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

melisr@student.ethz.ch
eholmer@student.ethz.ch
pmichal@student.ethz.ch

December 2, 2015

Extracting Representative Elements

We use k-means approach in both mappers and reducer to extract representative elements.

Mapper

Each mapper uses **MiniBatchKMeans** implementation, found in `sklearn.cluster.MiniBatchKMeans`, to extract representatives from batches of input data. Batch size used was 750 data points, which we empirically found to be the biggest batch size not exceeding the servers memory limit. The algorithm uses **kmeans++** for initialization and this is repeated 20 times to find the best one, as measured by the kmeans objective function, before proceeding with the optimisation algorithm. Each mapper returns 750 centroids.

Reducer

Reducer receives centroids from all the mappers, and runs **MiniBatchKMeans** once more, again with batch size 750, but this time returning only 100 centroids.

Member contributions

Rik Meils: Experimented to find optimal parameters of the **MiniBatchKMeans** function

Erik Holmer: Explored alternative solutions

Michal Porvaznik: Implemented first version of the algorithm