L3S Research Center
Large Scale Data Mining, SS 2016
Dr. Avishek Anand
Assignment 2, due: **29 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. Suppose we have a universal set U of n elements, and we choose two subsets S and T at random, each with m of the n elements. What is the expected value of the Jaccard similarity of S and T ?

Problem 2.

1. If we use the stop-word-based shingles, and we take the stop words to be all the words of three or fewer letters, then what are the shingles in the sentence **'Even if we hash them to four bytes each, the space needed to store a set is still roughly four times the space taken by the document.'**
2. What is the largest number of $k-$shingles a document of n bytes can have? You may assume that the size of the alphabet is large enough that the number of possible strings of length k is at least as n.

Problem 3. Suppose we want to use a MapReduce framework to compute minhash signatures. If the matrix is stored in chunks that correspond to some columns, then it is quite easy to exploit parallelism. Each Map task gets some of the columns and all the hash functions, and computes the minhash signatures of its given columns. However, suppose the matrix were chunked by rows, so that a Map task is given the hash functions and a set of rows to work on. Design Map and Reduce functions to exploit MapReduce with data in this form.

Problem 4. Approximate the S-curve $1 - (1 - s^r)^b$ when s^r is very small.
Hint: $(1 + a)^b = e^{ab}$, where a is very small.

Problem 5. For the (r, b) pair:

1. $r = 6$ and $b = 20$
2. $r = 3$ and $b = 10$
3. $r = 35$ and $b = 60$

compute the threshold, that is, the value of s for which the value of $1 - (1 - s^r)^b$ is exactly $1/2$. How does this value compare with the estimate of $(1/b)^{1/r}$ that was suggested in the lecture?

Problem 6. On the space of nonnegative integers, which of the following functions are distance measures? If so, prove it; if not, prove that it fails to satisfy one or more of the axioms.

1. max(x,y) = the larger of x and y.
2. diff(x,y) = $|x - y|$ (the absolute magnitude of the difference between x and y).
3. sum(x,y) =$x + y$.
4. Jaccard distance
5. shortest path between a pair of nodes in an weighted (weights are non-negative) undirected graph.