Mining massive Datasets WS 2017/18

Problem Set 2

Rudolf Chrispens, Marvin, Daniela Schacherer

November 10, 2017

Exercise 02

1. MapReduce Pseudocode for k-means
   The k-means algorithm consists of two steps. The first step computes for each mean μ_i the set of points that are closest to it. In the second step new means are computed using the priorily determined sets. These two phases correspond to the Map- and the Reduce phase of MapReduce. The Map phase computes the squared distance to all means for each point x in the dataset and returns a key-value pair $(i,(x,1))$ where i is the index of the mean. The Reduce phase then simply computes the sum of the vector points for each key.
   Pseudocode:
   - **Map** for every point x: return $(argmin_i(||x-\mu_i||,(x,1))$
   - **Reduce** for every elements with key i: return $(i,(x+y,s+t))$ with x and y being the data points and s and t being the counts.

2. MapReduce Pseudocode for Inverted Indexing
   - **Map**: for every keyword in the given list the Mapper should perform the following:
     if $keyword$ in $text_i$: return $(keyword, doc_i)$
   - **Reduce**: for every keyword add the documents indices to a list and finally return $(keyword, [doc_i, doc_j, ..$

3. MapReduce Pseudocode
   When one dataset is small and every mapper has access to it the joining can already be part of the mapping phase. Let R with tuples (a,b) and S with tuples $b,c)$ be the datasets, and R is the smaller one. We want to join on b. Every mapper gets tuples from S in this form: (S,a,b)
   - **Map**: for every tuple of R:
     if b in (S,a,b): return (a,b,c)
   - **Reduce**: in the reduce phase we now only have to collect all the joined tuples.