~~Librarian, AIKTC~~

AIKTC/KRRC/SoET/ACKN/QUES/2022-23/ **Date:** 25/01/23.

School: SoET-REV. C-Scheme **Branch: COMP. ENGG.** **SEM: VII**

To,
Exam Controller,
AIKTC, New Panvel.

Dear Sir/Madam,

Received with thanks the following **Semester/Unit Test-I/Unit Test-II (Reg./ATKT)** question papers from your exam cell:

Sr. No.	Subject Name	Subject Code	Format		No. of Copies
			SC	HC	
1	Machine Learning	CSC701		✓	
2	Big Data Analytics	CSC702		✓	
3	Department Level Optional Course-3	CSDC 701X		✓	
4	Department Level Optional Course-4	CSDC 702X		✓	
5	Institute Level Optional Course-1	ILO 701X			

Note: SC – Softcopy, HC - Hardcopy

10.30 am CO-R-19 12/12/22

Sem-VII - CBGS 19 - Reg

Time: 03 Hours **Marks: 80**

Note: 1. Question 1 is compulsory
2. Answer any three out of the remaining five questions.
3. Assume any suitable data wherever required and justify the same.

Q1 a) What is function of Map Tasks in the Map Reduce framework? Explain with the help of an example. [5]

b) Demonstrate how business problems have been successfully solved faster, cheaper and more effectively considering NoSQL Google's MapReduce case study. Also illustrate the business drivers and the findings in it. [5]

c) Why is HDFS more suited for applications having large datasets and not when there are small files? Elaborate. [5]

d) Explain the concept of bloom filter with an example [5]

Q2 a) Name the three ways that resources can be shared between computer systems. Name the architecture used in big data solutions and describe it in detail. [10]

b) Write a map reduce pseudo code for word count problem. Apply map reduce working on the following document: [10]

"This is an apple. Apple is red in color".

Q3 a) Suppose the stream is 1, 3, 2, 1, 2, 3, 4, 3, 1, 2, 3, 1. Let $h(x) = 6x + 1 \bmod 5$. Show how the Flajolet- Martin algorithm will estimate the number of distinct elements in this stream. [10]

b) Consider the following data frame given below: [10]

subject	class	marks
1	1	56
2	2	75
3	1	48
4	2	69
5	1	84
6	2	53

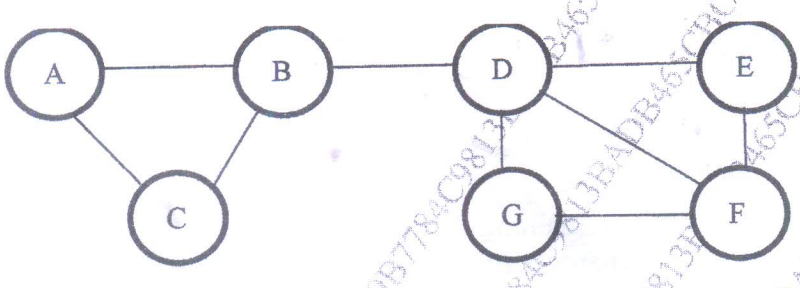
i. Create a subset of subject less than 4 by using subset () function and demonstrate the output.

ii. Create a subset where the subject column is less than 3 and the class equals to 2 by using [ ] brackets and demonstrate the output.

Q4 a) What are the Core Hadoop components? Explain in detail. [10]

b) With a neat sketch, explain the architecture of the data-stream management system. [10]

Q5 a) Determine communities for the given social network graph using Girvan- Newman algorithm. [10]



b) The data analyst of Argon technology Mr. John needs to enter the salaries of 10 employees in R. The salaries of the employees are given in the following table: [10]

Sr. No.	Name of employees	Salaries
1	Vivek	21000
2	Karan	55000
3	James	67000
4	Soham	50000
5	Renu	54000
6	Farah	40000
7	Hetal	30000
8	Mary	70000
9	Ganesh	20000
10	Krish	15000

i. Which R command will Mr. John use to enter these values demonstrate the output.
ii. Now Mr. John wants to add the salaries of 5 new employees in the existing table, which command he will use to join datasets with new values in R. Demonstrate the output.

Q6 a) i. Write the script to sort the values contained in the following vector in ascending order and descending order: (23, 45, 10, 34, 89, 20, 67, 99). Demonstrate the output. [10]
ii. Name and explain the operators used to form data subsets in R.

b) How recommendation is done based on properties of product? Elaborate with a suitable example. [10]

---

10 am

CO - R-19

8/12/22

Sem-VII-C-19-Reg.

Duration: 3hrs **[Max Marks:80]**

N.B. : (1) Question No 1 is Compulsory.
(2) Attempt any **three** questions out of the remaining **five**.
(3) All questions carry equal marks.
(4) Assume suitable data, if required and state it clearly.

