Big Data Analytics SS 2020
Übungsblatt 2

Aaron Winziers - 1176638

9. Mai 2020

Task 01

20 lines in the file contained improperly formatted data. These were ignored and not used in the calculations. Below is a table of the calculated values.

	First Name	Last Name
Rec	0.3848	0.7632
Save	0.9970	0.9996
C	12,083,443	1,561,279
Number of Blocks	18,168	29,830

Rec The two blocking strategies differ greatly in their recall values. With a recall of 0.7632, blocking by last name was almost twice as successful as blocking by first name in placing the name pairs into the same block. This is likely because first names are abbreviated much more often than last names.

Another aspect that affected both groups considerably was the swapping of names, meaning that the order of first and last name was incorrect.

Save Both strategies allowed for a large percentage of calculations to be saved; through first name blocking 99.7% and through last name blocking 99.96%. This still means that roughly 10.5 million more calculations must be performed when blocking by first name.

This difference can be explained through the previously mentioned abbreviation of first names. Each time a name is abbreviated, it is placed into a block that contains all other instances of a name starting with a particular letter that are also abbreviated. As a result, 26 blocks (assuming characters are limited to just the English alphabet) are created that can quickly become quite large. The number of calculations that need to be performed grow quadratically with the size of the block, meaning that these blocks have a strong negative impact on the performance gain brought by the blocking.