DATA PREPROCESSING AND ANALYSIS
(CSEN 5231)

Time Allotted : 3 hrs **Full Marks : 70**

Figures out of the right margin indicate full marks.

Candidates are required to answer Group A and any 5 (five) from Group B to E, taking at least one from each group.

Candidates are required to give answer in their own words as far as practicable.

Group – A
(Multiple Choice Type Questions)

1. Choose the correct alternative for the following: **10 × 1 = 10**

(i) Which is the correct order for pre-processing in Natural Language Processing?
(a) tokenization->stemming->lemmatizatio
(b) lemmatization->tokenization->stemming
(c) stemming -tokenization->Quartile deviation
(d) none of above.

(ii) Bag of Words in text pre-processing is a
(a) Feature Extraction Technique (b) Feature Scaling Technique
(c) Feature Selection Technique (d) None of above.

(iii) When conducting an ANOVA, FDATA will always fall within what range?
(a) Between 0 and infinity (b) Between 0 and 1
(c) Between -infinity and +infinity (d) Between -1 and +1.

(iv) A _____ is a line that provides an approximation of the relationship between the variables.
(a) sparkline (b) trendline (c) gridline (d) none of the above

(v) In one-way ANOVA, which of the following is used within the F-ratio as a measurement of the variance of individual observations?
(a) The Sum of Square of Treatments (SSTR)
(b) The Treatment Mean Square (MSTR)
(c) The Residual Sum of Squares (SSE)
(d) The Mean Sum of Squares (MSE).

(vi) Hierarchical clustering is one of the most famous clustering techniques used in
(a) unsupervised machine learning (b) supervised learning
(c) rote learning (d) reinforcement learning.

(vii) Which of the following step is performed by data scientist after acquiring the data?
(a) Data Cleansing (b) Data Integration
(c) Data Replication (d) All of the mentioned.

(viii) Data pre-processing helps in increasing the quality of data by filling in missing incomplete data and
(a) increasing the quality of data by filling in missing incomplete data
(b) smoothing noise
(c) resolving inconsistencies
(d) all of the above.

(ix) With respect to testing of hypothesis, which one of the following is correct?
(a) Type I error is falsely rejecting a null hypothesis
(b) Type II error is falsely rejecting a null hypothesis
(c) Type I error is falsely accepting a null hypothesis
(d) Type II error is falsely rejecting a null hypothesis.

(x) Basic Visualization operations are
(a) Graphical operation (b) Set operation
(c) Data operation (d) All the above.

Group – B

2. (a) What is the difference between structured data and semi structured Data? *[(CO4)(Remember/LOCQ)]*
(b) What is the need for parsing? Explain a mechanism to store unstructured data? *[(CO2)(Understand/IOCQ)]*
6 + (3 + 3) = 12

3. (a) Why is data preprocessing important? *[(CO3)(Remember/LOCQ)]*
(b) Discussed the steps involved in data preprocessing. *[(CO3)(Remember/LOCQ)]*
4 + 8 = 12

Group – C

4. (a) What are data cleaning problems? Classify data quality problems in data sources. *[(CO2)(Remember/LOCQ)]*
(b) How to deal with heterogeneous and missing data? *[(CO3)(Understand/IOCQ)]*
(3 + 3) + 6 = 12

5. (a) Briefly discuss the strategies for data transformation. *[(CO4)(Understand/LOCQ)]*
(b) Normalize the group of data: 1000, 2000, 3000, 9000.
Using min-max normalization by setting min=0 and max=1. *[(CO1)(Understand/IOCQ)]*
(c) Perform the z-score normalization on the following set of data:
12, 13, 15, 16, 19, 22, 28, 40, 60, 134. *[(CO1)(Understand/IOCQ)]*
6 + 3 + 3 = 12

Group – D

6. (a) Suppose each housewife uses the washing powder which comparative statistics t-test should a scientist of *Hindustan Unilever Ltd* perform.

Surf Excel Quick Wash	
Names	Wash Percentage
Mina	89%
Ruby	80%
Sarmishta	85%
Natasha	75%

Surf Excel Easy Wash	
Names	Wash Percentage
Mina	75%
Ruby	80%
Sarmishta	60%
Natasha	65%

Explain the t-test mechanism that will be performed. *[(CO4)(Analysis/HOCQ)]*

(b) Explain 3 central tendency mechanism with an example of your own. *[(CO2)(Understand/IOCQ)]*

(c) Explain Mann-Whitney test mechanism. *[(CO1)(Remember/LOCQ)]*

4 + 5 + 3 = 12

7. (a) The manufacturer states that the average lifetime of a light bulb with extra brightness is 70 hours. Competitive firm believes that it is in fact lower, so decided to prove that the manufacturer's claim is not correct. Randomly selected 20 light bulbs and found that their average life was 67 hours and the standard deviation was 5 hours. Significance level α = 0:05 verify whether the manufacturer's claim is actually incorrect or not. *[(CO4)(Analyze/HOCQ)]*

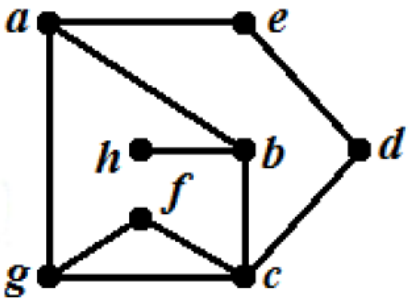
(b) What do mean by null hypothysis? *[(CO5)(Remember/LOCQ)]*

(c) What are the differences between K means and hierarchical clustering? *[(CO5)(Remember/LOCQ)]*

6 + 2 + 4 = 12

Group – E

8. (a) Represent the following graph using list, and, matrix representations: *[(CO5)(Understand/IOCQ)]*



(b) Discuss the steps in designing visualization of data. *[(CO5)(Remember/LOCQ)]*

6 + 6= 12

9. (a) What are the tools for time series analysis. Explain how do you use one of the tools. *[(CO1)(Remember/LOCQ)]*

(b) Explain one of the techniques to find correlation between two parameters in a dataset. *[(CO2)(Understand/HOCQ)]*

6 + 6 = 12

Cognition Level	*LOCQ*	*IOCQ*	*HOCQ*
Percentage distribution	*53.12*	*30.21*	*16.67*

Course Outcome (CO):

After the completion of the course students will be able to

1. Acquire knowledge in a broad range of methods based on statistics and informatics for data preprocessing and analysis and tools for visualizing the main characteristics of data.
2. Understand the whole process line of gathering relevant data, preprocessing the data, performing exploratory analysis on the data and visualizing the implicit knowledge extracted from data.
3. Apply suitable methods for unveiling the underlying structure of the data, testing underlying assumptions in various fields.
4. Analyze the results of experiment with the help of various visualization tools and statistical tests.
5. Evaluate the performance of not only a computational method after obtaining different results by using different parameter values in order to choose the correct parameter value, but also, all similar methods in order to find out the best performing algorithm for a dataset.
6. Get familiar with relevant literatures, derive theoretical properties of the existing methods and come up with novel approach or pipeline for analyzing data across various fields by solving assignment problems

**LOCQ: Lower Order Cognitive Question; IOCQ: Intermediate Order Cognitive Question; HOCQ: Higher Order Cognitive Question*