

RAMAKRISHNA MISSION VIDYAMANDIRA

(Residential Autonomous College affiliated to University of Calcutta)

M.A./M.Sc. THIRD SEMESTER EXAMINATION, JANUARY 2023

SECOND YEAR [BATCH 2021-23]

COMPUTER SCIENCE

Paper : CC12

Date : 11/01/2023

Time : 11.00 am – 1.00 pm

Full Marks : 50

Answer **any five** questions :

[5×10]

1. a) Define two sample t-test. In which cases One-sample, two-sample, or paired t-tests are applicable?
- b) What is the statistical significance of p-value ?
- c) A certain bag of fertilizer advertises that it contains 7.25 kg , but the amounts these bags actually contain is normally distributed with a mean of 7.4 kg and a standard deviation of 0.15 kg . The company installed new filling machines, and they wanted to perform a test to see if the mean amount in these bags had changed. They took a random sample of 50 bags and observed a sample mean and standard deviation of 7.36 kg and 0.12 kg respectively.
 - i) What conclusion should be made using a significance level of $\alpha = 0.05$?
 - ii) How would the conclusion have changed if they had instead used a significance level of $\alpha = 0.01$?

The t-distribution critical values are given as following :-

confidence level									
D.F.	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
40	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496

[3+2+5]

2. a) A pharmaceutical company is testing a new drug to see if it helps reduce the time to recover from a fever. They decide to test the drug on three different races (Hindu, Muslim, and Christian) and both genders (male and female). They randomly select five test subjects from each of those six treatments. The response variable is the time in minutes after taking the medicine before the fever is reduced. The recover times are recorded as follows :-

	Male	Female
Hindu	53, 72, 43, 56, 52	46, 51, 33, 47, 41
Muslim	33, 30, 26, 25, 29	18, 21, 34, 40, 24
Christian	54, 49, 59, 39, 55	25, 29, 47, 26, 28

Using two-way ANOVA test, show how the different factors affect the response variable for significance level of $\alpha = 0.05$.

Critical values of F for the 0.05 significance level are given as following :-

	1	2	3	4	5	6	7	8	9
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.38	2.32
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30

- b) What are the features of RDD in Spark?
3. a) Explain the role of different type of nodes to write a new file into HDFS .
- b) What is the difference among Rdd, Dataframes and Datasets ?
- c) How can you judge a classification with the help of a confusion matrix?

[8+2]

[3+2+5]

4. a) Find out the edit distance and alignment between the two strings “CALCULATOR” and “COMPUTER”, considering an equal cost (say, 1) for insertion, deletion, and substitution. [8+2]
- b) State and explain Zipf’s Law.
5. a) Naive Bayes classifier has an important similarity to language modelling. Explain.
- b) Given the following training documents, compute which class the test document belongs to.

	Doc	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?

- c) Discuss how you could compute class-specific (i.e. per class) precision, recall, and accuracy when dealing with multiple classes. [2+6+2]
6. a) Explain in details the masked language model
- b) How does masked language model differ from Causal language Model? [7+3]
7. Write short notes on **any two** of the following : [2×5]
- Word tokenization
 - Semantic Analysis
 - Markov Chain
 - Derivational Morphology
 - Uses of TF - IDF

_____ × _____