Package ‘KOR.addrlink’

March 5, 2024

Type Package

Title Matching Address Data to Reference Index

Version 1.0.1

Date 2024-03-02

Author Daniel Schürmann [aut, cre]

Maintainer Daniel Schürmann <d.schuermann@2718282.net>

Depends R (>= 3.4)

Imports stringdist, stringi

LazyData true

Description Matches a data set with semi-structured address data,
e.g., street and house number as a concatenated string,
wrongly spelled street names or non-existing house numbers to a
reference index. The methods are specifically designed for German
municipalities ('KOR'-community) and German address schemes.

License GPL-3

Encoding UTF-8

URL https://git-kor.stadtdo.de

BugReports https://git-kor.stadtdo.de/stadt-dortmund/adressdaten/-/issues

NeedsCompilation no

Repository CRAN

Date/Publication 2024-03-05 11:10:13 UTC

R topics documented:

`KOR.addrlink-package` *KOR.addrlink*

Description

Geocode address data from German municipalities

Details

- `split_address` Splits strings into street, house number and addional letter
- `split_number` Splits strings into house number and addional letter
- `addrlink` Matches splitted address data to reference table

Matching is based on street name, house number and additional letter.

Author(s)

Daniel Schürmann

`addrlink` *Merge Data To Reference Index*

Description

Takes two data.frames with address data and merges them together.

Usage

```
addrlink(df_ref, df_match,
col_ref = c("Strasse", "Hausnummer", "Hausnummernzusatz"),
col_match = c("Strasse", "Hausnummer", "Hausnummernzusatz"),
fuzzy_threshold = 0.9, seed = 1234)
```

Arguments

`df_ref` data.frame with address references

`df_match` data.frame with addresses to be matched

`col_ref` character vector of length three, naming the df_ref columns which contain the steet names, house numbers and additional letters (in that order)

`col_match` character vector of length three, naming the df_match columns which contain the steet names, house numbers and additional letters (in that order)

`fuzzy_threshold` The threshold used for fuzzy matching street names

`seed` Seed for random numbers

Details

The matching is done in four stages.

Stage 1 (qAdress = 1). This is an exact match (highest quality, qscore = 1)

Stage 2 (qAdress = 2). Exact match on street name, but no valid house number could be found. Be aware that random house numbers might be used. Consider setting your own seed. qscore indicates the match quality. See `match_number` for details.

Stage 3 (qAdress = 3). No exact match on street name could be found. Street names are fuzzy matched. The method "jw" (Jaro-Winkler distance) from package stringdist is used (see stringdist-metrics). If 1 - [Jaro-Winkler distance] is greater than fuzzy_threshold, a match is assumed. The highest score is taken and house number matching is done as outlined in Stage 2. qscore is fuzzy_score*[house number score].

Stage 4 (qAdress = 4). No match (qscore = 0)

Value

A list

`ret` The merged dataset

`QA` The quality markers (qAdress and qscore)

Author(s)

Daniel Schürmann

See Also

`split_address`, `split_number`

`Adressen` *Address data from the city of Dortmund*

Description

This data set gives all the addresses in the city of Dortmund.

Usage

```
Adressen
```

Format

A data.frame

STRNAME	character	street name
STRSL	numeric	street number
HNR	numeric	house number
HNRZ	character	additional letter
RW	numeric	longitude
HW	numeric	latitude
UBZ	numeric	subdistrict number

Source

https://open-data.dortmund.de

`df1` *Example dataset 1*

Description

This dataset contains separate street and house number information.

Usage

```
df1
```

Format

A data.frame

gross_strasse	character	street names
hausnr	character	house number and additional letter
Var1	numeric	Variable 1
Var2	character	Variable 2

Source

Dortmunder Statistik

df2 *Example dataset 2*

Description

This dataset contains concatenated street and house number information.

Usage

```
df2
```

Format

A data.frame

Adresse	character	street name, house number and addional letter
Var1	numeric	Variable 1
Var2	character	Variable 2

Source

Dortmunder Statistik

helper_split_address *Splits A Single Address Into Street, House Number And Additional Letter*

Description

This is an internal function. Please use `split_address`

Usage

```
helper_split_address(x, debug = FALSE)
```

Arguments

`x`	A character vector of length 1
`debug`	If true, print(x)

Value

A list with three elements

strasse	Extracted street name
hnr	Extracted house number
hnrz	Extracted extra letter

Author(s)

Daniel Schürmann

See Also

split_address

---

helper_split_number *Splits A Single House Number Into House Number And Additional Letter*

Description

This is an internal function. Please use split_number

Usage

```
helper_split_number(x, debug = FALSE)
```

Arguments

x	A character vector of length 1
debug	If true, print(x)

Value

A data.frame with two elements

Hausnummer	Extracted house number
Zusatz	Extracted extra letter

Author(s)

Daniel Schürmann

See Also

split_number

`l1score` *Calculate L1-Distance Based Scores*

Description

Reversed normalized absolute distance from zero.

