Census Record Linkage with Household Structures

This document provides general information on using the code (general_code.R) to perform census record linkage with household structures. This algorithm takes in individual links and calculates the household similarity of each link to determine the true match given the presence of multiple ambiguous links. The household similarity is also used to find links that are missed by the individual linkage.

Preparing the data:
The goal is to link the "left" dataset to the "right" dataset by processing links from individual linkage performed by any individual linkage algorithm or built-in packages such as dtalink and RecordLinkage.
The linkage requires the following identification variables for both "left" and "right" dataset:

- personID
- householdID

The linkage is based on the following variables:

- First name (character)
- Last name (character)
- Birth year (numeric)

Records with missing values in these three variables will be dropped.
The linkage also allows one or more user-defined blocking variables such as gender and location.

User should provide the following files and directories:

1. *individual_links*: a *.csv* file that contain the links from individual linkage. The file should have three columns. The first column should be named ***"similscore"*** and contain individual similarity of values between 0 and 1. The second column should contain the personID of the "left" dataset and the third column should contain the personID of the "right" dataset.
2. *left_file*: a *.csv* file that contains the following columns with the exact names:
    - ***"personID"***
    - ***"householdID"***
    - ***"namefrst"***
    - ***"namelast"***
    - ***"birthyr"***

The file can also include other columns such as blocking variables.

3. *right_file*: a *.csv* file with the same format as *left_file.*

Linkage parameters:

At the top of the "general_code.R" file, user should set the following parameters:

1. *output_directory*: the desired directory of the output file.
2. *leftname*: the name of the "left" dataset, can be different from the *.csv* file name.
3. *rightname*: the name of the "right" dataset, can be different from the *.csv* file name.
4. *blocking*: names of the blocking variables, should be consistent with the column names in the input files.
5. *firstname_cutoff*: a numeric value between 0 and 1 that specifies the string similarity threshold for first names.
6. *lastname_cutoff*: a numeric value between 0 and 1 that specifies the string similarity threshold for last names.
7. *age_cutoff*: an integer value that specifies the maximum accepted age difference.
8. *requirements*: an integer value between 1 and 3 that specifies the number of rules among 5-7 that a pair needs to satisfy to be matched.
9. *percent_match*: a numeric value between 0 and 1 that specifies the household similarity level required for a pair to be matched.
10. *lowest_similarity*: a numeric value between 0 and 1 that specifies the minimum required individual *"similscore"*.