

Parallelize K Means

Jennifer Parnell



September 3, 2018

1. Pseudocode:

Select and clone 4 k clusters randomly

Integer MAXITER = data count

Integer datMatrix = [k][MAXITER]

Integer NumIter = 0

Boolean Changed = TRUE

While (NumIter<MAXITER) AND (Changed = TRUE)

For each data point

CalculateDistance(i, j)

Assign j to closest Mini

Update clusters(i,j)

CalculateDistance(Acluster, Apoint)

d(i, j) = Minimum(d(i - 1, j) + 1, d(i, j - 1) + 1, d(i - 1, j - 1))

Integer Minimum (int x, int y, int z)

Integer mini

Mini = x

If y < Mini

Mini = y

If z < Mini

Mini = z

Return Mini

UpdateCluster(Cluster, PointsBelongingToCluster, Changed)

Integer n

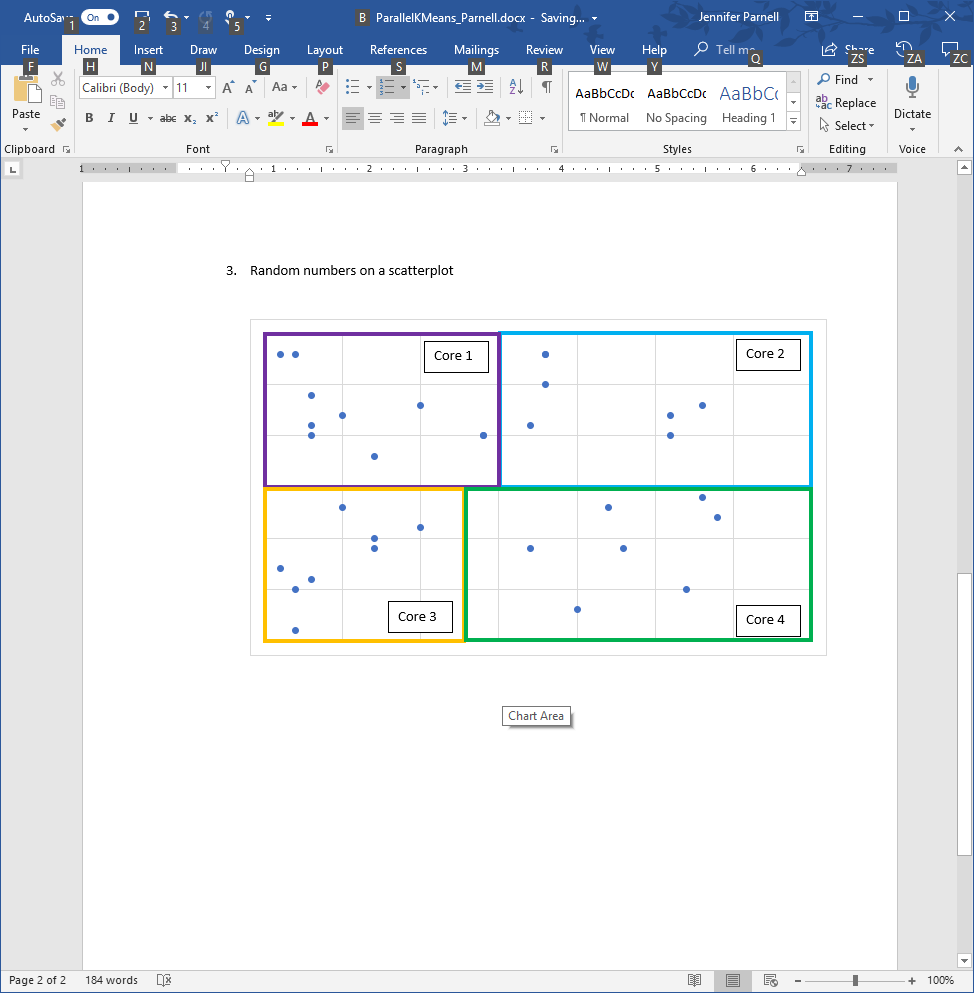
For (j=0;j<p; j++)

FeatureTotal += (i,j)

n++

Clusterj = FeatureTotal/n

1. The parallelization of integer data using a matrix in C. Four randomly selected data items are stored in the matrix (k) then each data integer is compared to the k items. Whichever k the item is closest to it will be added to that section of the matrix. This continues until all the data has been sorted.
2. Random numbers on a scatterplot



Core 3

Core 4

Core 2

Core 1

Data is broken down into 4 random groups which then can use each of its 4 threads to further sort the data into small groups. All four cores can run this at the same time on their threads.