NATIONAL INSTITUTE OF TECHNOLOGY PATNA
Department of Computer Science and Engineering
MID SEMESTER EXAMINATION, Mar 2024
M. Tech. (CSE) 2nd Sem/ PhD

Course Name: Data Visualization Techniques **Max. Marks: 60**
Course Code: CS540203 **Maximum Time: 3 hours**

Instruction:

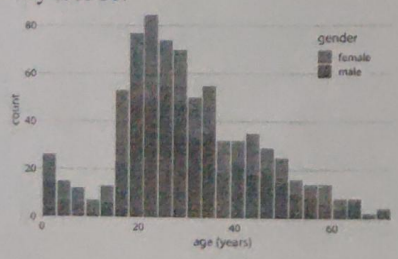
1. **Attempt all questions. The total number of questions is Five.**
2. **Assume any suitable data, if necessary.**
3. **Answer all the questions in the order as appeared in the question paper and put all sub-parts of a question in one place.**

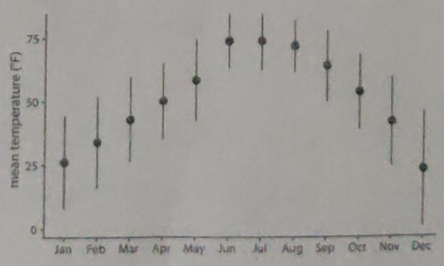
BL: 1-Remember, 2-Understand, 3-Apply, 4-Analyze, 5-Evaluate

S.N.	Questions	Marks	CO	BL
1	(a) Explain the importance of color choice in data visualization. Provide examples of how appropriate color selection can enhance understanding and interpretation of data.	5	1	2
	(b) Explain the role of Python libraries such as Matplotlib, Seaborn, and Plotly in data visualization. Compare and contrast these libraries, highlighting their strengths and weaknesses.	5	3	4
	(c) Discuss any five data distribution visualization plots with suitable examples.	5	2	2
2	(a) The following table indicates the data on the number of patients visiting a hospital in a month. Using the measuring tendency process, find the following: (I) Mean of patients visiting the hospital in a day (II)Median of the patients visiting the hospital in a day	9	4	3

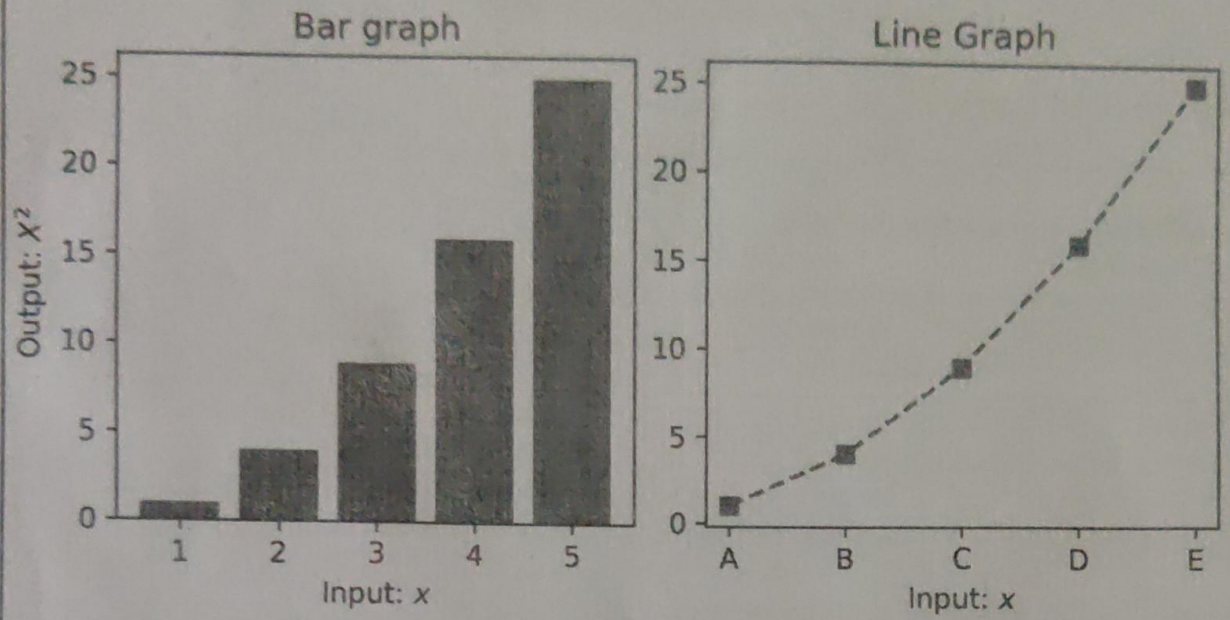
Number of patients	Number of days visiting the hospital
1-10	3
11-20	7
21-30	10
31-40	8
41-50	5
51-60	3

