##

Old questions:

Q1:

Set 1

What are backup tasks in Hadoop and what are they used for?

Explain how to Hadoop ensures fault tolerance in Map-reduce tasks.

Consider the following scenario and report the task completion times for each task assuming that there are 4 worker nodes each capable of running one task at a time, each mapper task takes 2 units of time to complete and each reducer task takes 2 units of time to complete after it has received the last input record.

There are 3 mapper tasks M1, M2, M3, and 2 reducer tasks R1, R2. M1 and M2 start at t = 0. M3, R1 and R2 starts at t = 1. M2 fails at t = 3, and R2 fails at t = 4. There are no other task failures.

Set 2

What are backup tasks in Hadoop and what are they used for?

Explain how to Hadoop ensures fault tolerance in Map-reduce tasks.

Consider the following scenario and report the task completion times for each task assuming that there are 4 worker nodes each capable of running one task at a time, each mapper task takes 2 units of time to complete and each reducer task takes 2 units of time to complete after it has received the last input record.

There are 2 mapper tasks M1, M2 and 2 reducer tasks R1, R2. M1 and M2 start at t = 0. R1 starts at t = 2, R2 starts at t = 3. M1 fails at t = 1. R1 fails at t = 4. There are no other task failures.

Q2:

Set 1

Write a spark program in Scala / pseudocode for computing a **maximal common substring** present in at least k of the given strings. Assume that the strings are given as records in an RDD. A maximal common substring (MCS) between k strings is a string of maximal length which is a substring of all k strings. For example, DE is a MCS of the below 3 strings but BDE is not MCS.

ABCDE

BCDE

BDE

Set 2

Write a spark program in Scala / pseudocode for computing k nearest neighbors for each data point in an input dataset. Assume that the datapoints are given as records of an input RDD. Note that you cannot store O(n) datapoints in one record of any intermediate RDD, where n is the total number of datapoints. For example, for k = 2 the following input should have the given output.

Input:

(1,1)

(1,2)

(1,3)

(1,4)

(1,5)

Output:

(1,1), ((1,2),(1,3))

(1,2), ((1,1),(1,3))

(1,3), ((1,2),(1,4))

(1,4), ((1,3),(1,5))

(1,5), ((1,3),(1,4))

Q3:

Set 1

Write a pytorch program for the following problem:

Given $\{(x_i, y_i), i = 1, \dots, n\}$ where x_i is a feature vector of d dimensions and y_i is a real number, write a program to learn the parameters of the model $\hat{y}_i = W^T x_i + b$, by minimizing the loss function $l(W, b) = \sum_i (y_i - \hat{y}_i)^2$.

List all the operators you have used in your program and their gradients.

What is the gradient for the tensorflow operator y = tf.maximum(a,b) with respect to a ?

Set 2

Write a pytorch program for the following problem:

Given $\{(x_i, y_i), i = 1, \dots, n\}$ where x_i is a feature vector of d dimensions and y_i is a class label from {0,1}, write a program to learn the parameters of the model $\hat{y}_i = \sigma(W^T x_i + b)$, where $\sigma(a) = \frac{1}{1+e^{-a}}$, by minimizing the loss function $l(W, b) = \sum_i (y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i))$.

List all the operators you have used in your program and their gradients.