Mining massive Datasets WS 2017/18

Problem Set 2

Rudolf Chrispens, Marvin, Daniela Schacherer

November 5, 2017

Exercise 01

See hand written solution.

Exercise 02

In order to also maintain information about which points are in which cluster one could store each Point p as
(*vector sum of points in cluster, number of points in cluster, list of data points in the cluster*).

Pseudocode for step 3 and 4 of subroutine *merge* from lecture 3:

- Find best pair, merge those two points/clusters and compute new cluster center

  ```
  d, (p,q), (ip, iq) = bestPair
  sum = p[0] + q[0]
  count = p[1] + q[1]
  points = p[2].append(q[2])
  newCenter = (sum, count, points)
  ```

- Filter p and q from the inCluster
- Re-number centroid index in inCluster
- Add new cluster to the outCluster