21BDH201 Reg No : .....

CHOICE BASED CREDIT SYSTEM

M.Sc. SECOND SEMESTER DEGREE EXAMINATION AUGUST 2022

BIG DATA ANALYTICS

Big Data Architechure & Hadoop Ecosystem

Duration:3 Hrs **Max Marks:70**

PART - A

I. Answer any EIGHT of the following: **(2×8= 16)**

a) Highlight the concept of ETL. Where is it useful?

b) Name the default port numbers on which Name Node, Job Tracker and Task Tracker run in Hadoop.

c) Discuss the data model followed by HBase to store data.

d) Describe the use of the Expand command in Apache Cassandra.

e) Enumerate any four Apache projects that fit into the Hadoop Ecosystem.

f) Enumerate the benefits of Solr.

g) Discuss the use of Apache Docker for a big data environment.

h) Enumerate the models supported by Avro.

i) Highlight the significant features of Flume in distributed systems.

j) Outline the role of Nagios in a distributed environment.

PART - B

Answer any FOUR questions : **(6×4= 24)**

2. With a suitable example demonstrate why Shuffle and Sort are intermediary steps in MapReduce execution flow.

3. Distinguish between the three different forms of big data with suitable examples.

4. Explain how YARN divides the functionality of a Job Tracker into different services.

5. Discuss the Key-value data stores and Document data store models of NoSQL.

6. Explain how companies use Big Data for marketing citing suitable examples.

7. Describe the three fundamental configuration units in Apache Chef.

PART - C

Answer any THREE questions : (10×3= 30)

8. Consider a database Course with details of Course ID, Course name, Course duration, Course instructor and Course category. Write MongoDB operations to demonstrate the following. 1. Insert multiple documents. 2. Display all the documents. 3. Display the students enrolled for the course Python Programming 4. To display the students who have enrolled for courses of duration 6 months. 5. To update the existing course Java to Networking with Java.

9. Consider a Restaurant review database with details of Restaurant Name, Address, Cuisine type, and the Rating given by customers for the Restaurant. Demonstrate creation of this database using Blur. Display the deatils of all restaurants which are in the city of Mangaluru.

10. Consider the dataset of India's population census 2020. The data is collected from each house, both urban and rural. Describe the contents of the dataset and write 10 diverse analysis that can be carried out with the data.

11. With an illustration, explain in detail the functionality of the two main components of HDFS.

12. Explain in detail the services provided by the secure gateway Apache Knox and Apache Sentry for components in the Hadoop Ecosystem.

****

21BDH202 Reg No : .....

CHOICE BASED CREDIT SYSTEM

M.Sc. SECOND SEMESTER DEGREE EXAMINATION AUGUST 2022

BIG DATA ANALYTICS

Machine Learning and Deep Learning

Duration:3 Hrs **Max Marks:70**

PART - A

I. Answer any EIGHT of the following: **(2×8= 16)**

a) What are the different methods to map the data into higher dimensional space in Support Vector Machines?

b) What is the purpose of scikit-learn?

c) Describe the importance of adding interaction features and polynomial features to enrich feature representation.

d) How do you analyze the behavior of classifiers at different thresholds?

e) What is the use of loss functions?

f) Compare sigmoid and tanh functions.

g) Describe the need for softmax regression.

h) Describe the method of constructing convolution kernel in case of input data with multiple channels.

i) With suitable python code, describe how to visualize decision trees.

j) Identify the relationship between model complexity and dataset size.

PART - B

Answer any FOUR questions : **(6×4= 24)**

2. How do you compute k for K Nearest Neighbors algorithm?

3. Would you use PCA on large datasets or there is a better alternative?

4. How does deep learning work?

5. Describe the evolution of Convolutional Neural Network.

6. Explain the advantages of Cross Validation.

7. Write a short note on deferred initialization.

PART - C

Answer any THREE questions : **(10×3= 30)**

8. How do you differentiate between supervised and unsupervised learning?

9. Explain the various strategies used to identify the best features.

10. Write a note on Vanishing and Exploding Gradients.

11. How do you implement Convolutions for images? Explain.

12. Explain the clustering methods in detail with example.

****

21BDS203 Reg No : .....

CHOICE BASED CREDIT SYSTEM

M.Sc. SECOND SEMESTER DEGREE EXAMINATION AUGUST 2022

BIG DATA ANALYTICS

Multivariate Techniques for Data Analysis

Duration:3 Hrs Max Marks:70

PART - A

I. Answer any EIGHT of the following: (2×8= 16)

a) Mention any two applications of multivariate analysis.

b) What does ANOVA stand for and how do you calculate it?

c) Where is logistic regression used in real-life?

d) Give some real-life examples for multinomial logistic regression.

e) How do you calculate average linkage clustering?

f) Differentiate K-means Clustering and K-Nearest Neighbor Clustering methods.

g) Why do we use canonical correlation analysis?

h) What do you mean by Type 1 error? Describe with an example.

i) Derive the equation of logistic regression.

j) Describe non-metric measurement scale with its types.

PART - B

Answer any FOUR questions : (6×4= 24)

2. Find the correlation coefficient between the variables for the following data.

X	8	6	12	14	16	10
Y	15	10	20	25	30	20

3. What are the issues present in factor analysis? Explain each of them.

4. List out the steps included in the clustering process with an example.

5. Briefly explain the assumptions of ANCOVA?

6. Briefly explain the assumptions of Linear Discriminant Analysis.

7. Explain the effect of multicollinearity.

PART - C

Answer any THREE questions : **(10×3= 30)**

8. The following are the heights of 10 persons and one each of their sons

Father	158	160	163	165	167	170	167	172	177	181
Son	163	158	167	170	160	180	170	175	172	175

a. Find the regression equation.

b. Find the most probable height of a person whose father is 184 cms.

