22BDAH201 Reg No : .....

CHOICE BASED CREDIT SYSTEM

M.Sc. SECOND SEMESTER DEGREE EXAMINATION MAY/JUNE 2023

BIG DATA ANALYTICS

Big Data Architechure & Hadoop Ecosystem

Duration:3 Hours Max Marks:70

PART A

I. Answer any FOUR of the following (4×5= 20)

a) Describe the suitability of clustered computing as an approach for big data solutions.

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

c) With an illustration, describe the internal mechanism of a Kafka Cluster.

d) Using Parquet syntax, develop code to create an external table to store Movie reviews with reviewer name, movie title and rating details and load the Parquet format file in PIG environment.

e) Discuss the need for Data Transfer tools in the Hadoop Ecosystem.

PART B

Answer any FIVE questions selecting at least one question from each unit

(5×10= 50)

UNIT-I

2. Describe the various charateristics and types of Big Data with suitable examples.

3. 1. Discuss how cluster computing is an effective approach for big data solutions. (5M)

   2. Explain the steps required to deploy a Big Data Solution. (5M)

UNIT-II

4. Consider a text file containing details of utility bills paid during one financial year. The following is an example of the text file: Electricity 2020-01 2000 Gas 2020-01 700 Telephone 2020-01 1150 Electricity 2020-02 2230 Gas 2020-02 850 Telephone 2020-02 1350 ..... Write the Mapper and Reducer code using Python Programming language to find the total amount paid for each utility.

5. Create a database with information on winners of various marathons assuming suitable data and perform query operations using Apache Cassandra.

a. Display the points scored by runners who participated in 10km category in ascending order.
b. Display the total points scored by all players in the 10 km category.
c. Display the total number of medals won by players from a particular country.
d. Demonstrate the creation of Indexes.
e. Display the medals won by athletes in the London Marathon for the recent 5 years.

UNIT-III

6. Explain the two approaches to indexing for data management in Hadoop.

7. Create a database with Festival details and perform query operations using Apache Hive.

a. Display the festivals in the month March.
b. Display the festivals which are celebrated for more than 5 days in Karnataka.
c. Display the details of Harvest festivals of India.
d. Display the count of festivals per region.
e. Display the festival which is celebrated for the maximum number of days.

UNIT-IV

8. Highlight the Machine Learning algorithms suported by MLlib in Apache Spark.

9. Highlight the three primary categories of tools for managing and monitoring the Hadoop architecture. Describe the services provided by Apache Ambari in the Hadoop cluster environment.

****

22BDAH202 Reg No : .....

CHOICE BASED CREDIT SYSTEM

M.Sc. SECOND SEMESTER DEGREE EXAMINATION MAY/JUNE 2023

BIG DATA ANALYTICS

Machine Learning and Deep Learning

Duration:3 Hours Max Marks:70

PART A

I. Answer any FOUR of the following **(4×5= 20)**

a) Explain why python language is prefered for machine learning and data analysis.

b) Develop a code to build K-means clustering model and its prediction for the make_blobs dataset.

c) Explain the concept of shuffle-split cross-validation with a suitable illustration.

d) Explain Loss function in multilayer perceptrons.

e) Elaborate on GPU capacity with respect to Neural Network.

PART B

I. Answer any FIVE questions selecting at least one question from each unit: **(5×10= 50)**

UNIT-I

2. Explain with illustration, the Linear regression model on make_wave dataset available in mglearn.

3. Discuss on the concept of generalization and overfitting in supervised machine learning with suitable example.

UNIT-II

4. Explain diagramatically RobustScaler and MinMaxScaler.

5. Explain the following dimensionality reduction techniques:
   a) PCA
   b) Non-Negative Matrix Factorization

UNIT-III

6. Elaborate on Model Selection, Underfitting, and Overfitting with respect to Neural Networks.

7. Discuss on any two concrete situations where covariate or concept shift might not be obvious.

UNIT-IV

8. Explain with an illustration, Cross-correlation computation with Multiple Input Channels.

9. Explain the implementation of Convolutions for images in two-dimensional cross-correlation operation.

****

22BDAH203 Reg No : .....

CHOICE BASED CREDIT SYSTEM

M.Sc. SECOND SEMESTER DEGREE EXAMINATION MAY/JUNE 2023

BIG DATA ANALYTICS

Multivariate Techniques for Data Analysis

Duration:3 Hours Max Marks:70

PART A

I. Answer any FOUR of the following **(4×5= 20)**

a) Explain the procedure to fix Multi-Collinearity during model building.

b) Provide a summary of 05 uses of Multiple regression model.

c) Elaborate on direct or inverse correlation with suitable examples.

d) Explain different distance measures under cluster analysis.

e) Differentiate between discriminant analysis versus logistic analysis.

