

Package ‘LinkOrgs’

January 19, 2024

Title LinkOrgs: Algorithms for Organizational Record Linkage

Version 0.01

Description An R package for organizational records using the algorithms of Jerzak & Libgobner (2023+). The linkage is done based on organizational names and using half a billion open collaborated records on those names from LinkedIn users. It also contains functions implementing string matching performance metrics, as well as a fast, parallized version of fuzzy string matching.

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,Rfast,stringdist,doMC,parallel,glmnet,parallel,stringr,dplyr,fastmatch,reticulate

RoxygenNote 7.2.3

R topics documented:

AssessMatchPerformance	1
BuildBackend	3
BuildTransfer	3
FastFuzzyMatch	4
LinkOrgs	5
url2dt	7
Index	9

AssessMatchPerformance
<i>AssessMatchPerformance</i>

Description

Computes the true/false positive and true/false negative rates of a candidate matching based on a ground-truth (preferably human-generated) matched dataset.

Usage

```
AssessMatchPerformance(
  x,
  y,
  z,
  z_true,
  by,
  by.x = by,
  by.y = by,
  openBrowser = F
)
```

Arguments

<code>x, y</code>	data frames to be merged
<code>z</code>	the merged data frame to be analyzed. Should contain <code>by</code> , <code>by.x</code> , and/or <code>by.y</code> as column names, depending on usage.
<code>z_true</code>	a reference data frame containing target/true matched dataset. Should contain <code>by</code> , <code>by.x</code> , and/or <code>by.y</code> as column names, depending on usage.
<code>by, by.x, by.y</code>	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)` candidate match pairs.

Examples

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

# Obtain match performance data
PerformanceMatrix <- AssessMatchPerformance(x = x,
                                             y = y,
                                             z = z,
                                             z_true = z_true,
                                             by.x = "orenames_x",
                                             by.y = "orenames_y")

print( PerformanceMatrix )
```

BuildBackend	<i>Build the environment for LinkOrgs machine learning models. Builds a conda environment in which jax, optax, equinox, and jmp are installed.</i>
--------------	--

Description

Build the environment for LinkOrgs machine learning models. Builds a conda environment in which jax, optax, equinox, and jmp are installed.

Usage

```
BuildBackend(conda_env = "LinkOrgsEnv", conda = "auto")
```

Arguments

conda_env	(default = "LinkOrgsEnv") Name of the conda environment in which to place the backends.
conda	(default = auto) The path to a conda executable. Using "auto" allows reticulate to attempt to automatically find an appropriate conda binary.

Value

Builds the computational environment for LinkOrgs. This function requires an Internet connection. You may find out a list of conda Python paths via: `system("which python")`

Examples

```
# For a tutorial, see
# github.com/cjerzak/linkorgs-software/
```

BuildTransfer	<i>A primarily internal function which builds the organizational record linkage models used in Libgober and Jerzak (2023+).</i>
---------------	---

Description

A primarily internal function which builds the organizational record linkage models used in Libgober and Jerzak (2023+).

Usage

```
BuildTransfer()
```

FastFuzzyMatch	<i>FastFuzzyMatch</i>
----------------	-----------------------

Description

Performs fast fuzzy matching of strings based on the string distance measure specified in `DistanceMeasure`. Matching is parallelized using all available CPU cores to increase execution speed.

Usage

```
FastFuzzyMatch(
  x,
  y,
  by = NULL,
  by.x = NULL,
  by.y = NULL,
  return_stringdist = T,
  onlyUFT = T,
  qgram = 2,
  DistanceMeasure = "jaccard",
  MaxDist = 0.2,
  AverageReference = NULL,
  AveMatchNumberPerAlias = NULL,
  openBrowser = F,
  ReturnProgress = T,
  ReturnMaxDistThreshold = F
)
```

Arguments

<code>x, y</code>	data frames to be merged
<code>by, by.x, by.y</code>	specifications of the columns used for merging. We follow the general syntax of <code>base::merge</code> ; see <code>?base::merge</code> for more details.
<code>...</code>	For additional options, see “Details”.

Details

`FastFuzzyMatch` can automatically process the `by` text for each dataset. Users may specify the following options:

- 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 `MaxDist` to control the maximum allowed distance between two matched strings
- Set `AveMatchNumberPerAlias` to control the maximum allowed distance between two matched strings. Takes priority over `MaxDist` if both specified.
- 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.

Examples

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

# Perform merge
linkedOrgs_fuzzy <- FastFuzzyMatch(x = x,
                                   y = y,
                                   by.x = "orenames_x",
                                   by.y = "orenames_y")
```

LinkOrgs

LinkOrgs

Description

Implements the organizational record linkage algorithms of Libgober and Jerzak (2023+) using half-a-billion open-collaborated records.

