University of Augsburg, Institute of Computer Science SS 2019
Prof. Dr. P. Fischer 21. June 2019
J. Kastner, L. Rudenko Sheet 8

Analyzing Massive Data Sets

Exercise 1: BFR - variance (homework)

Last week we have dealt with the BFR-algorithmus (Sheet 7, Exercise 4). The clusters in this approach are stored as an array with the numerical values. One of these values is the **variance** σ^2. For the calculation of the variance in the dimension i we have used the following formula: $SUMSQ_i/N - (SUM_i/N)^2$. Prove that this formula is correct.

Exercise 2: Boolean Retrieval (homework)

The following query Q is given:

$$Q : Bread\ AND\ NOT\ (game\ AND\ work)$$

Furthermore, the following corpus K of documents is given:

- $D_1 = \{$bread, celebration, game$\}$
- $D_2 = \{$game, work$\}$
- $D_3 = \{$bread, gladiator$\}$

Turn the query Q to the **disjunctive normal form (DNF)** and evaluate each single conjunction of the DNF using the **boolean model** on the corpus K. Specify the result of query Q.

Note that negations in the disjunctive normal form are only allowed for single literals!

Exercise 3: Boolean Retrieval (live)

Consider the following **documents** D_1, D_2, D_3, D_4, D_5.

- $D_1 = \{$cat, pet, dog$\}$
- $D_2 = \{$cat, bird, duck$\}$
- $D_3 = \{$duck, chicken, bird$\}$
- $D_4 = \{$tiger, cat, lion$\}$
- $D_5 = \{$duck, chicken, bird$\}$

a) Is it possible to specify a boolean query for each document that returns **exactly this one document**? Under what conditions does this work?

b) **Evaluate** the following **queries** using the **boolean retrieval model** and **indicate** the **relevant documents**:

   i) $Q_1 =$ 'cat' AND 'chicken'

   ii) $Q_2 =$ 'tiger' AND 'lion'

   iii) $Q_3 =$ 'bird' OR 'duck'

   iv) $Q_4 =$ 'cat' AND NOT 'bird'

c) **Evaluate** following **queries** using the **boolean retrieval model** and **indicate** the **relevant documents**. First of all **transfer** following queries into the **Disjunctive Normal Form (DNF)** and the **Conjunctive Normal Form (CNF)**.

   i) $Q_5 =$ 'cat' AND (('lion' AND 'duck') OR 'bird')

   ii) $Q_6 =$ (('pet' AND 'cat') OR ('cat' AND 'duck')) AND ('cat' OR 'bird')

Exercise 4: Fuzzy IR-model (live)

Look again at the documents D_1, D_2, D_3, D_4 from Exercise 3.

a) Determine the **Jaccard indices** for the terms in the documents $D_j, j = 1, ..., 4$ to get a notion of **term similarity**.

b) Compute **fuzzy degree of membership** $W(D_j, t_i)$ for each term t_i in each document D_j.

c) Compute the result of the queries $Q_i, i = 1, ..., 6$ from Exercise 3 in the fuzzy model.