Data Interview Problem V2

1. Algorithmic Problem Solving (10 points)

   You are given A, a set of N integers [a_1, a_2, ..., a_n], and each of the integer a_i ranges between A_low and A_high. You want to make sure the difference between the maximum and minimum to be less than or equal to MAX_DIFF and to enforce this, you can add or subtract each number by some integer, but the cost of the operation is proportional to

   the sum of the square of the changes. In other words, you want transform A into another set of integers B = [b_1, b_2, ..., b_n] where max(B) – min(B) <= MAX_DIFF while minimizing the cost = Σ(b_i – a_i)^2.

   Problem Summary:

   Problem Input:

   >size_of_A, MAX_DIFF

   > a set of numbers A = [a_1, ..., a_n] with A_low <= min(A) <= max(A) <= A_high (each number is separated by new line)

   example input:

   4 10

   1

   5

   10

   15

   Problem Solution:

   > a set of numbers B = [b_1, ..., b_n] with max(B)-min(B) <= MAX_DIFF which minimizing the cost = Σ(b_i – a_i)^2> the cost to transform A into B

   Problem Output:

   > The cost to transform A into B

   example output:

   8

   Full Example:

   Input:

   4 10

1

5

10

15

Solution:

→From the input, we know that MAX_DIFF=10 and A=[1,5,10,15].

→ The optimum B for minimizing the cost is: B = [3, 5, 10, 13]

→ with cost = (3-1)^2 + (5-5)^2 + (10-10)^2 + (15-13)^2 = 8

Output:

8

Notes: Your program only need to output the solution which is the minimum cost

Could you design an algorithm (you can use Python, C++, Java, or Scala) to solve this problem given the input restriction:

a. (3 points)

   1 < n < 1000

   1 <= A_low <= a_i <= A_high <= 100

   MAX_DIFF <= A_high - A_low

b. (7 points)

   1 < n < 1000

   1 <= A_low <= a_i <= A_high <= 1000000000

   MAX_DIFF <= A_high - A_low

2. Big Data (10 points)

   a. (5 points) Suppose you are given a text file containing the name (string with at most 256 characters) and the age (integer) of all people in the world (~7 billion people). Now suppose you only have a budget laptop with only 1 GB of RAM. How can you sort this data (first by age then by name) and output the result to a single file? You can use either **C++**, **Java**, **Python**, or **Scala** but you are NOT allowed to use any external library.

      Example (with fewer data input):

      Input:

      person_A 24

      person_B 15

      person_C 52

person_D 15

Output:

person_B 15

person_D 15

person_A 24

person_C 52

b. (2.5 points) Now, suppose you are the owner of an online shop. You have been running your shop for a long time, and have a list of 1 million blacklisted names and phone numbers. Each line contains a single-word name, followed by a space, and then by the corresponding phone number.

Example of blacklist.txt:

=================

Andi 1341441

Melisa 8565467

Aslam 2908345

=================

You want to build an API server that receives a name and a phone number at the input, and then outputs a flag (boolean) whether the name and the corresponding phone number is in the blacklist. How would you write these two functions below to optimize the latency for each API call (no need to write an API server):

- initialize(blacklist)

This function takes string input, which is the filename of the blacklist you have, and is called when the API server starts.

- check_blacklist(name, phone_number)

This function takes 2 arguments: name (string) and phone number (int). This function is called whenever there is a call to the API, and returns a boolean of whether the name and phone number inputted is in the blacklist.

You can use **C++**, **Java**, **Python**, or **Scala** but, once again, you are NOT allowed to use any external library.

c. (2.5 point) Continuing the blacklist on (b), now suppose you are being attacked by hacker. There are a lot of fake name and phone number that you want to blacklist. The blacklist grow into 10 billions of name and phone number! In this case, the blacklist can't longer fit in your RAM. Could design the system that can implement the blacklist to scale to 10 billions name and phone? (Give the concept and design is enough, no need for actual implementation)

3. Machine Learning (10 points)

   a. (2 points) Suppose you have so many features available, but you can't use all of them directly because of computational cost. What would you do in this kind of situation?

   b. (4 points) Suppose you have 10000 close-up images of people's faces; each with its gender label. You used all the labeled images to train a gender classifier model. After you created the model, you tested the accuracy using the training data and you got a very good accuracy. When you deploy it to production, however, the accuracy drops considerably. What are the possible causes to such problem? What will be your proposal to handle them?

   c. (4 points) Again on gender classification, but now you have text data from tweeter. Given a tweet, you want to know whether the person posting the tweet is a male or a female. You intend to use SVM as the classifier. How would you process the raw text data from tweeter such that it can be used as the input to the SVM?

4. Statistics (10 points)

   a. (2.5 points) What is the expected number of die rolls required to get three same consecutive outcomes (for example: a 111, 222, etc) if we use a 6-sided fair die?

   b. (2.5 points) We throw 8 dice and take the sum of the highest 4 outcomes. What is the probability that the sum equals to 24?

   c. (2.5 points) We throw 6 dice at the same time. What is the expected number of distinct outcomes? Example: if the outcomes are 1, 2, 1, 3, 4, and 2, then there are 4 distinct outcomes (i.e. 1, 2, 3, and 4).

   d. (2.5 points) We throw a die 10000 times and record the sum of the outcomes. Approximate the probability that the sum is in the range of [34500, 35500]. (The approximation must be within +/- 0.03%)