	(b) Compute the central tendency (mean, median and mode) of the following data. 2, 4, 3, 5, 6, 5, 4, 7, 8, 6, 7, 5, 4, 3, 5, 6, 7, 3, 1, 4	6	2	3
3	(a) What are symmetric and asymmetric binary numbers? Give one example of each.	4	1	1
	(b) How is dependency-oriented data different from non-dependency-oriented data? Explain the following dependency-oriented data with their precise definition, examples and supporting visualization tools and techniques. (I) Multivariate Time-Series Data (II) Network and Graph Data	6	1,4	4
4	(a) The following figures will not be considered good figures. Give your point of view on why it is so.	4	4	3





	(b) What is Geospatial data? Discuss the basic components of Geospatial data required for visualization, such as Poles, Equator, Longitude, latitude, and Altitude. Also, discuss some supporting tools to visualize such data.	6	4,5	1

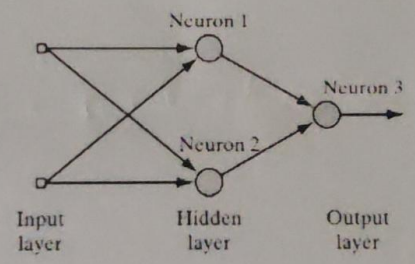
5	Write the Python code to plot the given data considering bar and line graphs in a single figure. X= [1, 2, 3, 4, 5]  	10	2,3	3

All the best

NATIONAL INSTITUTE OF TECHNOLOGY PATNA
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
END SEMESTER EXAMINATION, Jan-June, 2024

Programme: M. Tech (Data Science and Engineering) PhD **Semester**: 2nd
Course Code: CS540202 **Course Name**: Deep Learning
Full marks:60

Answer all questions.
The use of *calculator* is allowed.

Q. no.	Question	Marks	CO	BL
1	a) Discuss the momentum based gradient descent algorithm to train a MLP in detail with all necessary equations.	06	CO1	Remembering
	b) How does the use of regularizers reduce the chance of overfitting in any neural network? Explain with suitable example and proper justification.	06	CO1	Analysis
2	a) How does the continuous Hopfield network lead to the generation of stochastic neurons? Explain. How the energy function is measured for the network containing stochastic neurons?.	05	CO2	Remembering
	b) Use the back-propagation algorithm for computing a set of synaptic weights and bias levels for a neural network structured as in the following figure to solve the XOR problem. Assume the use of a logistic function for the nonlinearity 	07	CO1	Application
3	a) Why was the restricted Boltzmann machine (RBM) invented? Which situations demand the use of RBM that cannot be solved using the normal Boltzmann machine? Explain with proper justification	06	CO3	Analysis
	b) Discuss the architecture of a deep belief network with suitable diagram and the algorithm to train this network.	06	CO3	Remembering
4	a) Discuss with the help of suitable diagrams the structure of an RBFN based on the interpolation theory and the structure of a practical RBFN.	05	CO4	Remembering
	b). How is the structure of the practical RBFN influenced by the use of the *K*-means clustering algorithm? How are the centers of	07	CO4	Understanding , Analysis

	different clusters determined by the *K*-means algorithm to initialize the positions of the RBF centers? Discuss with all the necessary equations.			
5	a) What are the advantages of using second-order recurrent neural network over the fully connected recurrent neural network? Discuss with the help of the architectures of the network and all necessary equations.	07	CO5	Analysis
	b) Show that the recurrent multilayer perceptron model can be represented by the state space model: $$\mathbf{x}_{n+1} = \mathbf{f}(\mathbf{x}_n, \mathbf{u}_n)$$ $$\mathbf{y}_n = \mathbf{g}(\mathbf{x}_n, \mathbf{u}_n)$$ where u_n denotes the input, y_n denotes the output, x_n denotes the state, and f(.,.) and g(.,.) denote vector-valued nonlinear functions	05	CO5	Analysis, Application

