作业5

实验要求

在MapReduce上实现K-Means算法并在小数据集上测试。

可以使用附件的数据集，也可以随机生成若干散点的二维数据（x, y）

要求用Matlab或者R语言等工具可视化散点图。设置不同的K值和迭代次数，可视化聚类结果，看是否符合直观。

运行说明

1. 由于之前网站上下载的 Instance.txt打开时格式有问题，在网上查找并修改了Random代码生成新的数据集Instance.txt
2. 进行kmeans聚类，设置不同的k和迭代次数

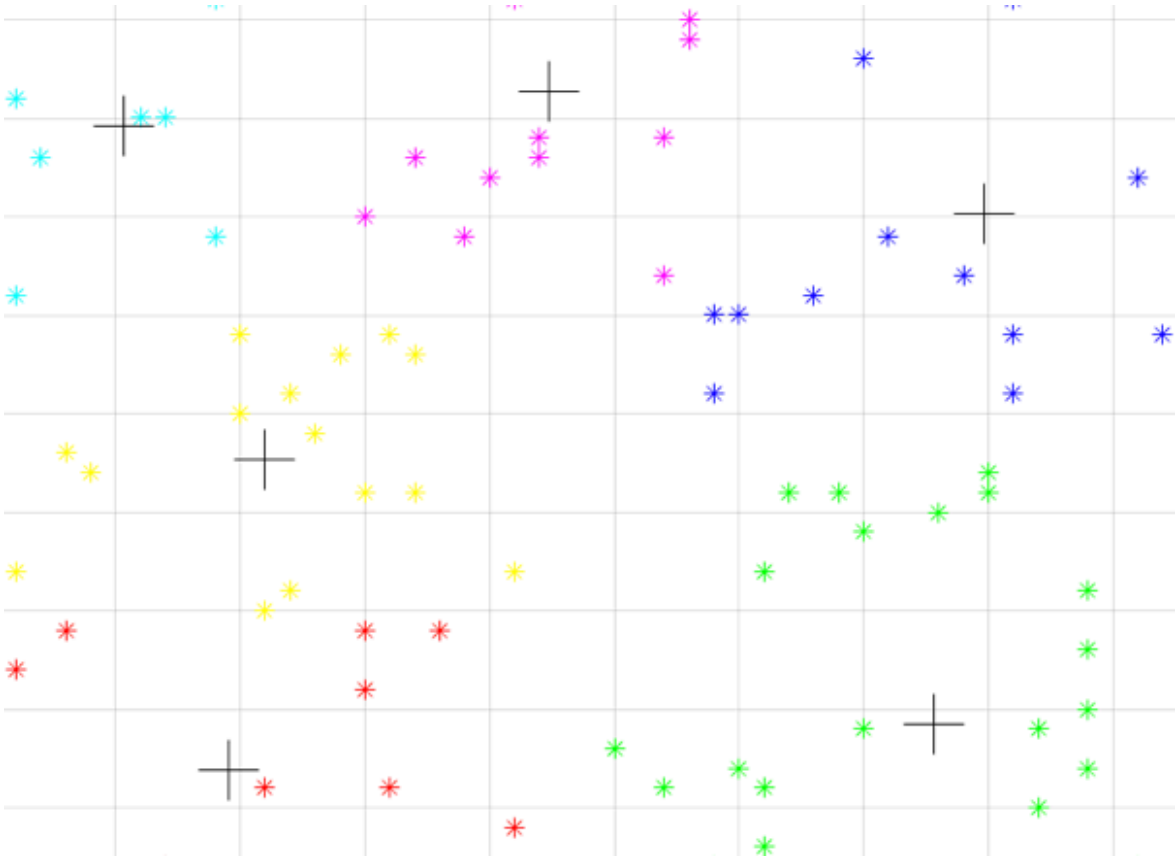
   map阶段计算所有点到中心的欧式距离；reduce阶段找到最小距离并输出其分类

   进行了k=4迭代次数=2；k=4迭代次数=10；k=6迭代次数=10三次实验，输出文件位于output文件夹中
3. Prodeal.java将步骤2输出结果处理成matlab可处理的数据形式
4. 运行kmeans.m将数据集变为可视化散点图

结果说明

1. 输出文件中前面代表改点坐标，后面代表改点所属的集群

   ```
   1,31    1
   44,7    4
   31,6    4
   28,44   2
   22,39   2
   35,19   4
   ```
2. 可视化散点图中不同颜色表示不同集群，黑色十字为集群中心



ps

全部文件上传至github，地址为https://github.com/BubblyDong/homework5