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

Trier, May 8, 2024

Assignment 2 of the course Big Data Analytics

Summer semester 2024

Deadline 10:15h on May 15, 2024

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. The provided CSV file 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 many 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 May 27, 2024.

General remarks:

- The tutorial group takes place on Mondays at 14:15 in F55 on a (roughly) bi-weekly basis.
- The first meeting of the tutorial group is on May 27, 2024.
- To be admitted to the final exam, you need to acquire at least 50% of the points in the assignments.
- It is required to submit in groups of size 3; only one submission is sufficient for the whole group. Groups must be chosen in Moodle (see link on the course page in Moodle). Write the names of all group members on your solutions. Students without a group cannot submit.
- Solutions must be handed in before the deadline in Moodle (`https://moodle.uni-trier.de/`, course `BDA-24`) as as a PDF or, if submitting multiple files, as an archive (.zip or comparable). Submissions that arrive after the deadline will not be considered.
- Graded versions of your submissions will be returned in Moodle until the following tutorial.
- Announcements regarding the lecture and the tutorial group will be done in the area of the lecture in StudIP.