Matching Algorithm

For this matching algorithm in assignment 3, I used Jaro-Winkler distance to find the similarity between book titles in Table A and Table B. The Jaro-Winkler distance is a string metric that measures the edit distance between two strings, the values ranging from 0 to 1. A value of 1 indicates the two strings are identical, while a value of 0 signifies no similarity. The Jaro similarity is calculated using the following formula:

$$Jaro\ similarity = \frac{1}{3}\left(\frac{m}{|s1|} + \frac{m}{|s2|} + \frac{m-t}{m}\right)$$

Where m is the number of matching characters, t is half of the number of transpositions, and |s1| and |s2| are the lengths of strings s1 and s2 respectively. For example, if s1 = “arnab” and s2 = “raanb”, the Jaro Similarity is 0.8667.

First, I read the CSV files tableA.csv and tableB.csv in my code. For each row in Table A, I calculated the Jaro-Winkler similarity between its title and every title in Table B. If the similarity score is greater than 0.90, I recorded the following information into a list called "matches": ID, Table A ID, Table B ID, Table A title, Table B title, Table A author, and Table A rating. Finally, after all iteration is completed, the marches array is written in a file called “tableC.csv.”

To determine an appropriate threshold for the similarity score, I performed multiple tests to minimize errors. I first set the threshold to 0.70 resulting in 3,654 rows in Table C. At this threshold, many book titles from table A were matched to multiple titles in Table B. For instance, "American Psycho" from Table A was matched to "American Psycho," "Mexican Gothic," and "Amnesia" from Table B, as all of these pairs had similarity scores exceeding 0.70. Ultimately, a threshold of 0.90 yielded 120 matches with significantly fewer errors.

Table A and Table B both contain the columns: ID, title, author, and rating. Each table has 1,000 rows. The total number of tuple pairs in the Cartesian product of A and B is 1,000 * 1,000 = 1,000,000. The final number of tuple pairs in Table C is 120.

Reference

https://www.youtube.com/watch?v=vmpoIIVZrgM

https://www.geeksforgeeks.org/jaro-and-jaro-winkler-similarity/