Data Mining: Learning from Large Data Sets - Fall Semester 2015

mmarti@student.ethz.ch
trubeli@student.ethz.ch

December 2, 2015

Extracting Representative Elements

For this project we used a parallel version of k-means on our dataset. In a first step we designed an algorithm that could run on an arbitrary number of map processes and that could be combined in one reduce process.

Firstly, we choose k center randomly. Each map process gets a seed for the random number generator in order to make sure that the same centers are picked in each map process. The map process reads on the standard input and assigns each data point to one of the k_i centers. The reduce process sums all the data points assigned to each center k_i and compute the mean of this subset. This gives a new center which is used for these data points.