NATIONAL INSTITUTE OF TECHNOLGY PATNA
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
END-SEMESTER EXAMINATION - MAY, 2024

M.Tech – Data Science & Engineering

CS540213 – Recommendation Systems

Max.Marks:60 IInd Semester Time: 3 hours

Note: Answer All questions, and all parts of the same question must be answered at the same place

Q. No	Question	Marks	CO	BL
1	a) Draw the high-level architecture of Content-based recommendation system and discuss about each and every module?	07	CO2	Understand
	b) Differentiate content-based recommendation system and collaborative recommendation system.	03	CO2	Understand
2	a) What is long tail phenomenon and mention any two issues with respective recommendation system can be inferred from long tail phenomenon.	04	CO1	Remember
	b) Differentiate user-based collaborative filtering and item-based collaborative filtering. List any two advantages of user-based collaborative filtering over item-based collaborative filtering. List any two limitations of user-based collaborative filtering over item-based collaborative filtering.	06	CO3	Understand
3	Mention and explain the two ways used for estimating the posterior of the raing r_{uj} (for user ***u*** and item ***j***) in Naive Bayes Collaborative Filtering Algorithm (NBCF). Consider the rating matrix with 7 items and 6 users, where ***1*** indicates user likes the item, ***-1*** indicates user dislikes the item and ***?*** indicates missing value. Predict all the missing entries by using NBCF.	10	CO4	Apply & analyze

ItemID→ USerID ↓	i_1	i_2	i_3	i_4	i_5	i_6	I_7
U_1	1	-1	-1	1	1	-1	1
U_2	1	1	1	?	-1	-1	1
U_3	-1	1	-1	1	1	1	1
U_4	1	1	1	1	1	-1	-1
U_5	?	1	1	-1	1	1	1
U_6	1	-1	1	1	-1	1	1

4	a) Differentiate the following i) Novelty and serendipity ii) Confidence and trust	02	CO5	Understand
	b) Mention the division of the rating matrix in Netflix Prize data set and discuss about each division.	02	CO5	Remember
	c) **Match the following**	06	CO5	Remember

1) Sparsity	i) Averages Precision and Recall with bias toward the weaker value.
2) Precision	ii) The average deviation between computed recommendation scores and actual rating values for all evaluated users and all items in their testing sets.
3) Recall	iii) The proportion of relevant instances that are retrieved
4) Accuracy	iv) The proportion of true results (both true positives and true negatives) among the total number of cases examined
5) F1	v) The ratio of empty and total entries in the user – item matrix.
6) MAE	vi) The proportion of retrieved instances that are relevant

5	a) Define the following terms related to the attacks in recommendation system **i)** Push attack **ii)** Nuke attack **iii)** Size of attack **iv)** Filler item	04	CO6	Remember
	b) Let I_F be the ratings for the filler items, I_T be the ratings for the target item, r_{max} is the maximum rating value, r_{min} is the minimum rating value. Match the following:	06	CO6	Apply & Evaluate

1) Random Attack	i) I_F = Most disliked items rated r_{min} $I_T = r_{min}$
2) Segment Attack	ii) I_F = Popular items rated r_{max} and others rated randomly $I_T = r_{max}$
3) Bandwagon Attack	iii) I_F = Normal distribution around system mean $I_T = r_{min} / r_{max}$
4) Love/Hate Attack	iv) I_F = If it is less than the global mean value of all items then it is set to r_{min}, otherwise $r_{min} + 1$ $I_T = r_{max}$
5) Popular Attack	v) $I_F = r_{min}$ vi) $I_T = r_{max}$
6) Reverse Bandwagon Attack	vii) $I_F = r_{max}$ $I_T = r_{min}$

6	a) Define latent factor model?	02	CO6	Remember
	b) Consider a utility matrix ***R*** consists of six users and 12 items.	08	CO6	Apply

1		3			5			5		4	
		5	4			4			2	1	3
2	4		1	2		3		4	3	5	
	2	4		5			4			2	
		4	3	4	2					2	5
1		3		3			2			4	

Assume that we applied an oracle based latent factor model and we got the following matrices U and V.

