L3S Research Center
Large Scale Data Mining, SS 2016
Dr. Avishek Anand
Assignment 1, due: **19 April 2016**

You must return your assignment sheet and have a correct solution in order to present in the exercise groups. Please write legibily! Do no forget to put your name and matriculation number on your solution!

Problem 1.
Recall the Word Count example from the lecture.

1. How do we need to change it to count the number of sentences a word occurres in, instead of the total number of occurrences?
2. You now get a second list of already computed word occurrences as input in the same format: (word, count). Design a MapReduce algorithm that combines these two lists by computing the union, with the counts for the same word summed up.
3. How can you compute the intersection and only keep words that occur in both lists with the counts summed up? Design a MapReduce job for this task.

(Please describe the corresponding map and reduce steps or modify the code accordingly.)

Problem 2. How many and which MapReduce jobs are required for the Pig example from the lecture? (Please describe the corresponding map and reduce (and shuffle) steps.)

Problem 3. Design MapReduce algorithms to take a very large file of integers and produce as output -

1. The largest integer.
2. The average of all integers.
3. The same set of integers, but with each integer appearing only once.
4. The count of the number of distinct integers in the input.

Hint: Make use of combiners and sorting in the shuffle phase. (Please describe the corresponding map and reduce steps.)

Problem 4. Can the matrix multiplication algorithm shown in the lecture be achieved with only one MapReduce job? (Please describe the corresponding map and reduce steps.)