Package ‘LinkOrgs’

June 22, 2021

Title LinkOrgs: Algorithms for Organizational Record Linkage

Version 0.0

Authors 'Brian Libgober <libgober@gmail.com > [aut,cre], Connor Jerzak <cjerzak@g.harvard.edu> [aut, cre]'

Description An R package for organizational records using the algorithms of Jerzak & Libgober (2021). The linkage is done based on organizational names and using half a billion open collaborated records on those names from LinkedIn users.

Depends R (>= 3.3.3)

License
Creative Commons Attribution-Noncommercial-No Derivative Works 4.0, for academic use only.

Encoding UTF-8

LazyData true

Maintainer 'Connor Jerzak' <connor.jerzak@gmail.com>

Imports data.table,plyr

RoxygenNote 7.1.1

R topics documented:

`FastFuzzyMatch` *FastFuzzyMatch*

Description

Performs fast fuzzy matching of strings based on the string distance measure specified in `control`.

Usage

```
FastFuzzyMatch(x,y,by,...)
```

Arguments

`x, y`	data frames to be merged
`by, by.x, by.y`	specifications of the columns used for merging. See `?base::merge` for more details regarding syntax.
`control`	A list specifying how to process the alias text. See “Details”.

Details

LinkIt can automatically process the alias text for each dataset. Users may specify the following options in the `control` list:

- Set `DistanceMeasure` to control algorithm for computing pairwise string distances. Options include `"osa"`, `"jaccard"`, `"jw"`. See `?stringdist::stringdist` for all options. (Default is `"jaccard"`)
- Set `FuzzyThreshold` to control the maximum allowed distance between two matched strings
- Set `qgram` to control the character-level q-grams used in the distance measure. (Default is `2`)
- Set `RemoveCommonWords` to TRUE to remove common words (those appearing in > 10% of aliases). (Default is `FALSE`)
- Set `NormalizeSpaces` to TRUE to remove hanging whitespaces. (Default is `TRUE`)
- Set `RemovePunctuation` to TRUE to remove punctuation. (Default is `TRUE`)
- Set `ToLower` to TRUE to ignore case. (Default is `TRUE`)
- Set ’PreprocessingFuzzyThreshold‘ to some number between 0 and 1 to specify the threshold for the pre-processing fuzzy matching step.

Value

z The merged data frame.

`LinkOrgs` *LinkOrgs*

Description

Implements the organizational record linkage algorithms of Jerzak and Libgober (2021).

Usage

```
LinkOrgs(x, y, by ...)
```

Arguments

`x, y`	data frames to be merged
`by, by.x, by.y`	character vector(s) that specify the column names used for merging data frames `x` and `y`. The merging variables should be organizational names. See `?base::merge` for more details regarding syntax.
`algorithm`	character; specifies which algorithm described in Jerzak and Libgober (2021) should be used. Options are `"markov"`, `"bipartite"`, and `"ml"`. Default is `"ml"`, which uses a machine learning approach to predicting the match probability.
`control`	A list specifying how to process the alias text and how to compute string distances. See “Details”.

Details

LinkOrgsautomatically processes the name text for each dataset (specified by by, by.x, and/or by.y. Users may specify the following options in the control list:

- Set DistanceMeasure to control algorithm for computing pairwise string distances. Options include "osa", "jaccard", "jw". See ?stringdist::stringdist for all options. (Default is "jaccard")
- Set FuzzyThreshold to control the maximum allowed distance between two matched strings
- Set qgram to control the character-level q-grams used in the distance measure. (Default is 2)
- Set RemoveCommonWords to TRUE to remove common words (those appearing in > 10% of aliases). (Default is FALSE)
- Set NormalizeSpaces to TRUE to remove hanging whitespaces. (Default is TRUE)
- Set RemovePunctuation to TRUE to remove punctuation. (Default is TRUE)
- Set ToLower to TRUE to ignore case. (Default is TRUE)

Value

z The merged data frame.

MatchPerformance *MatchPerformance*

Description

Automatically computes the true/false positive and true/false negative rates based on a ground-truth (preferably human-generated) matched dataset.

Usage

```
MatchPerformance(x,y,by,...)
```

Arguments

x, y	data frames to be merged
z	the merged data frame to be analyzed. Should contain by,by.x, and/or by.y as column names, depending on usage.
z_true	a reference data frame containing target/true matched dataset. Should contain by,by.x, and/or by.y as column names, depending on usage.
by, by.x, by.y	character strings specifying of the columns used for merging.

Value

ResultsMatrix A matrix containing the information on the true positive, false positive, true negative, and false negative rate, in addition to the matched dataset size. These quantities are calculated based off all possible nrow(x)*nrow(y) match pairs.

Examples

```
# Create synthetic data
x_orgnames <- c("apple","oracle","enron inc.","mcdonalds corporation")
y_orgnames <- c("apple corp","oracle inc","enron","mcdonalds co")
x <- data.frame("orgnames_x"=x_orgnames)
y <- data.frame("orgnames_y"=y_orgnames)
z <- data.frame("orgnames_x"=x_orgnames[1:2], "orgnames_y"=y_orgnames[1:2])
z_true <- data.frame("orgnames_x"=x_orgnames, "orgnames_y"=y_orgnames)

# Obtain match performance data
performanceMat <- MatchPerformance(x = x,
                                   y = y,
                                   z = z,
                                   z_true = z_true,
                                   by.x = "orgnames_x",
                                   by.y = "orgnames_y")
print( performanceMat )
```

Index