CS6240-SHUBHAM DEB

NOTE: My output is of the format List(Brightness) but with the corresponding coordinates also.

EXECUTION STEPS IN SCALA

1) First we get the z-deimension from the imageFile by processing the input file to read the image and get the z dimension from the image

2) I parse the image and by passing the image path and the z dimension to the parser which then retrieves me a list containing the coordinate of the pixel and the distance value.

3) Then I converted the list of string from Java into an RDD of String by parallelizing the list

4) I converted the above RDD into a pairRDD by splitting on **":"** and then filtering on the distance value of the pixel.

If the pixel distance is 2 or 3, I am ignoring that pixel and then I am mapping each node to key and value based on the value. If the value<2, then we make it as **"0"** means foreground, else it is **"1"** or background.

5) Again here I am loading the image file and and converting it into a scala List by using asScala.toList.

6) The RDD is then split on **":"** and then I am filter the pixels based on the proximity to the boundary .

Then each of those filtered pixels is processed to get the neighboring pixels which are then appended to each pixel and then we return an entireRDD containing the neighborNodes' brightness as the value and the coordinate as the key.

5) Then I join the imageRdd and the distRDD on the key to get(coord, adjNodesBrightness, dist).

6) I only get the values of the aboveRDD as the key is not required for training and then saved it as a text file.

7) I continue this process for each of the images i.e each dist and image file.

Discuss how you represent image and distance data in your Spark Scala program

I represented distance data as an RDD(Coordinate of pixel, Distance) and I represented image data as RDD(Coordinate of pixel, list of adjNodes).

What is stored in each row of the RDD/DataSet?

In each row of distance data, I am storing coordinate of the pixel with the corresponding distance from the nearest foreground pixel.
In each row of image data, I am storing RDD(Coordinate of pixel, List of adjacent nodes).

Do you store each image in a separate RDD/DataSet?
I am storing each image in 2 RDD's first containing the distance information and the the second containing the brightness values of the adjacent nodes.
Then I joined on these 2 RDD's to get the brightness values as well as the ditance info by just getting the values of the corresponding joined RDD.