Computer Lab - Practical Question Bank

FACULTY OF COMMERCE, OSMANIA UNIVERSITY

B.Com (Computer Applications) CBCS Semester – VI

Data Analytics Practical Question Bank DSE-603 (c)

1) To setup and install Hadoop in Pseudo-Distributed Mode.

2) Install, Deploy & configure Apache Spark.

3) Implement the following file management tasks in Hadoop: Adding Files and Directories, Retrieving Files, Deleting Files

4) Exploring various shell commands in Hadoop.

5)To install and run Apache Pig in Ubuntu so as to work with Hadoop.

6)To explore basic Data Processing operators in Pig.

7)To explore Hive with its basic commands: create, alter, and drop databases, tables, views, functions and indexes.

8)Data analytics using Apache Spark on Amazon food dataset, find all the pairs of items frequently reviewed together.

9)To perform data analysis on weather dataset using MapReduce.

10)To count the number of lines in a document using MapReduce Programs.

Find the average, max and min temperature for each year in NCDC data set?

11)Write Pig Latin program to sort, group, join, project, and filter your data. Find out Number of Products Sold in Each Country.

12)Implement word count / frequency programs using Map Reduce

13) Implement an MR program that processes a weather dataset

14) Implement Linear and logistic Regression

15) Implement SVM / Decision tree classification techniques

16) Implement clustering techniques

17) Visualize data using any plotting framework

18) Implement an application that stores big data in Hbase / MongoDB / Pig Using Hadoop / R.

19) READING AND WRITING DIFFERENT TYPES OF DATASETS

Reading different types of data sets (.txt, .csv) from web and disk and writing in file in specific disk location.

20)Reading Excel data sheet in R.

21)Reading XML dataset in R

22)VISUALIZATIONS find the data distributions using box and scatter plot.

23)VISUALIZATIONS find the outliers using plot.

24) VISUALIZATIONS Plot the histogram, bar chart and pie chart on sample data

25) REGRESSION MODEL Import a data from web storage. Name the dataset and now do Logistic Regression to find out relation between variables that are affecting the admission of a student in ainstitute based on his or her GRE score, GPA obtained and rank of the student. Also check the model is fit or not. require (foreign), require(MASS)

26)MULTIPLE REGRESSION MODEL Apply multiple regressions, if data have a continuous independent variable. Apply on above dataset.

27) REGRESSION MODEL FOR PREDICTION Apply regression Model techniques to predict the data on above dataset

28) CLASSIFICATION MODEL Install relevant package for classification.

29) CLASSIFICATION MODEL Choose classifier for classification problem.

30) CLASSIFICATION MODEL Evaluate the performance of classifier.

31) To get the input from user and perform numerical operations (MAX, MIN, AVG, SUM, SQRT, ROUND) using in R.

32) To perform data import/export (.CSV, .XLS, .TXT) operations using data frames in R.

33) To get the input matrix from user and perform Matrix addition, subtraction, multiplication, inverse transpose and division operations using vector concept in R.

34) To perform statistical operations (Mean, Median, Mode and Standard deviation) using R.

35) To perform data pre-processing operations

i) Handling Missing data ii) Min-Max normalization

36) To perform dimensionality reduction operation using PCA for Houses Data Set

37) To perform Simple Linear Regression with R.

38) To perform K-Means clustering operation and visualize for iris data set

39) Write R script to diagnose any disease using KNN classification and plot the results.

40) To perform market basket analysis using Association Rules

41) Creation of a Data Warehouse.

42)Apriori Algorithm

43)One Hierarchical clustering algorithm.

44)Bayesian Classification

45)Decision Tree Induction algorithm

46)Support Vector Machine

47) Applications of classification for web mining.

48) Case Study on Text Mining.

49) Visualize data using any plotting framework

50) Study and implementation of Data Visualization with ggplot2

LIST OF SOFTWARE:

- Hadoop
- YARN
- R Package
- Hbase
- MongoDB
- Excel