U =

.2	-.4	.1
.5	.6	-.5
.5	.3	-.2
.3	2.1	1.1
-2	2.1	-.7
.3	.7	-1

V^T =

-.9	2.4	1.4	.3	-.4	.8	-.5	-2	.5	.3	-.2	1.1
1.3	-.1	1.2	-.7	2.9	1.4	-1	.3	1.4	.5	.7	-.8
.1	-.6	.7	.8	.4	-.3	.9	2.4	1.7	.6	-.4	2.1

Find the approximate matrix P (assume k =3).

********** ALL THE BEST **********

NATIONAL INSTITUTE OF TECHNOLGY PATNA
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
End Semester Examination Jan - May 2024

M. Tech. (DS) 2nd Semester. *Max.Marks: 60*

Date: 16-05-24 ***Time: 3 Hrs.*** ***CS540221 – Big Data Analytics***

Instructions:
1. Attempt all questions.
2. Assume any suitable data, if necessary. (Any other Instruction need to provide by the concerned faculty)

	Questions	Marks	CO	BL
1	a. What is Hadoop? Explain the architecture of Hadoop 2.x in detail. b. Explain Hbase architecture and its components in detail. c. Explain Hive architecture and its services in detail.	[5+5+5]	CO1 CO2 CO4	Remember Understand
2	(a) Explain the Spark architecture in detail. (b) What is Resilient distributed dataset (RDD)? Expain in detail. (c) Explain the execution of spark application in detail.	[5+5+5]	CO2 CO3	Apply Understand Remember
3	Implement the following relational algebra operations using MapReduce algorithm, and explain each algorithms with an example. i) union ii) intersetion iii) difference iv) projection v) groupBy with aggregation	[10]	CO3	Analyze, apply
4	Design MapReduce algorithms for the following algorithms. a) ANOVA test. b) K-Nearest Neighbor (KNN) Read the following table [**Table 1: Student preformance dataset**] and analyze the performance of each gender in various subjects by applying MapReduce based ANOVA test. Also apply MapReduce based KNN to predict the gender if the test data is [**math_score: 55, science_score: 75, english_score: 65**].	[20]	CO3	Analyze Apply

Table 1: Student preformance dataset

gender	math_score	science_score	english_score
female	72	72	74
female	69	90	88
female	90	95	93
male	47	57	44
male	76	78	75
female	71	83	78
female	88	95	92
male	40	43	39
male	64	64	67

National Institute of Technology Patna

End Semester Exam, Session: Jan_June 2024

Program: M.Tech/Ph.D. **Semester: 2nd** **Department: CSE**
Subject Name: Natural Language Processing **Subject Code: CS540201**
Time: 3 hrs **Full Marks: 60**

Assume any missing data and/or conditions. All questions are compulsory and the question paper is of two pages

Sl. No.	Question	CO	BL
1	Consider the costs of operations **insertion** and **deletion** to be 1 each while the cost of substitution to be 2 in edit distance calculation. Calculate the edit distance between strings **GOING** and **COMING**. Also specify the number of insertions, deletions and substitutions for the optimal alignment of both words.	CO1	Understand
2	Consider the following corpus C_3 of four sentences. $< s >$ three friends amar akbar and anthony are reading book $< /s >$ $< s >$ amar is reading malgudi days $< /s >$ $< s >$ akbar is reading a detective book $< /s >$ $< s >$ anthony is reading a book by rk narayan $< /s >$ a. Assume a bi-gram language model. Calcuate P($< s >$ amar is reading a book $< /s >$). b. Consider the same Bi-gram model, this time with Laplace (Add-one) smoothing. Calculate P($< s >$ akash is reading story book $< /s >$). c. Consider the same Bi-gram model, what is the Perplexity of the sentence $< s >$ akar is reading a detective book $< /s >$. – laplace	CO2	Apply
3	Find the part of speech (POS) tags for the words of the sentence "Jane will spot will" given the following data for training where the letter in bracket indicates the word's POS tag. Mary(N) Jane(N) can(M) see(V) will(N). Spot(N) will(M) see(V) Mary(N). Will(M) Jane(N) spot(V) Mary(N). Mary(N) will(M) pat(V) spot(N). a. Draw transition, emission and initial probabilities. b. Estimate the POS tags for words of a sentence **Jane will spot will** using the Viterbi algorithm.	CO3	Analyze

