NATIONAL INSTITUTE OF TECHNOLOGY PATNA

Department of Computer Science and Engineering

MID SEMESTER EXAMINATION, Mar 2024

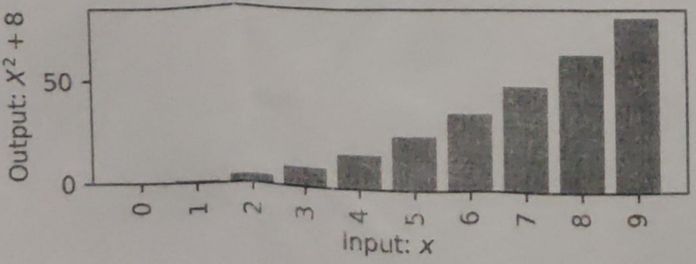
M.Tech.(CSE) 2nd Sem/ PhD

Course Name: Data Visualization Techniques **Max. Marks: 30**

Course Code: CS540203 **Maximum Time: 2 hours**

Instruction:

1. Attempt all questions.
2. Assume any suitable data, if necessary.
3. 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.

<table>
<tr><th>S.N.</th><th>Questions</th><th>Marks</th><th>CO</th><th>BL</th></tr>
<tr><td>1(a)</td><td>Briefly describe the data visualization and its importance in enterprise.</td><td>3</td><td>CO1</td><td>Remember</td></tr>
<tr><td>1(b)</td><td>Describes the following:
i) Data Objects ii) Attributes iii) Nominal data type iv) Ratio-based data type</td><td>4</td><td>CO1</td><td>Apply</td></tr>
<tr><td>1(c)</td><td>What do you understand by dependency-oriented data? Explain Multivariate time series with suitable examples.</td><td>3</td><td>CO4</td><td>Understand</td></tr>
<tr><td>2(a)</td><td>Explain the five-number summary (Minimum, Q1, Medium, Q2, Maximum) and plot the following data in the form of boxplot.
3, 0, 12, 0, 2, 0, 26, 0, 7, 5, 5, 2, 1, 1, 2</td><td>5</td><td>CO3</td><td>Apply</td></tr>
<tr><td>2(b)</td><td>Let's say we have recorded data about research articles having attributes like publishers and article types. Compute the correlation between Publisher and Article types.
<table>
<tr><th>Index</th><th>Publishers</th><th>Article Types</th></tr>
<tr><td>1</td><td>Elsevier</td><td>Book</td></tr>
<tr><td>2</td><td>Elsevier</td><td>Book</td></tr>
<tr><td>3</td><td>Elsevier</td><td>Journal</td></tr>
<tr><td>4</td><td>Wiley</td><td>Journal</td></tr>
<tr><td>5</td><td>Wiley</td><td>Conference</td></tr>
<tr><td>6</td><td>Wiley</td><td>Book</td></tr>
<tr><td>7</td><td>Elsevier</td><td>Book</td></tr>
<tr><td>8</td><td>Wiley</td><td>Book</td></tr>
</table>
</td><td>5</td><td>CO2</td><td>Apply</td></tr>
<tr><td>3(a)</td><td>What are the different ways to create plots in the matplotlib? Write the Python code to generate the figure given below. Consider the function f(x)=x^2 +8, where x ranges from 0 to 9, figure size (4,2), and ticks have a rotation of 90-degree at x-axis.


</td><td>6</td><td>CO3</td><td>Apply</td></tr>
<tr><td>3(b)</td><td>Briefly explain the cartesian coordinate system and polar coordinate system with suitable examples.</td><td>4</td><td>CO3</td><td>Understand</td></tr>
</table>

All the best

NATIONAL INSTITUTE OF TECHNOLOGY PATNA

Department of Computer Science and Engineering

MID SEMESTER EXAMINATION, March 2024

B.Tech.(CSE) 4nd Sem

Course Name: Bioinformatics | **Max. Marks: 30**

Course Code: CS540210 | **Maximum Time: 2 hours**

Instructions:

1. Attempt all questions.
2. Assume any suitable data, if necessary.
3. 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
1.a)	Why UniGene is important in biological data retrieval?	2	CO1	Remember
b)	Which biological database can be used to retrieve and compare sequences of the same gene from three species?	1		
c)	What CATH stands for? Why it is used?	2		
2.	Differentiate between the following:	5	CO1	Remember
a)	Transcription and translation			
b)	mRNA and rRNA			
c)	tBLASTn and BLASTx			
d)	RefSeq and GenBank			
e)	PAM and BLOSUM			
3. a)	Why tRNA is important protein synthesis?	2	CO2	Understand
b)	Can the base sequence of an mRNA be predicted from the amino acid sequence of its polypeptide product?	1		
c)	How BLAST algorithm works?	2		
4. a)	Is DNA synthesis semidiscontinous? If so, why?	2	CO2	Understand
b)	How progressive method of multiple sequence alignment works? Illustrate it using suitable diagram.	5		
c)	How the scoring function can be implemented using sum of pairs method to align four DNA sequences?	3		
5.	Given a set of sequence pairs, a and b: a: CTCGT b: CTAAGT Determine the "best" global alignment between them via trace-back procedure using Needlemann-Wunsch algorithm	5	CO3	Apply

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

NATIONAL INSTITUTE OF TECHNOLOGY PATNA

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

MID SEMESTER EXAMINATION, March, 2024

Programme: M. Tech (Data Science and Engineering)/PhD **Semester**: 2nd

Course Code: CS540202 **Course Name**: Deep Learning

Full marks:30

Answer *all* questions.

