CS677 final proposal

What is the computation and why is it important?

k-means clustering is rather easy to apply to even large data sets, particularly when using heuristics such as Lloyd's algorithm. It has been successfully used in market segmentation, computer vision.

Abstraction of computation:

1. Calculate the Euclidean distance of two points to set clusters.
2. In each cluster, calculate the average distance to its center, and calculate the total distance of all cluster(add).

$$J_{SSE} = \sum_{i=1}^{k} \sum_{x \in D_i} || \boldsymbol{x} - \mu_i ||^2$$

Suitability for GPU acceleration:

1. Amdahl’s Law: Each points distance is independent, and permanent , it is the core calculation in the k-means algorithm. we can use parallel computing.
   a. Synchronization and Communication:
2. Data structure: struct Cluster{float *cluster_x, float *cluster_y} to calculate distance, it may used the const memory, use int *cluster_info to update the Tuple, it may use the shared memory. The distance calculation and the cluster_info update should be synchronize.

Difficulties:

Copy Overhead: If the data size is large, copy the data from CPU to GPU may take a long time.

How to update the resets on GPU, the use of shared memory and const memory.