

RAMAKRISHNA MISSION VIDYAMANDIRA
(Residential Autonomous College affiliated to University of Calcutta)
M.A./M.Sc. THIRD SEMESTER EXAMINATION, DECEMBER 2024
SECOND YEAR [BATCH 2023-25]

Date : 03/12/2024
Time : 11 am – 1 pm

COMPUTER SCIENCE
Paper : CC 12

Full Marks : 50

Answer **any five** of the following questions:

[5×10]

1. a) 50 years of annual record is used to compute the mean annual rainfall at Sundarban. The mean is equal to 1460 mm. Is the population mean significantly different from 1500 mm at a level of significance of 0.05? Assume the population standard deviation as 150 mm. Given that $Z_{0.025} = \pm 1.96$.

- b) Briefly explain the Mapreduce job with an example.

- c) What do you mean by RDD transformation and action? Explain with example.

(3+4+3)

2. a) For an aptitude test for 20 respondents, among which 10 were men and 10 of them were women, the scores for each of the respondents are recorded in the following table below:-

Men	90	79	80	85	54	64	72	76	80	65
Women	85	80	70	60	82	91	90	50	65	56

Using appropriate hypothesis testing, check whether the distribution of scores is the same for men as women for 5% significance level. Critical values of the Mann-Whitney U distribution are given as following:-

n1	n2	0.10	0.05	0.025	0.01	0.005	0.001
10	9	62	66	70	74	77	82
10	10	68	73	77	81	84	90

- b) A botanist plants 40 seeds and lets them grow for two months under different conditions for sunlight exposure and watering frequency. After two months, he records the height of each plant. The results are shown below:

	Sunlight Exposure			
Watering Frequency	None	Low	Medium	High
Daily	4.8	5	6.4	6.3
	4.4	5.2	6.2	6.4
	3.2	5.6	4.7	5.6
	3.9	4.3	5.5	4.8
	4.4	4.8	5.8	5.8
Weekly	4.4	4.9	5.8	6
	4.2	5.3	6.2	4.9
	3.8	5.7	6.3	4.6
	3.7	5.4	6.5	5.6
	3.9	4.8	5.5	5.5

Now he wants to know whether or not plant growth is influenced by sunlight exposure. Is sunlight exposure statistically significant at alpha level 0.05? Given that F-critical value for degree of freedom 3 at 5% significance level is 2.90112.

(5+5)

3. a) What are the advantages of using Dataframes over traditional data structures in data analysis?
b) Provide an example of how a Dataframe simplifies operations on large datasets.
c) Describe the key features of HDFS and explain how it supports fault tolerance and scalability in big data processing. (2+3+5)
4. a) What is Named Entity Recognition (NER) systems and how they are useful in NLP? Explain with example.
b) Describe the process of feature engineering in NLP what are some common features extracted from text data for machine learning models.
c) What are the main Challenges in parsing natural language sentences? Explain role of syntax trees. (4+3+3)
5. Find out the edit distance and alignment between the two strings “METAMORPHOSIS” and “HOMOMORPHISM”, considering an equal cost (say, 1) for insertion, deletion, and substitution. (10)
6. a) If a model has high precision but low recall, what does this imply about its performance?
b) Given the following training documents, compute which class the test document belongs to.

	Doc	Words	Class
Training	1	software update release	Technology
	2	government election law	Politics
	3	new tech gadgets	Technology
	4	fitness diet exercise	Health
	5	debate policy government	Politics
	6	healthy living diet	Health
Test	7	tech policy update	?

You are given the following prior probabilities:

$$P(\text{Technology}) = 0.33$$

$$P(\text{Health}) = 0.33$$

$$P(\text{Politics}) = 0.33$$

(2+8)

7. Write a short note with examples of **any two** of the following: [2×5]
 - a) Zipf's law and its application in natural language
 - b) Use of tf-idf in feature engineering for text data
 - c) JSON and its use in computational NLP
 - d) Role of Laplace Smoothing in Naïve Bayes classification
 - e) Precision and Recall in classification tasks

_____ × _____