Report on the Data Linkage project

Data linkage is bringing information from different sources together about a product or an entity to create a linked dataset. Here the two datasets abt_small.csv and buy_small.csv hold information about their products. An algorithm is developed in this part of the assignment to link the same products in their databases based on mutual information.

In part 1a, the data linkage is performed without blocking. Both the datasets contain the name of the product and its description. Only a few of the products in the dataset buy_small.csv have some description linked to it. Hence an algorithm is set up which checks for the existence of the product description and concatenates the product name and description only if the description exists . If a product is found without any description, it only saves the product name. This is done for all the products in the dataset. These concatenated strings act as a pool of choices used during linking the ids from both the databases. Since every product in the abt_small.csv dataset is linked to a description, a concatenated string of product name and description is produced. Each of these strings which represent a product from the abt_small.csv dataset is then compared to the pool of strings representing the products from the buy_small.csv dataset. This process is done using the **fuzzywuzzy** library's **'fuzz.token_set_ratio'** which considers duplicate words as a single word while comparing the strings. The match with the highest score is selected and checked for a threshold >= 70. This threshold indicates a substantial match between the strings. If the comparison score crosses this threshold, the corresponding ids of the strings are matched and saved in a data frame like this –

	idABT	idBuy
1	580	202812620
2	6726	203111433
3	9546	208455792

The recall and precision values come out to be 0.67 and 0.66 repectively which indiciate that the overall performance can be substantially improved. This can be done using more rigorous string matching.

In part 1b, blocking method is implemenetd to link the buy.csv and abt.csv datasets. Blocking is a process in which the dataset is divided into blocks in which the comaparions are carried out to link the enitites. For the given datasets, intitally the blocking is implemented on the buy.csv dataset where the products are blocked by their names and then by their manufacturer. This gives a more precise segregation of the data into blocks. Since the abt.csv dataset does not contain the manufacturer name, the names are found using tokenizing the product name by selecting the first word as majority of the product names start with the manufacturer name. Then the blocks identified using the first dataset are applied to the abt.csv dataset hence implementing the blocking method. Two csv files are produced (abt_blocks.csv and buy_blocks.csv) which contain the product ids with their repective blocks as follows-

	block_key	product_id
1	Sony	552
2	BOSE	580
3	Sony	4696

The PC and RR turn out to be 0.94 and 0.946 which indicate a good implementation of the blocking method.