K-means remains one of the most popular algorithm in clustering methods over the past decades. The advantage of k-means is its simplicity in implementing. however, this algorithm is not good enough in terms of clustering quality and efficiency. So recent work has been focused on finding better representation of the data and better initialization procedure. We choose to corporate spectral clustering with parallel version of k-means++ initialization method in to k-means algorithm so as to achieve better clustering.

# Algorithm Outline:

In the following description, assume a set of data points X = {x1, x2, …, xn} in the d dimensional space, and our objective is to divide them into k different clusters.

## Spectral Clustering:

1. Calculate the affinity matrix A , where ,  is a scaling parameter which can be manually assigned or selected during clustering procedure.
2. Define a diagonal matrix D, whose i-th diagonal entry is the sum of A’s i-th row, and define .
3. Find the k largest eigenvectors of L, , and form a matrix Z by horizontally stacking them in columns.
4. Form matrix Y by normalizing each row of X so that each row in Y has unit norm, i.e. 
5. Treat each row of Y as a new data point, and cluster them into k clusters via k-means.
6. Assign the original point  to the same cluster as i-th row of Y.

## Parallel k-means++ algorithm:

Define a cost function of X with respect to the set of centers C as: 

1. First sample a single point uniformly at random from data set Y, add it to the center set C
2. Calculate  and denote it as 
3. Then update C iteratively. In each iteration, sample each points in Y independently, with inclusion probability of  , where l is a oversampling factor chosen manually, . Then update the center set by adding all these newly chosen points. Repeat this for  times.
4. For each point x in the center set C, assign a weight  that equals the number of points closer to x than any other points in C.
5. Finally, recluster the weighted points in C into k clusters.