The use of *calculator* is allowed.

Q. no.	Question	Marks	CO	BL
1	Discuss the Gradient Descent algorithm to train a MLP in detail with all necessary equations. Consider both the cases of any neuron *i* when it is present in the output layer and when it is present in the hidden layer.	07	CO1	Remembering
2	a) Discuss the Hopfield network algorithm (both *training* and *recall*) with proper explanation.	05	CO2	Remembering
	b) Suppose that a five-node Hopfield network has stored the patterns $X^{(1)}$=+++++, $X^{(2)}$=+-+-+, $X^{(3)}$=+++-- If the new pattern -+++- is presented to the net, find out the stored pattern to which the net converges.	07	CO2	Application
3	a) Which drawback of discrete Hopfield network has led to the generation of continuous Hopfield network? Why are stochastic neurons used in a continuous Hopfield network? Why specifically was the Boltzmann machine invented where both Boltzmann machine and Hopfield network contain stochastic neurons?	07	CO2, CO3	Analysis
	b) Mathematically derive the Boltzmann learning rule with all necessary equations.	04	CO3	Understanding

NATIONAL INSTITUTE OF TECHNOLGY PATNA
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
MID-SEMESTER EXAMINATION - MARCH, 2024

M.Tech – Data Science & Engineering

CS540213 – Recommendation Systems

Max.Marks:30 IInd Semester Time: 2 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) We want to design a recommendation system for a restaurant that serves the entire world. The restaurant has over 5000 food items and one million customers, but its rating database has only 10 million ratings. Which of the following would be a better recommendation system, and justify your answer?
(i) User-based collaborative filtering.
(ii) Item-based collaborative filtering.
(iii) Content-based recommendation system.
(iv) Popularity-based recommendation system (Discussed in the class)

3M | CO1 CO2 CO3 | Create, Evaluate &Analyze

b) Suppose the restaurant is using the recommendation system you suggested in the above (a). Consider a customer who rated only two items, ***Rasgulla*** and ***Gulab Jamun***, and both ratings are five and five on a scale of five. Which of the following items is most likely to be recommended with proper justification:
(i) Jalebi **(ii)** Motichoor Laddu **(iii)** It depends only on other users' ratings.
(iv) Samosa **(v)** It depends only on other item's ratings

3M | CO1 CO3 | Analyze

c) Assume that there is a very close neighbours $v_1, v_2, \ldots v_k$ of u. The users $v_1, v_2, \ldots v_k$ have not rated the target item ***t*** yet. Consider the number of nearest neighbours is ***k***. Propose an efficient solution for the user-based collaborative filtering to predict the rating for the item ***t*** for user ***u***.

4M | CO1 CO3 | Apply

2 a) Let ***n*** be the number of users, ***m*** be the number of items, ***p*** be the number of nearest neighbours, and ***k*** be the number of recommendations for each user. Assume that the ***j**th* user has rated all the items and ***i**th* item has rated by all users. Compute the time complexity for finding the similarity between users and similarity between items, and also compute the space complexity for storing similarity values in both cases. Also, find the time complexities for predicting the k recommendations for each user in both user-based collaborative filtering and item-based collaborative filtering.

3M | CO3 | Remember & understand

b) Justify the statement, "The cold start problem regarding the new items can be addressed using the content-based recommendation system."

3M | CO2 | Analyze

c) Consider the following dataset consisting of five features and a class label (likes/dislikes):

4M | CO2 | Remember, Understand, Apply

Keyword **Song Id**	Drums	Guitar	Beat	Orchestra	Classical	Like/ Dislike
201	1	0	1	0	1	Like
202	0	0	1	0	1	Like
203	0	0	1	1	0	Dislike
204	1	1	0	1	1	Dislike
205	1	1	0	1	1	Like

Using the Bayes Method, predict the class label for the following two test samples: **(i)** (1, 1, 1, 0, 1) **(ii)** (1, 0, 1, 0, 0)

3 a) Consider the following dataset consisting of 16 observations: Remember & Apply

