National University of Computer and Emerging Sciences, Lahore Campus

Course:	Data Analysis and Visualization	Course Code:	DS3001
Program:	BS (Data Science)	Semester:	Fall 2022
Duration:	60 Minutes	Total Marks:	45
Paper Date:	27-Sept-2022	Weight	15 %
Section:	A	Page(s):	8
Exam:	Midterm 1	Roll No.	

Instruction/Notes: Attempt all questions. Answer in the space provided. You can ask for rough sheets but will not be attached with this exam. Answers written on rough sheet will not be marked. Do not use pencil or red ink to answer the questions. In case of confusion or ambiguity make a reasonable assumption.

Q1: You have to remove noise from a **5-bit** grayscale image. As you know images are collection of pixels so I have picked a 5x5 subset of the 1024x1024 grayscale image. This subset contains some noise pixels as well. [20 Marks]

1. You have to identify the type of noise? [2 Marks]
2. What filter you will choose and why? The filter size must be 3x3. [4 Marks]
3. What will be the output after applying filter (**Convolution**) on the subset of the image? [12 Marks]
   a. Original size of image must remain same after applying filter? [2 Marks]
   b. What kind of padding you are going to use (if needed)?
   c. You have to take care of object geometry/structure as well during filtering process.

Input Image (Subset):

7	7	12	7	12
7	7	7	7	7
7	7	12	7	7
7	7	7	7	7
12	7	7	7	12

Noise Type: Gaussian → 2

Filter Type: Linear Filter → 2+2

Padding Type: Extended → 2 marks

Filter Values:

1/9	1/9	1/9
1/9	1/9	1/9
1/9	1/9	1/9

2 correct cell = 1 mark

Output Image:

7	8	8	9	9
7	8	8	9	8
7	8	8	8	7
8	8	8	8	8
9	8	7	8	9

2 cells = 1 mark

Possible Variations

$$\begin{bmatrix} 7 & 7 & 7 \\ 7 & 7 & 7 \\ 7 & 7 & 7 \end{bmatrix} = 7 \qquad \begin{bmatrix} 7 & 7 & 7 \\ 7 & 7 & 7 \\ 7 & 7 & 12 \end{bmatrix} = 8$$

$$\begin{bmatrix} 7 & 7 & 7 \\ 7 & 7 & 7 \\ 7 & 12 & 12 \end{bmatrix} = 8 \qquad \begin{bmatrix} 7 & 7 & 7 \\ 7 & 7 & 7 \\ 12 & 12 & 12 \end{bmatrix} = 9$$

$$\begin{bmatrix} 7 & 7 & 7 \\ 7 & 7 & 12 \\ 12 & 12 & 12 \end{bmatrix} = 9$$

Q2: In text-based processing model, the very basic model named Bag of Words (BoW) used to extract features from given corpus. [13 Marks]

1. How BoW works?
2. Is there any limitation to BoW? If yes then state it.
3. Transform following text to into BoW format that will be acceptable by classifiers.

Text Corpus:

Text1 = "Data science is an interdisciplinary field focused on extracting knowledge from typically large data set"

Text2 = "Data science is an interdisciplinary field focused on applying the knowledge and insights from that data to solve problems"

Data => 0
science => 1
is => 2
an => 3
interdisciplinary => 4
field => 5
focused => 6
on => 7
extracting => 8
knowledge => 9
from => 10
typically => 11
large => 12
set => 13
applying => 14
the => 15
and => 16
insights => 17
from => 18
that => 19
to => 20
solve => 21
problems => 22

3 marks

Unique Vocabulary

Brief Algorithms steps how Bog Works? [5 Marks]

Selects Unique Vocabulary from the corpus, counts its frequency and makes a feature vector.

↳ 2 marks

Text1's BoW representation: [2 Marks]

Sentence 1
2 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0

23 dimensional feature Vector

Text2's BoW representation: [2 Marks]

Sentence 2
2 1 1 1 1 1 1 1 0 1 0 0 0 0 1 1 1 1 1 1 1 1 1

Limitations: [4 Marks]

=> 24 dimensional fv

Increase in Corpus results in huge feature vector size.
Context Free Algorithm

Data ≠ data

Working for the above Question

Q3: We have data as show below there are some missing values, your task is to substitute suitable values in missing places. Nature of data is time series. [2+10 Marks]

1. Identify what method you will use to impute missing values? Linear Regression
2. Show details steps for processing towards imputation to missing values.

↳ 2 marks

Dataset:

Note: Data points format ➔ (X, Y)

Serial Number	X	Y
1	1	2
2	3	4
3	5	6
4	7	8
5	9	10
6	11	12
7	13	14
8	15	16
9	17	18
10	19	20

4 => 2.5

2.5 <= 7

14 => 2.5

17 <= 2.5

2.5 marks

$y = ax + b$

$a = 1.04$

$b = 0.44$

Working for the above Question

Formula for regression(if needed)

$$a = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

$$b = \frac{1}{n}(\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i)$$

Rough Work

Rough Work