Usage

```
LinkOrgs(
  x,
  y,
  by = NULL,
  by.x = NULL,
  by.y = NULL,
  algorithm = "bipartite",
  conda_env = NULL,
  conda_env_required = F,
  ReturnDiagnostics = F,
  ReturnProgress = T,
  ToLower = T,
  NormalizeSpaces = T,
  RemovePunctuation = T,
  MaxDist = NULL,
  MaxDist_network = NULL,
  AveMatchNumberPerAlias = 10,
```

```

AveMatchNumberPerAlias_network = 2,
DistanceMeasure = "jaccard",
qgram = 2,
openBrowser = F,
ReturnDecomposition = F
)

```

Arguments

<code>x, y</code>	data frames to be merged
<code>by, by.x, by.y</code>	character vector(s) that specify the column names used for merging data frames <code>x</code> and <code>y</code> . The merging variables should be organizational names. See <code>?base::merge</code> for more details regarding syntax.
<code>algorithm</code>	character; specifies which algorithm described in Libgober and Jerzak (2023+) should be used. Options are "markov", "bipartite", "ml", and "transfer". Default is "ml", which uses a machine-learning approach using Transformer netes and 9 million parameters to predict match probabilities using half a billion open-collaborated records as training data.
<code>conda_env</code>	character string; specifies a conda environment where JAX and related packages have been installed (see <code>?LinkOrgs::BuildBackend</code>). Used only when <code>algorithm='ml'</code> or <code>DistanceMeasure='ml'</code> .
<code>conda_env_required</code>	Boolean; specifies whether conda environment is required.
<code>ReturnDiagnostics</code>	logical; specifies whether various match-level diagnostics should be returned in the merged data frame.
<code>ml_version</code>	character; specifies which version of the ML algorithm should be used. Options are of the form "v1", "v2", "v3"... Highest version currently supported is "v1" (11M parameters).
<code>...</code>	For additional specification options, see "Details".

Details

LinkOrgs automatically processes the name text for each dataset (specified by `by` or `by.x`, and `by.y`). Users may specify the following options:

- Set `DistanceMeasure` to control algorithm for computing pairwise string distances. Options include "osa", "jaccard", "jw". See `?stringdist::stringdist` for all options. Default is "jaccard". To use the combined machine learning and network methods, set `algorithm` to "bipartite" or "markov", and `DistanceMeasure` to "ml".
- Set `MaxDist` to control the maximum allowed distance between two matched strings
- Set `MaxDist_network` to control the maximum allowed distance between two matched strings in the integration with the LinkedIn network representation.
- Set `AveMatchNumberPerAlias` to control the maximum allowed distance between two matched strings. Takes priority over `MaxDist` if both specified.
- Set `AveMatchNumberPerAlias_network` to control the maximum allowed distance between two matched strings in the integration with the LinkedIn network representation. Takes priority over `MaxDist_network` if both specified.
- 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.

Examples

```
#Create synthetic data
x_orenames <- c("apple","oracle","enron inc.,"mcdonalds corporation")
y_orenames <- c("apple corp","oracle inc","enron","mcdonalds co")
x <- data.frame("orenames_x"=x_orenames)
y <- data.frame("orenames_y"=y_orenames)

# Perform merge
linkedOrgs <- LinkOrgs(x = x,
                       y = y,
                       by.x = "orenames_x",
                       by.y = "orenames_y",
                       MaxDist = 0.6)

print( linkedOrgs )
```

url2dt

url2dt

Description

Downloads a .zip file from a URL as a data.table from a URL.

Usage

```
url2dt(url, target_extension = ".csv")
```

Arguments

url character string with the URL housing the data object.

target_extension (default = ".csv") character string describing the target extension of the file in the downloaded .zip folder.

Details

url2dt downloads a zipped .csv file and loads it into memory based on the input URL.

Value

z The downloaded data object from the URL.

Examples

```
# Example download
```

```
my_dt <- url2dt(url="https://www.dropbox.com/s/iqf9ids77dckopf/Directory_LinkIt_bipartite_Embeddings.csv.zi
```


Index

AssessMatchPerformance, [1](#)

BuildBackend, [3](#)

BuildTransfer, [3](#)

FastFuzzyMatch, [4](#)

LinkOrgs, [5](#)

url2dt, [7](#)