Please read below for instructions how to run the MapReduce program.

First make sure that the data file are present in the right places, hadoop namenode and datanodes are up.

Assuming the HDFS contains a directory output where the output will be written.
That can easily be changed by changing the path in runKMeans.sh file.

1. **Creating k-random centroids**
    a. Run **setup.py** file, enter the name of the input file and value of k.
    b. This will create a file called "centroids.txt" that will hold k random values from the input file.

2. **Running map reduce**
    a. Run the bash script **runKMeans.sh**, enter the name of the input file as input. This script does multiple things.
        i. Copy the input file to HDFS
        ii. Call the map reduce streaming API with correct parameters
        iii. Loop till convergence

3. **Analyizing and Plotting**
    a. Run the **results.py** file, entering the name of the input file when prompted.
    b. The script will print the Rand co-efficient and Jaccard co-efficient alongwith showing the plot.