PROJECT #1 - MapReduce/Hadoop - DUE DATE: April 2nd, 2017

In this project you will implement KNN clustering algorithm in MapReduce and apply it on synthetic data that you will create.

1. Install Hadoop

2. Create using whatever language you want a very large text file, containing at least 1M data points in the form of (x, y), where x and y are real numbers. The generation of should be biased toward the creation of three clusters. In other words, choose a-priori three centers (x1,y1), (x2,y2) and (x3,y3) and generate the rest of the data points around these, using some random distance following a skewed distribution (towards 0)

3. Move your file to HDFS

4. Implement KNN algorithm as described in the class and apply it in your data set