PART B

Answer any FIVE questions selecting at least one question from each unit

(5×10= 50)

UNIT-I

2. Explain any 05 applications of Multivariate Techniques.

3. Explain 5 real-world cases where logistic regression is effectively used.

UNIT-II

4. Differentiate between CFA and EFA and discuss the pros and cons of Factor Analysis.

5. Using PCA convert the following dataset and find the first four components.

X	5	6	8	9
Y	6	5	7	8

UNIT-III

6. Explain the designing of conjoint analysis.

7. Derive the cannonical correlation between the random vector X and Y.

UNIT-IV

8. Explain the advantages and disadvantages of ANCOVA.

9. Elaborate on the need of MANOVA and its assumptions.

****

22BDAS207 Reg No : .....

CHOICE BASED CREDIT SYSTEM

M.Sc. SECOND SEMESTER DEGREE EXAMINATION MAY/JUNE 2023

BIG DATA ANALYTICS

Natural Language Processing

Duration:3 Hours **Max Marks:70**

PART A

I. Answer any FOUR of the following **(4×5= 20)**

a) Develop a Python program to demonstrate the concept of rare word removal.

b) Explain the generation of parse tree for an example sentence using a suitable Context Free Grammer.

c) Describe the concept of topic modeling in NLP Applications.

d) Describe the working of a typical Dialog System.

e) Develop a Web Scrapper that crawls and scraps 3 levels of headers from a given website.

PART B

Answer any FIVE questions selecting at least one question from each unit:

(5×10= 50)

UNIT-I

2. Develop a python program to demonstrate the application of various types of Sequential taggers as Backoff taggers.

3. 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.

UNIT-II

4. a. Consider the sentence "The manager preferred the morning train through Kunigal". Explain the working of Dependency parsing for the given text with suitable illustration.

   b. Explain the concept of chunking with a suitable example. (5+5)

5. Explain the concepts of Named Entity Recognition and Relation extraction. Demonstrate through a python program the generation of NERs for a text file using NLTK.

UNIT-III

6. Develop a python program to iterate over a list of sentences and rank the sentences by a score based on the fraction of tokens being entities as compared to to a normal token. Generate a tuple with details of sentence number and score for all the sentences.

7. Consider two documents - Doc1 containing the phrase "Big Data Analytics" and Doc2 containing the phrase "Data Analytics". Explain the working of Vector Space Model for information retrieval with a suitable illustration for the query string "Hadoop" given the two documents.

UNIT-IV

8. Consider a tagged dataset containing both spam and non spam messages. Each message is tagged as Spam or NotSpam accordingly. Develop basic functions using Python to perform text preprocessing and classify the text using Naive Bayes classifier.

9. Describe in detail the various steps during the data flow in Scrapy.

****

22BDAE215 Reg No : .....

CHOICE BASED CREDIT SYSTEM

M.Sc./M.Com SECOND SEMESTER DEGREE EXAMINATION MAY/JUNE 2023

Data Analytics using Python

Duration:3 Hours **Max Marks:70**

PART A

I. Answer any FOUR of the following **(4×5= 20)**

a) Elaborate on the concept of classes using a python program.

b) Develop a python code for the function isnull() and specify the expected output.

c) Explain pandas dataframe. Develop a python code for creating a pandas dataframe.

d) With an example demonstrate the merging of dataframes based on keys in Pandas.

e) Develop a python program to demonstrate plotting of a sine and cosine graph using matplotlib.

PART B

Answer any FIVE questions selecting at least one question from each unit:

(5×10= 50)

UNIT-I

2. Elaborate on the concept of tuples with suitable examples. Develop a python code for tuple creation.

3. Elaborate on break and continue statements with suitable examples.

UNIT-II

4. Elaborate on Python numpy Comparison Operators.

5. Given the statements: import numpy as np and x = np.arange(10), what is the output of the following?

```
x
x[:5]
x[5:]
x[4:7]
x[::2]
x[-1]
x[1::2]
x[::-1]
x[5::-2]
x[-2]
```

UNIT-III

6. Consider the following string : An investment in knowledge, pays the best interest. Develop python code to perform the following: 1. Split the given comma-seperated string into pieces. 2. Check if the word "knowledge" is present in the string. 3. Find the first occurence of the letter "i". 4. Count the total number of occurrences of the characters "in". 5. Replace the space between words with a comma.

7. Consider an Iris dataset (The dataset contains details of 3 species of the Iris flower and the features are Sepal Legth, Sepal Width, Petal Length, Petal width, Species Name). Develop code using python to demonstrate the following. 1. To display basic statistical properties of the dataset. 2. To check for null values. 3. To have a count of the null values. 4. To display the first two rows of the dataset. 5. To display details of flowers having sepal length greater than 2.5.

UNIT-IV

8. Elaborate on Matplotlib library.

9. Develop a Python program to create a pie chart of medal achievements of five most successful countries in 2020 Summer Olympics. Add a suitable title to the chart. Display the medal details suitably.

****