Technische Universität Berlin
Database Systems and Information Management Group
Prof. Dr. Volker Markl, Sebastian Schelter, Christoph Boden, Thomas Bodner

AIM3: Scalable Data Mining and Data Analysis
Summer term 2012

Third Assignment

Linear Algebra on Parallel Processing Platforms

Due on May 30th

Search as Linear Algebra

1. **Search as matrix vector multiplication**

   We have seen that search can be expressed as multiplication of the corpus matrix with a query vector. In *de.tuberlin.dima.aim3.assignment3.SearchAsMatrixVectorMultiplication* you have to implement the vectorization of the corpus, which is supplied in a textfile holding the following data per line:

   documentID;terms.

   A query file is supplied that needs to be mapped to the vector for the multiplication. The result of the job is a vector whose entries denote the number of matched terms for each document.

   Hint: Use *de.tuberlin.dima.aim3.assignment3.Dictionary* to map terms to matrix dimensions.

2. **Inverting an index as matrix transposition**

   When searching documents, one usually builds a so called *Inverted Index*, a data structure that for each term holds the documents containing it. Conceptually the inverted index is equivalent to the transposed corpus matrix, which points from terms to documents. In *de.tuberlin.dima.aim3.assignment3.MatrixTransposition* you have to implement matrix transposition for this task.

Deadline

Source code for the exercises is available at *https://github.com/dimalabs/scalable-dataming-class*.

Upload your solution to ISIS in the form of a patch file until noon of May 30th.