Q1. Solve any **four** from following. [20]

a. What are the issues in Machine learning?
b. Explain Regression line, Scatter plot, Error in prediction and Best fitting line.
c. Explain the concept of margin and support vector.
d. Explain the distance metrics used in clustering.
e. Explain Logistic Regression

Q2. a. Explain the steps of developing Machine Learning applications. [10]
b. Explain Linear regression along with an example. [10]

Q3. a. Create a decision tree using Gini Index to classify following dataset. [10]

Sr. No.	Income	Age	Own Car
1	Very High	Young	Yes
2	High	Medium	Yes
3	Low	Young	No
4	High	Medium	Yes
5	Very High	Medium	Yes
6	Medium	Young	Yes
7	High	Old	Yes
8	Medium	Medium	No
9	Low	Medium	No
10	Low	Old	No
11	High	Young	Yes
12	Medium	Old	No

b. Describe Multiclass classification. [10]

Q4. a. Explain the Random Forest algorithm in detail. [10]
b. Explain the different ways to combine the classifiers. [10]

Q5. a. Compute the Linear Discriminant projection for the following two-dimensional dataset. X1= (x1, x2) = {(4,1), (2,4), (2,3), (3,6), (4,4)} and
X2= (x1, x2) = {(9,10), (6,8), (9,5), (8,7), (10,8)} [10]
b. Explain EM algorithm. [10]

Q6. Write detailed note on following. (**Any two**) [20]

a. Performance Metrics for Classification
b. Principal Component Analysis for Dimension Reduction
c. DBSCAN

****

Paper / Subject Code: 42175 / NATURAL LANGUAGE PROCESSING (DLOC - III)

30am CO (R-19) 14/12/22

Time: 3 Hours Sem- VII - C-19- Reg. Max. Marks: 80

N.B. (1) Question No. 1 is compulsory
(2) Assume suitable data if necessary
(3)Attempt any three questions from remaining questions

Q.1	Any Four	20[M]
a	Differentiate between Syntactic ambiguity and Lexical Ambiguity.	[5M]
b	Define affixes. Explain the types of affixes.	[5M]
c	Describe open class words and closed class words in English with examples.	[5M]
d	What is rule base machine translation?	[5M]
e	Explain with suitable example following relationships between word meanings. Homonymy, Polysemy, Synonymy, Antonymy	[5M]
f	Explain perplexity of any language model.	[5M]
Q.2 a)	Explain the role of FSA in morphological analysis?	
Q.2 b)	Explain Different stage involved in NLP process with suitable example.	[10M]
Q.3 a)	Consider the following corpus <s> I tell you to sleep and rest </s> <s> I would like to sleep for an hour </s> <s> Sleep helps one to relax </s> List all possible bigrams. Compute conditional probabilities and predict the next ord for the word "to".	[5M]
Q.3 b)	Explain Yarowsky bootstrapping approach of semi supervised learning	[5M]
Q.3 c)	What is POS tagging? Discuss various challenges faced by POS tagging.	[10M]
Q.4 a)	What are the limitations of Hidden Markov Model?	[5M]
Q.4 b)	Explain the different steps in text processing for Information Retrieval	[5M]
Q.4 c)	Compare top-down and bottom-up approach of parsing with example.	[10M]
Q.5 a)	What do you mean by word sense disambiguation (WSD)? Discuss dictionary based approach for WSD.	[10M]
Q.5 b)	Explain Hobbs algorithm for pronoun resolution.	[10M]
Q.6 a)	Explain Text summarization in detail.	[10M]
Q.6 b)	Explain Porter Stemming algorithm in detail	[10M]

****

10am Sem-VII-C-19-Reg. CO (R-19) 16/12/22

Duration : 3 Hours **Marks : 80 Marks**

N.B. : (1) Question No 1 is Compulsory.
(2) Attempt any three questions out of the remaining five.
(3) All questions carry equal marks.
(4) Assume suitable data, if required and state it clearly.

Q.1 Solve any four.
a. Compare and contrast Boolean Model vs Vector Space Model. 5
b. Specify the significance of User Relevance feedback in an IR system. 5
c. Explain inverted file indexing with suitable examples. 5
d. Explain the process of Structured Text retrieval model. 5
e. Illustrate different types of keyword-based queries. 5

Q.2 a. Draw the taxonomy of IR models and explain any one IR modeling technique. 10
b. What is the significance of **tf** and **idf** ? How can you calculate **tf** and **idf** in a vector model? 10

Q.3 a. Explain the various system related issues faced in Information retrieval systems and how they can be refined for a deployed system. 10
b. State the different types of queries. Explain the pattern matching query concept with an example . 10

Q.4 a. What is the role of suffix array and suffix tree in information retrieval system with example. 10
b. What is Latent Semantic Indexing model ? Write the advantages of Latent Semantic Indexing Model? 10

Q.5 a. Define Multimedia information retrieval. Discuss indexing and searching. 10
b. What is the difference between Unranked Retrieval models and Ranked Retrieval models.

Q.6 Write short notes on any two. 20
a. Information Retrieval in digital libraries.
b. Sequential Searching.
c. Flat browsing vs Hypertext Browsing model.
d. Distributed Information Retrieval.
