Package 'LinkIt'

June 22, 2021

Title An Algorithm for Dataset 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 estimating record linkage.

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.

`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.

`LinkIt` *LinkIt*

Description

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

Usage

```
LinkIt(x, y, by, by.x = by,by.x = 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.

`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

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`)

Value

`z` The merged data frame.

`MatchPerformance` *MatchPerformance*

Description

Record linkage description.

Usage

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

Arguments

`x, y`	data frames to be merged
`z`	the merged data frame to be analyzed
`by, by.x, by.y`	character strings specifying of the columns used for merging.
`z_true`	a reference data frame containing target/true matched dataset.

Details

Details

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.

Index