Instructions:

1. **Write your name and roll number on the question paper.**
2. **Use of any electronic device or reading material is not allowed.**
3. **Write appropriate explanation/justification/steps wherever necessary.**
4. **Any form of plagiarism will be penalized.**
5. **If anything is not clear, there is a mistake, or some question is incomplete, make appropriate assumptions, mention them and proceed.**

1. Consider a set of 10 elements $\mathcal{S} = \{a, b, c, d, e, f, g, h, i, j\}$. Create a sequence of 20 random and distinct subsets using these elements, with each subset containing 4–8 elements. Answer the following questions based on this.

   (a) Assume that each element in $\mathcal{S}$ is a character and each subset created above is a word. Assume that this sequence of words is arriving in a streaming manner. Calculate the joint probability of the tenth and twentieth word based on the words that have arrived before them. [6 **Marks**]

   (b) Suppose we need to sample around 50% of the streaming words such that the joint probabilities as calculated above are close to the approximate joint probabilities computed after sampling. [16 **Marks**]
   (i) Propose a hash function to perform this hashing and justify why this will be an appropriate hash function for the given task.
   (ii) Apply this hash function and calculate the joint probability of the tenth and twentieth word as in the previous case, and also calculate the error in the probability calculation for both of them.

2. Assume that some data is stored in a distributed manner on a parameter server in appropriate format. Write a pseudo-code of the following algorithms (*any four of your choice*) describing all the necessary steps and assumptions. Also draw a supporting figure to illustrate the steps if required. [40 **Marks**]
   (a) k-Nearest Neighbours
   (b) User-user collaborative filtering
   (c) Page-rank using the power iteration method (using the simplest update rule, the one discussed before the Google's page-rank algorithm)
   (d) Apriori algorithm
   (e) LSH

3. Assume that there is an online store with a fixed set of items. It requires a user to first login before buying any item, and records the log of items bought by each user in a particular session. In this case, each (purchase) session of a user can be represented as a shingle such that the indices corresponding to the items purchased are denoted by 1 and the rest by 0. Over a period of time, each user would have done one or more purchases, where each purchase is represented by a shingle as described above. In other words, over a period of time, there would be a set of shingles corresponding to each user indicating his/her purchase history. To compress these shingles, they are represented using low-dimensional signatures obtained using permutation vectors as in LSH. Now, suppose the online store needs to recommend an item to user based on the similarity of her purchase history (which is represented by a set of signature vectors) with that of other users. Propose a measure of similarity to calculate user-user similarity in this case (scaled to a fixed range of $[-1, 1]$), and justify it using an example. [14 **Marks**]

4. Answer the following question w.r.t. your project of this course.
   (a) What was the title of your course project? Write the name of your team member if the team size was 2. [1 **Mark**]
   (b) Describe in detail what challenge(s) you have addressed in this project, and why they are worth addressing. [4 **Marks**]
   (c) Till what extent do you think the above challenges have been addressed? What more do you think you could have done? [4 **Marks**]

(d) Describe the individual contribution of each team member in the project. [1 **Mark**]
(e) Approximately on which date did you start working on the project? How much time (in hours) had you planned to devote initially, and how much time (in hours) were you able to devote eventually? What were the reasons for deviation(s) in your plan, if any? Plot an approximate month-wise bar-plot for the four months (Jan, Feb, Mar, Apr) denoting the approximate amount of time (in hours) spent on the project per month. Justify. [2 **Mark**]
(f) How many hours of compute (assuming one CPU core) in total was consumed in your project? Comment on whether this is justified w.r.t. the final outcome(s) of your project. [1 **Mark**]
(g) How many marks would you give to yourself out of 100 for the work done on the project? Justify your answer in detail. [2 **Marks**]