fastLink tutorial

Emanuel Ben-David

2023-06-03

What is fastLink?

- **fastLink** is an **R** package for Fast Probabilistic Record Linkage
- fastLink to some degree can handle missing data
- It is amenable to parallel computing, using the number of cores in the computer
- **splink** is essentially a translation of fastLink to python language, with spark capability, for **python users**.
- It is written by **Ted Enamorado** [aut, cre], Ben Fifield [aut], Kosuke Imai

Linking two files

We would like to link two data files: dfA and dfB. The main function for linking is fastlink().

fastLink(dfA, dfB, varnames, stringdist.match, stringdist.method, numeric.match, partial.match, cut.a, cut.p, . . .)

- "**varnames**": vector of matching variables.
- Must be present in both dfA and dfB
- "**stringdist.match**": vector of string variables in "varnames"
- "**stringdist.method**": default is jw" for Jaro-Winkler, other options are "jaro" for Jaro, and "lv" for edit.

Linking files with fastlink() continued

- **"numeric.match": numeric** variables for numeric matching
- **"partial.match":** string variables among "**stringdist.match" variables** for partial matching.
- "**cut.a**": lower bound for full string-distance match, ranging between 0 and 1. Default is 0.94
- "**cut.p**": Lower bound for partial string-distance match, ranging between 0 and 1. Default is 0.88
- "**n.cores**": number of cores to parallelize over. Default is NULL.
- We can use "getMatches()" function to get the matches.
- The arguments for "getMatches()" are:
- dfA: files A
- dfB : file B
- fl.out: the output of "fastlink()" in the setp above

Slide with R Output

- For demo, we uplaod file_a and file_b.

‘ Let’s see the variables in data files.

```
##  [1] "CredentialNumber" "LastName"          "FirstName"
##  [5] "CredentialType"   "Status"            "BirthYear"
##  [9] "FirstIssueDate"   "LastIssueDate"     "ExpirationDa
## [13] "LastInitial"
```

```
##  [1] "CredentialNumber" "LastName"          "FirstName"
##  [5] "CredentialType"   "Status"            "BirthYear"
##  [9] "FirstIssueDate"   "LastIssueDate"     "ExpirationDa
## [13] "LastInitial"
```

- We link these two data files by blocking on “BirthYear”.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```