**homework2**

**1 Data process**

1. 为了得到每个基站每个小时的总流量，对时间（秒）除以3600，得时间（小时），then map(key is location ID and time(h)) and reducebykey(add traffic volumns) get sum of traffic volumns for each BS per hour.(P.S. 校正了homework1的代码，重新计算)
2. 为了将一个月的数据压缩为一周，对时间（小时）除以(7\*24) 取余数，将1）的结果map(key is location ID and time(week)) and reducebykey(average of traffic volumns)。
3. 标准化，（traffic volumns (每小时)-平均值）/标准差

~~结果：有缺失值？有些基站某些时间点没有数据~~

输出：hw2\_sum\_traffic\_volumns\_per\_hour\_of\_allBSs\_foldtoWeek\_norm.txt

**2 cluster : kmeans**

定义类的数量k，迭代次数x

初始化k个类的中心

每次迭代：

计算每个点到每个聚类中心的距离

该点属于距离该点最近的聚类中心所在类

每个类：

计算每个类新的中心

如果 新旧两个类的中心距离小于阈值，

则退出迭代

**3 result and virtualization （k=20）**

|  |  |  |
| --- | --- | --- |
| centroid | radius | center\_point |
| 0 | 4.738381 | 19 |
| 1 | 6.056489 | 2 |
| 2 | 16.11395 | 2 |
| 3 | 7.195818 | 25 |
| 4 | 4.80702 | 41 |
| 5 | 16.09102 | 36 |
| 6 | 0.393806 | 261 |
| 7 | 20.19489 | 5 |
| 8 | 1.601039 | 130 |
| 9 | 8.542803 | 4 |
| 10 | 7.871914 | 2 |
| 11 | 9.61304 | 90 |
| 12 | 4.142821 | 39 |
| 13 | 5.387921 | 43 |
| 14 | 13.99257 | 59 |
| 15 | 8.158211 | 72 |
| 16 | 6.598582 | 64 |
| 17 | 2.566839 | 34 |
| 18 | 10.18753 | 20 |
| 19 | 3.710234 | 13 |



