Aufgabe 6-1  Efficient cosine similarity for parallel systems

The cosine similarity function is commonly defined as:

\[ \cos(\varphi) := \frac{x \cdot y}{||x|| \cdot ||y||} \]

The angle \( \varphi \) can be used as a pseudo distance function.

Of particular importance is this distance function for text data, which are usually high-dimensional and sparse. If the data vector has been normalized in a previous step (i.e. \( ||v|| = 1 \)), this formula becomes:

\[ \cos_{\text{norm}}(\varphi) = x \cdot y = \sum_{i=0}^{n} x_i y_i \]

(a) What is the complexity of this distance function, if vectors \( x \) and \( y \) are both sparse and very high dimensional, particularly compared with the Euclidean distance?

(b) Assuming only \( x \) is sparse, but \( y \) (e.g. a centroid) is dense. How does this affect the computational complexity?

(c) To calculate pairwise similarity in a large database, we transpose the vectors and process them iteratively (e.g. using Hadoop). What is the advantage of this approach?

(d) A similar trick can be applied to Euclidean distances applied to sparse vectors. To achieve this the second binomial theorem can be used: \( (a - b)^2 = a^2 - 2 \cdot a \cdot b + b^2 \). Describe how this formula can be applied here.
Aufgabe 6-2 Privacy Preservation in Standard Classifiers

Given the following classifiers: decision trees, nearest neighbor classification, support-vector-machines, and naive bayes.

- Discuss whether pre-trained classifiers can be distributed to third parties without giving access to parts of the training set.
- How could encountered problems be solved?

Aufgabe 6-3 Parallele Association Rules

Discuss the advantages and disadvantages of horizontal and vertical distributions in the parallel generation of association rules.

Aufgabe 6-4 Parallel Naive Bayes Classification with Map Reduce

Describe a program which calculates all required probabilities for a Naive Bayes classifier using MapReduce. Assume that each class can be modeled by a multivariate axis-parallel normal distribution and that the training set $D$ is given as tuples $< ID, object >$ with $object$ having attributes $c$ and $v$. Let $ID$ be a key for each object, $c \in C$ be the class, and $v \in \mathbb{R}^d$ be a feature vector.

Specify a function for the mapper and a function for the reducer in pseudo-code.