User_Id	Movie_Id	Rating
1	101	3
2	102	4
4	102	2
3	101	5
2	103	1
3	102	1
6	102	2
4	103	4
6	104	5
5	101	3
5	102	5
1	103	3
5	104	2
1	104	2
3	103	5
4	104	5

i) Convert the given dataset into a utility matrix. 2M CO1, CO3

ii) Using the ***cosine similarity*** and ***pearson correlation coefficient***, find the similarities between ***user 4*** and every other user over the raw ratings only. Based on these similarity values, which measure is more efficient for the recommendation systems and why? 3M CO3

iii) Using the similarities calculated in (ii), compute the predicted rating for movie *101* by ***user 4*** using the suitable prediction function, and your prediction function must use the similarity values between users (**Consider the number of similar users/items is four**). 2M CO3

b) Consider a scenario where two users have only a small number of common ratings. Can you mention the issue with this scenario in user-based collaborative filtering? Also, suggest a solution to overcome the mentioned issue. 3M CO3 Create

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

NATIONAL INSTITUTE OF TECHNOLGY PATNA

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Mid Semester Examination Jan - Mar 2024

M. Tech. (DS) 2nd Semester. *Max.Marks: 30*

Date: 21-03-24 *Time: 2 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	What is Big Data? Explain 3Vs of Big Data in detail.	[5]	CO1	Remember
2	a). What is HDFS? Expalin Read and write pipeline on HDFS for a large file. b). Explain MapReduce in detail with all its components.	[5+5]	CO2	Apply Understand Remember
3	Design MapReduce algorithms to take a very large file of integers and produce as output for the following. Also expain each algorithms with example. a) The largest integer. b) The average of all the integers. c) The same set of integers, but with each integer appearing only once. d) The count of the number of distinct integers in the input. e) The set of prime numbers in the input file.	[3*5]	CO2, CO3	Apply Understand

National Institute of Technology Patna

End Semester Exam, Session: July-Dec-2023

Program: M.Tech/Ph.D. **Semester:** 1st

Subject Name: Natural Language Processing **Department: CSE**

Time: 2 hrs **Subject Code: CS540201**

Full Marks: 30

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

1. How many operation will be required to convert Intention to execution. [Consider each operations to be of unit cost]. Show the appropriate locations of the operations. [6]

2. Given the following term document matrix and the number of words in each document, compute the TF*IDF score for each word/document. [6]

Term/Doc	Doc1	Doc2	Doc3	Doc4	Doc5	Doc6	Doc7	Doc8	Doc9	Doc10
Car	3	0	0	5	12	0	0	2	8	1
Auto	8	6	0	12	0	0	9	1	3	10
best	0	1	7	0	1	5	12	0	0	0
Doc Size	40	22	15	38	29	19	47	10	25	26

3. Certain named entity recognition system models each word in the input text as one symbol in the following alphabet Σ: [9]
   - A Uppercase word (e.g. IBM)
   - C Capitalized word (e.g. John)
   - f Functional word (e.g. the, and, a, an, in, of, by, ...)
   - a Lowercase word (e.g. will)
   - 9 Number or code (e.g. 12)
   - p Punctuation (e.g. , . : ;)

   For instance, the sentence: **Tomorrow, John will be 12 years old. He likes music by Adam and the Ants.** would be encoded as **C p C a a 9 a a p C a a f C f f C p**.

   a. Use the given sample to estimate the probabilities $P(xy)$ and $P(xyz)$ for each observed bigram/trigram.

   b. Compute the smoothing of the obtained probabilities using Laplace's Law. Give also the probability for unseen events.

   c. Compute the language model $P(z|xy)\ \forall x, y, z \in \Sigma$ that would result from using each of the two previous estimations. Compare the results, discussing which option is more suitable to model these sequences.

4. Papazom.com also needs to match offers from different suppliers that correspond to the same product, as well as to match user queries with product descriptions. For this, they asked us to propose a similarity model able to establish how similar two product description are. For instance, given the product descriptions. [9]
   - s_1 smartphone Hoewai x23-A with latest super AMOLED display and 64Gb
   - s_2 smartphone x23-A with 64Gb and AMOLED charge indicator
   - s_3 Hoewai smartphone z21-B with super AMOLED display and 32Gb

   a. Represent each description as a word set, and compute $simjac(s_1, s_2)$, $simjac(s_1, s_3)$, and $simjac(s_2, s_3)$ using Jaccard similarity. Jaccard similarity is $simjac(x_1, x_2) = \frac{|x_1 \cap x_2|}{|x_1 \cup x_2|}$

   b. Represent each description as a word-bigram set (i.e set elements are not single words, but word bigrams in the sentece), and compute $simcos(s_1, s_2)$, $simcos(s_1, s_3)$, and $simcos(s_2, s_3)$ using Cosine similarity. Cosine similarity is $simcos(x_1, x_2) = \frac{|x_1 \cap x_2|}{\sqrt{|x_1|}\sqrt{|x_2|}}$

c. A Papazon.com user wrote the search Hoewai smartphone AMOLED display. Compute the similarities of this query with s_1, s_2, and s_3 with each of the above metrics (unigram Jaccard and bigram Cosine).