9. Write a note on Principal Component Analysis.

10. The following dataset contains eight items representing colored points on the x-y plane.

x	y	color
1	1	red
1	3	green
2	5	blue
3	5	green
4	1	blue
4	4	red
5	3	blue
5	4	green

Using this data as the training set, run the k-nearest neighbor method to decide the most likely color for a new item with x=3 and y=3.

i) Find the color assigned to the new item when k=1

ii) Find the color assigned to the new item when k=4

11. Define and explain MANCOVA.

12. To study the performance of three detergents and three different water temperatures, the following "whiteness" readings were obtained with specially designed equipment

Water temp	Detergent A	Detergent B	Detergent C
Cold water	57	55	67
Warm water	49	52	68
Hot water	54	46	58

Perform a two-way analysis of variance using 5 percent level of significance.

****

21BDS207 Reg No : .....

CHOICE BASED CREDIT SYSTEM

M.Sc. SECOND SEMESTER DEGREE EXAMINATION AUGUST 2022

BIG DATA ANALYTICS

Natural Language Processing

Duration:3 Hrs **Max Marks:70**

PART - A

I. Answer any EIGHT of the following: **(2×8= 16)**

a) Enumerate two most commonly used tokenizers.

b) Discuss the context in which stemming should be avoided in NLP.

c) Enumerate the important uses of POS tags.

d) Describe the need for Relation Extraction in NLP applications.

e) Outline the working of a Recursive-Descent Parser.

f) Outline the working of a Chart Parser.

g) Describe the typical process incorporated in a text classifier.

h) With an example, describe the application of NLP in Information Retrieval.

i) Discuss the applications of Logistic Regression algorithm in NLP.

j) With an example, explain the purpose for Web Scraping.

PART - B

Answer any FOUR questions : **(6×4= 24)**

2. Develop a python program using Genism library to remove stop words.

3. Describe the typical processes that are followed in a typical machine translation application with suitable illustration.

4. Summarize the advantages and disadvantages of Unigram, Bigram and Trigram taggers.

5. Consider the sentence "The President speaks about the educational reforms". Explain the concept of chunking for the given example.

6. Describe the working of a typical Dialog System.

7. Using Scikit-learn's CountVectorizer( ) convert a collection of sentences to a vector of term counts.

PART - C

Answer any THREE questions : **(10×3= 30)**

8. Discuss the applications of NLP in a. Routing Support Tickets. b. Sentiment Analysis. c. Business Data Analysis d. Customer Service Chatbots e. Interactive Voice Response (IVR) Systems.

9. With an illustration explain how topics are allocated to documents using topic modelling.

10. Using CFG for a suitable text of sentences, develop a python programme to generate the parse tree. Explain the grammar applied.

11. Explain the importance of 1. Machine Translation 2. Optical Character Recognition and 3. Language detection for NLP applications.

12. Consider a CSV file containing SMS messages, where each message contains two attributes, SMS_ID, SMS_DATA. Develop a python program to perform five data cleaning operations on the given file.

****

21BDE215 Reg No : .....

CHOICE BASED CREDIT SYSTEM

M.A/M.Com./M.Sc. SECOND SEMESTER DEGREE EXAMINATION AUGUST 2022

BIG DATA ANALYTICS

Fundamentals of Data Analytics

Duration:3 Hrs **Max Marks:70**

PART - A

I. Answer any EIGHT of the following: **(2×8= 16)**

a) What do you understand by the term function in Python?

b) Develop a python program to print even integers from 0 to 100.

c) Develop a python program to demonstrate sorting using the in-built sort function.

d) Describe the two methods used to print the correlation and covariance of a series as a dataframe.

e) Explain the use of seed() in numpy.random.

f) List the different types of data transformation in Pandas.

g) Explain the concept of pivoting in Pandas.

h) Discuss the concept of Regular Expression in Python programming.

i) Discuss the purpose of the matplotlib package in Python Programming.

j) Give an example to set the X and Y axis labels in pyplot.

PART - B

Answer any FOUR questions : **(6×4= 24)**

2. With a suitable example, explain the concept of classes in Python.

3. Explain the concept of Statements and Expressions in Python programming language with examples.

4. Explain the concept of sorting with a dataframe considering both index and axis.

5. Demonstrate the concatenate function using any three series of your choice.

6. Differentiate between
   1. index( ) and find( )
   2. split( ) and strip( ) when performing string manipulation.

7. Develop python code to create two subplots and generate a sine wave and a cosine wave in each of them.

PART - C

Answer any THREE questions : **(10×3=30)**

8. 1. Develop a python program to print the sum of the first N natural Numbers. (5M)
   2. Develop a python program to find the largest of N numbers. (5M)

9. 1.Compare and contrast lists and tuples in Python language with examples. (5M)
   2. Develop a Python program to count the even numbers, odd numbers in a given input list. (5M)

10. 1. Given the multidimensional array 'X' = [[3, 5, 2, 4], [7, 6, 8, 8], [1, 6, 7, 7]], write commands to print the following:

    1. Display the element in the third row, fourth column.
    2. Display the element in the second row, first column.
    3. Displays the first 2 rows and 3 columns.
    4. Display the reverse of the array. (5M)

    2. With an example demonstrate how to create

    1. a fixed type integer array in python
    2. to create a numpy integer array. (5M)

11. With suitable examples, explain the different methods to handle missing values in Pandas.

12. Develop a Python program to create a pie chart of the popularity of various programming Languages using the sample data given.
    Programming languages: C++, Java, Python, C, JavaScript
    Popularity: 24.4, 16.7, 18.8, 7.6, 19.3

****