4	Consider the context-free grammar given below. S → NP VP NP → NP PP | we | noodles | chopsticks PP → IN NP IN → with VP → V NP | VP PP V → eat a. Use the CKY algorithm to check whether the string **We eat noodles with chopsticks** can be generated by the above grammar. b. Draw the parse trees generated by the above algorithm.	CO3	Apply
5	Consider the story "Once upon a time there lived a poor widow and her son Jack / 1. One day, Jack's mother told him to sell their only cow / 0. Jack went to the market and on the way he met a man who wanted to buy his cow / 1. She said, "You fool! He took away your cow and gave you some beans!" She threw the beans out of the window / 0. Jack was very sad and went to sleep without dinner / 1." The 1's at the end of the sentence indicate that these sentences are included in the summary and 0's indicate that these sentences are not included in the summary. You are required to summarize stories like these by training a Convolutional Neural Network (CNN) having 2 layers. Assume an embedding technique to create sentence embedding of dimension 6. Apply 2-gram with 10 filters with a stride of 1 in both layers. There is a dense layer followed by an output layer with two neurons. a. Draw the structure of the convolution network. b. How many trainable parameters are there in the network? c. Find the intermediate result after both the convolution operations.	CO2	Analyze
6	What is a self-attention mechanism? How to calculate self-attention in a transformer model. Consider four words and apply the self-attention mechanism to these four words w1=[1, 0, 0], w2=[0, 1, 0], w3 = [1, 1, 0], w4 = [0, 0, 1] and report the attention weight of each words. For this w_q = [[.24, 0.1, .20], [0.1, 0.2, 0.6], [.3, .6, .5]] w_k = [[.20, 0.01, .27], [0.10, 0.02, 0.06], [.03, .6, .05]] and w_v = [[.26, 0.01, .23], [0.10, 0.11, 0.06], [.07, .30, .005]].	CO3	Evaluate

NATIONAL INSTITUTE OF TECHNOLOGY PATNA
END SEMESTER EXAMINATION Jan-June 2024
M.Tech. (CSE) 2nd Semester

Course Name: **Bioinformatics**
Course Code: **CS540210**

Duration: **3 Hrs.**
Full Marks: **60**

Instructions:

a) Attempt all questions.
b) Assume any suitable data, if necessary.
c) Answer all the questions in the order as appeared in the question paper and write all the sub-parts of a question in one place.

S.N.	Questions	Marks	CO	BL
Q1.	a) Why SWISS-PROT is important in biological data retrieval? b) What data types can be retrieved from KEGG database? c) What are genomic primary databases? Mention the significance of each database.	[2] [2] [6]	CO1	Remember
Q2.	Differentiate between the following: a) Monophyletic and paraphyletic trees b) Cladogram and phylogram c) Neighbourhood and global optimization-based methods of protein network analysis d) Pairwise and multiple sequence alignment e) Clustal Omega and T-coffee	[10]	CO2	Understand
Q3.	a) How to represent phylogenetic trees computationally? Give a suitable example. b) Draw a bifurcating phylogenetic tree showing operational taxonomic units (OTUs) and clade. Calculate number of rooted and unrooted trees if number of taxa is 6. c) How to identify motif/domain in sequence analysis?	[2] [6] [2]	CO2 & CO3	Understand & Apply
Q4.	a) How E-value can be used in BLAST? b) How iterative method can be implemented in multiple sequence alignment works? Illustrate it using suitable diagram. c) How decision trees can be used for protein-protein interactions?	[3] [3] [4]	CO2	Understand
Q5.	a) Given a set of sequence pairs, x and y: x: …WRNDCQEGSA… y: …WGQEGSIEA… Determine the "best" local alignment between them via trace-back procedure using Smith-Watermann algorithm. b) Given sequence x: TTGCAAACGC, construct the dot-plot against itself.	[5] [5]	CO3	Apply

Q6.	a) How to compute the reconstruction of a phylogenetic tree using neighbor-joining (NJ) method? Illustrate the steps. b) Consider the pairwise evolutionary distance matrix of the set of six OTUs set {a, b, c, d, e, f} as given below:	[5] [5]	CO2 & CO3	Understand & Apply

	a	b	c	d	e
b	2				
c	4	4			
d	6	6	6		
e	6	6	6	4	
f	8	8	8	8	8

Construct a phylogenetic tree using UPGMA method for the given data.

*****All the best*****