Assignment 5

I am submitting my actual source code for PIC along with its jar. Same is for the 2 bonus questions and their jars. Below are the run instructions

Run Instructions:

For PIC:

Please give 3 parameters for K, g and MaxIterations along with the command as follows:

~spark/bin/spark-submit –master local –class “plot1” target/scala-2.10/PIC-assembly-1.0.jar [k] [g] [MaxIterations]

Q1) [Photo taak] [time lihi]

Q2)[Photo taak] [time lihi]

Q3)[Photo taak] [Time lihi]

Q4) I could not execute it successfully. I tried on Azure with 52 cores but it didn’t work.

Q5) γ is a scaling factor, effectively the “width” of the gaussian kernel that modulates the size of the neighborhood around each data point. The larger the neighborhood, the stronger the connections will be to data points further away (and the harder it will be to differentiate them from closer data points). This means if we keep a larger scaling factor every data point will have a larger neighborhood. So more number of data points will get included in the neighborhood resulting in strongly connected components.

So comparing ques 1 and 2, graph for 1 is not distorted but graph for question2 is pretty distorted as compared to question 1. Distorted means that the inter cluster data points come in each other’s neighborhood.

For question 3, the two clusters are way too far from each other. So even if for an average value of g, the neighborhood span is not large enough to include data points of the other cluster.

Q6) The runtimes for the previous questions were not that much. I mean execution seemed quick to me but the bottleneck does exist. The bottleneck is in similarity calculation. The RDF kernel calculation may be the cause of the time issue.