Trier University
Department IV – Computer Sciences
Prof. Dr.-Ing. Ralf Schenkel (H 539, schenkel@uni-trier.de)

Trier, April 21, 2021

Assignment 2 of the course Big Data Analytics

Summer 2021

Deadline 10:15h on Wednesday, April 28, 2021

Task 1: Evaluation of Blocking Strategies 45 points

In this task, you are to evaluate blocking strategies for mentions of persons, based on their names. With this excessive you can find a CSV file that contains pairs of person names that are known to refer to the same person entity. Each line contains two mentions that should eventually end up in the same profile.

Consider the following blocking strategies:

- s_1 Use the first (given) name of a person to determine the blocks.
- s_2 Use the last (surname) name of a person to determine the blocks.

Process all mentions in the CSV file (both `OLD_NAME` and `NEW_NAME`. For each name determine given and surname for each mention.

- The given name is the first name part before the first space. E.g. Peter M. Jones, ChenLi Wang, Sammy, Karl-Heinz Schmidt.
- The surname name is the part after the last space. E.g. Peter M. Jones, ChenLi Wang, Sammy, Karl-Heinz Schmidt.
- Ignore all other name parts
- Ignore all cultural particularities that affect what is a given name and surname.

Compute the blocks for each strategy (in a programming language of your choice, I suggest Java or Python). E.g., for s_1 all names with the same given name end up in the same block.

Evaluate the strategies with the following evaluation measures:

Let P be the set of all name pairs from the CSV file. Let $P_O \subseteq P$ the set of these pairs that end up in the same block.

$$Rec := \frac{|P_O|}{|P|}$$

defines a basic measure on recall (i.e., how may of the pairs are grouped together)

It is also important to determine how many similarity computations are saved by using the blocking. Let C be the sum of the number of size two subsets in all blocks (the actual number of similarities that need to be computed after the blocking. With n_b the number of mentions in a block and B the set of all blocks

$$C := \sum_{b \in B} \frac{n_b * (n_b - 1)}{2}$$

then with n the number of all mentions

$$Save := 1 - \left(\frac{|C|}{\left(\frac{n*(n-1)}{2} \right)} \right)$$

describes the saving if this blocking is used.

- Compute *Rec* and *Save* for both strategies.
- Compare those values and briefly comment on the differences that you observe.
- Name at least two properties of the underlying name base that might affect the usefulness of the strategies outlines above.

Hand in: The program code to compute the results; the results and comments as described above.

These tasks will be discussed in the tutorial on April 28, 2021.