Fuzzy Matching R Shiny Application

The standalone fuzzy matching application is written in the R language using the R Shiny package. It consists of a rather large folder because it uses R portable (and chrome portable) so that you can run it on machines without R installed.

If you do have R installed, the necessary R Shiny files are in the **shiny** subfolder of the application folder; you won't need the rest of the files in the application folder. I expect that if you have R installed, you will be familiar with how to use the R files, so I have not included additional instructions for doing so.

Running the application

1. Move the zip file to your local machine and unzip it.

2. Open the folder and run the file "run.bat".

3. Wait for chrome portable to load the app.

4. Upload a BUS file using the "Browse" button. This must have all of the BUS columns, named properly, and must be formatted as a .xlsx file.

5. Select the margin of error for differing full names, defaulting to two characters (this is used in match type 4, as described below).

6. Click "Load file".

7. Download the data.

Input

The input file to the R Shiny application must follow the BUS format and be saved as a .xlsx file. The BUS format includes, at minimum the following four column names *exactly* (including case):

first name

middle name

last name

street address

Output

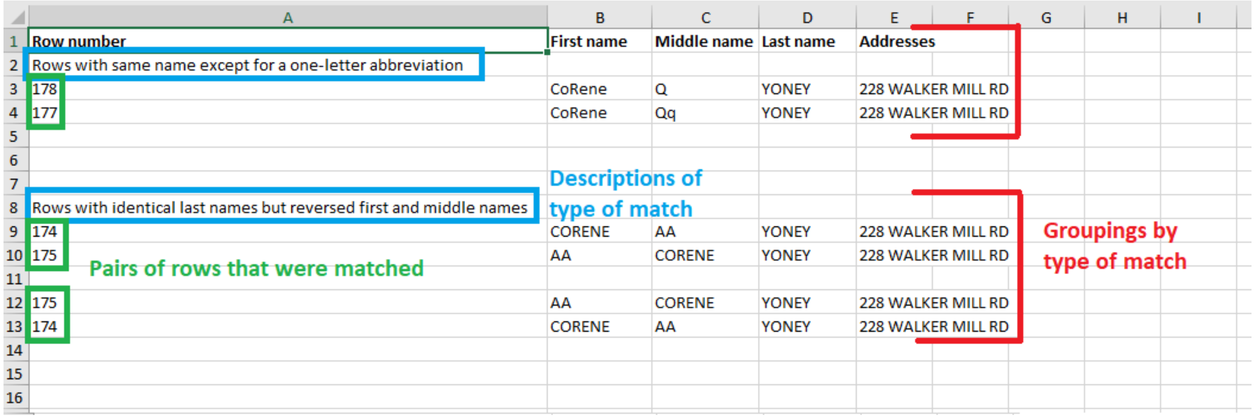
The downloaded file consists of an xlsx file with two sheets:

1. The “data” sheet contains the table from the original BUS file, with one additional column for each type of fuzzy match that is being checked.

- The name of the column describes the type of match.

- Each non-empty value in the column is a list of other rows that were matched to the row in question.

2. The “matches” sheet contains a list of matched pairs of rows. They are organized as follows:



	A	B	C	D	E
1	Row number	First name	Middle name	Last name	Addresses
2	Rows with same name except for a one-letter abbreviation				
3	178	CoRene	Q	YONEY	228 WALKER MILL RD
4	177	CoRene	Qq	YONEY	228 WALKER MILL RD
5					
6					
7					
8	Rows with identical last names but reversed first and middle names				
9	174	CORENE	AA	YONEY	228 WALKER MILL RD
10	175	AA	CORENE	YONEY	228 WALKER MILL RD
11					
12	175	AA	CORENE	YONEY	228 WALKER MILL RD
13	174	CORENE	AA	YONEY	228 WALKER MILL RD
14					
15					
16					

Descriptions of type of match

Pairs of rows that were matched

Groupings by type of match

Functionality

I am currently checking for the following types of matches for each pair of rows:

1. Rows A and B have identical first/middle/last names except that, for at least one of first/middle/last name,
   - Row A’s value consists of only a single letter.
   - Row B’s value consists of a full name (i.e. at least two letters long) that begins with row A’s value.
   Example: [Alice B Carol] and [Alice Bob Carol] are matched.
2. Rows A and B have identical values for one of first/middle/last name, but the values for their other two names are switched.
   Example: [Bob Alice Carol] and [Alice Bob Carol] are matched.
3. Rows A and B have identical first/middle/last names except that, for at least one of first/middle/last name,
   - They have the same value when both are converted to fully lower case, but otherwise do NOT have the same value. (caps rule)
   Example: [ALICE BOB CAROL] and [Alice Bob Carol] are matched.
   OR
   - Row A’s value contains a hyphen.
   - Row B has the same value as row A once the hyphen is removed. (hyphen rule)
   Example: [Alice Bob Carol] and [Al-ice Bob Carol] are matched.

4. Rows A and B have identical addresses, and their full names (concatenated first/middle/last names) are at most [margin of error] characters apart.
   - The distance calculation relies on the adist R function which uses edit distance, as described in the function documentation: https://stat.ethz.ch/R-manual/R-devel/library/utils/html/adist.html
   Example: if [error of margin] >= 2, then [Aliceaa Bob Carol] and [Alice Bob Carol] are matched.
5. Rows A and B have identical names for one or two of first/middle/last name, but for the remaining name column(s),
   - Row A’s value is blank.
   - Row B’s value is not blank.
   Example: [Alice Bob] and [Alice Bob Carol] are matched.