEMESTER EXAMINATIONS, APRIL/MAY 2019

PROGRAM ELECTIVE II : BIG DATA ANALYTICS [ICT 4005]

REVISED CREDIT SYSTEM

(30/04/2019)

Time: 3 Hours MAX. MARKS: 50

Instructions to Candidates:

- Answer ALL the questions.
- Missing data, if any, may be suitably assumed.

1A. With a neat diagram explain typical analytic architecture and describe the challenges of the current analytical architecture for data scientists. **5**

1B. Consider a csv file student.csv with schema (id: int, firstname: char array, lastname: char array, age: int, phone: char array, city: char array, gpa: int). Write a Pig script to compute the following:

i. Calculate the total number of student who have gpa >8.5

ii. To display the names of students who belong to city "Bangalore" **3**

1C. What are the key skill sets and behavioral characteristics of a data scientist **2**

2A. Consider a research study which was conducted to examine the differences between older and younger adults on perceived life satisfaction. A pilot study was conducted to examine this hypothesis. Ten older adults (over the age of 70) and ten younger adults (between the age 20 and 30) were give a life satisfaction test (known to have high reliability and validity). Scores on the measure range from 0 to 60 with high scores indicative of high life satisfaction; low scores indicative of low life satisfaction. The data are presented below in Fig.Q.2A.State the null and alternate hypothesis. Given the tabulated t- value as 1.734 using student t-test calculate the t-value and check whether there is a significant difference in older and younger adult life satisfaction test.

Older Adults	45	38	52	48	25	39	51	46	35	46
Younger Adults	34	32	19	27	57	41	24	39	46	56

Fig.Q.2A

5

2B. Explain the process of topic modelling. With example state the working of latent dirichlet allocation (LDA) model. **3**

2C. Describe the important terminologies used in hive architecture. **2**

3A. Given schema of the movie data as a csv file containing (Movie_id, Movie_name, Director, Collection, Ratings). Write hadoop map and reduce functions to compute the average ratings for a particular movie. **5**

3B. What are the key roles for a successful analytic project? Discuss what activities differentiate model planning and model building phase of data analytic life cycle? **3**

3C. ROC curve is considered as an efficient tool to evaluate classifiers. Justify **2**

4A. Describe linear regression model. Consider the data for the House Price with attribute length, width, no of rooms, kitchen area and price. Price being the predictor variable. Write R code for reading the file and building linear model and test it for predicting the price of a new house. 5

4B. Write R script for performing the following:
a. Create two vectors **X** (2, 4, 7) and **Y** (3, 2, 4).
b. Put the vectors **X** and **Y** into a matrix form **Z** and find the row sum and the column sum of matrix.
c. Replace the odd number of matrix Z with number 1. 3

4C. What is an analytic sandbox? Discuss the importance of analytic sandbox in big data projects? 2

5A. Consider a given dataset in the Table Q.5A which has attributes as Refund, Marital Status, Taxable Income, Evade. We now encounter a new example: Refund = no, Married, Taxable Income = 125. How should this example be classified using the Naive Bayes method? Show your computations.

Table Q.5A

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	no
2	no	married	100K	no
3	no	Single	70K	no
4	Yes	Married	120K	no
5	no	Divorced	95K	yes
6	no	married	60K	no
7	Yes	Divorced	220k	no

5

5B. Explain the concepts of data blocks and file system in hadoop distributed file system. 3

5C. State and explain the functionality of any two modules of MADlib library of in-database analysis. 2