Usage

```
l1score(x)
```

Arguments

`x` A numeric vector

Details

$$1 - \frac{|x|}{\max\{1, |x|\}}$$

Value

A numeric vector of the same length as x

Author(s)

Daniel Schürmann

`match_number` *Find Best House Number Match Within Given Street*

Description

This is an internal function. Please use `addrlink`

Usage

```
match_number(record, Adressen, weights = c(0.9, 0.1))
```

Arguments

`record` data.frame with one row and three columns (Strasse, Hausnummer, Hausnummernzusatz)

`Adressen` data.frame of all valid addresses (same columns as record data.frame)

`weights` The weighing factors between house number and additional letter

Details

If no house number and no additional letter is provided, a random address in the given street is selected (qscore = 0).

If only an additional letter but no house number is given and the letter is unique, returns the corresponding record (qscore = 0.05). Otherwise returns a random one as mentioned above (qscore = 0).

If no additional letter, but house number is provided and the maximum distance to a valid house number is 4, return the closest match as calculated by `l1score` (qscore is the result of l1score). Otherwise a random record is returned (qscore = 0).

If additional letter and house number are available and the house number distance is smaller then 4, calculates the l1scores of the house number distance and addional letters distance and selects the best match (qscore is the sum of both weighted l1scores). Otherwise a random record is selected (qscore = 0).

Value

A data.frame

`qscore`	The quality score of the match
`Strasse`	matched street
`Hausnummer`	matched house number
`Hausnummernzusatz`	matched additional letter

Author(s)

Daniel Schürmann

See Also

`addrlink`

`sanitize_street` *Clean Steet Names And Make Them Mergeable*

Description

This function replaces Umlauts, expands "str" to "strasse", transliterates all non-ascii characters, removes punctuation and converts to lower case.

Usage

```
sanitize_street(x)
```

Arguments

`x`	A character vector containing the steet names

Details

This is an internal function used in addrlink. Make sure house numbers have already been extracted. Use split_number or split_address for that. Only steet names can go into sanitize_street.

Value

A character vector of the same length as x containing the sanitized street names.

Author(s)

Daniel Schürmann

See Also

split_address, split_number, addrlink

---

split_address *Split Adresses Into Street, House Number And Additional Letter*

Description

This function takes a character vector where each element is made up from a concatenation of street name, house number and possibly an additional letter and splits it into its parts.

Usage

```
split_address(x, debug = FALSE)
```

Arguments

x	A character vector
debug	If true, all records will be printed to the console

Details

If the function fails, consider using debug = TRUE. This will print the record, which caused the error. Consider filing an issue on the linked git project (see DESCRIPTION).

Value

A data.frame with three columns

Strasse	A character column containing the extracted street names
Hausnummer	House number
Hausnummernzusatz	Additional letter

Note

For a more advanced, general purpose solution see libpostal.

Author(s)

Daniel Schürmann

See Also

`split_number`

Examples

```
split_address(c("Teststr. 8-9 a", "Erster Weg 1-2", "Ahornallee 100a-102c"))
```

`split_number` *Split house number into house number and additional letter*

Description

This function takes a character vector where each element is made up from a concatenation of house number and possibly an additional letter and splits is into its parts.

Usage

```
split_number(x, debug = FALSE)
```

Arguments

`x`	A character vector
`debug`	If true, all records will be printed to the console

Details

If the function fails, consider using `debug = TRUE`. This will print the record, which caused the error. Consider filing an issue on the linked git project (see DESCRIPTION).

Value

A data.frame with two columns

`Hausnummer`	House number
`Hausnummernzusatz`	Additional letter

Note

For a more advanced, general purpose solution see libpostal.

Author(s)

Daniel Schürmann

See Also

`split_address`

Examples

```
split_number(c("8-9 a", "1-2", "100a-102c"))
```

Index