CME 323: Distributed Algorithms and Optimization

Instructor: Reza Zadeh (rezab@stanford.edu)

HW#2

1. Using the Spark distribution and data you downloaded in hw1, follow the step by step instructions at the following URL:

   `https://databricks-training.s3.amazonaws.com/movie-recommendation-with-mllib.html`

   Note: To get the examples to run, you must first replace the file `machine-learning/scala/build.sbt` in your Spark distribution with the one from the following link:

   `https://web.stanford.edu/~rezab/dao/hw/build.sbt`

   Now, answer the following questions:

   (a) Set the rank to 8, and print the factor for the movie "Saving Private Ryan (1998)"

   (b) Set the rank to 5, and print the factor for the movie "Alien (1979)"

   Submit your code and answers.

2. Using the Spark distribution and data you already downloaded in hw1, follow the step by step instructions at the following URL, skipping to section 4.1 in the tutorial:

   `https://databricks-training.s3.amazonaws.com/graph-analytics-with-graphx.html#getting-started-again`

   Go to section 4.1 in the tutorial, and answer the following questions:

   (a) What are websites with the top 10 pageranks in the wikipedia dataset?

   (b) What are websites with the top 10 indegrees in the wikipedia dataset?

   Submit your code and answers.

3. Give the pseudocode for finding the shortest path between a given source and a destination using the Pregel paradigm. Assume each vertex has a unique ID and each edge has a weight of 1.