Entity Resolution

Instructions

1) Please solve the following problem using Python or R.

2) Please push the assignment onto a free, publicly available repository for review (e.g. GitHub or Gitlab). The repository should include:
   a) The final code used to solve the problem, preferably in one script
   b) A README.md file
   c) The output .csv file (see below for more details)

3) In the README.md file, please include:
   a) A description of the steps taken to resolve the entities
   b) Any assumptions made to solve the problem
   c) Libraries or packages used

4) You are free to use any open source library or package to complete the problem

Problem Description

"Entity resolution" is the problem of identifying which records in a database represent the same entity. When dealing with user data, it is often difficult to control the quality of the data inputted into the system. The poor quality of the data may be characterized by:

- Duplicated records
- Records that link to the same entity across different data sources
- Data fields with more than one possible representation (e.g. "P&G" and "Procter and Gamble")

In this assignment, you are provided with two datasets:

1. Scholar.csv
2. DBLP.csv

Each dataset contains the following columns:

Id[.csvName]	title	author	venue	year	ROW_ID

There are records that reference the same entity across the two datasets.

Bringing Everything Together

Your assignment is to resolve the records to their respective entities, and write a final .csv named "**DBLP_Scholar_perfectMapping_[YourName].csv**" that only contain the resolved entities. The final .csv file should include the following column headings:

idDBLP: The matched DBLP.csv id
idScholar: The matched Scholar.csv id
DBLP_Match: The ROW_ID of the DBLP.csv file
Scholar_Match: The ROW_ID of the Scholar.csv file
Match_ID: A final column that combines the number from DBLP_Match and Scholar_Match, separated by an underscore.

The first row is provided as an example:

idDBLP	idScholar	DBLP_Match	Scholar_Match	Match_ID
conf/vldb/Levy96	lDTPyBMtHVwJ	1996	14	1996_14

Emphasis will be placed on the method and logic used in resolving the entities, rather than on the final result.