Package 'uniqtag'

October 12, 2022

Type Package

Title Abbreviate Strings to Short, Unique Identifiers

Version 1.0.1

Description For each string in a set of strings, determine a unique tag that is a substring of fixed size k unique to that string, if it has one. If no such unique substring exists, the least frequent substring is used. If multiple unique substrings exist, the lexicographically smallest substring is used. This lexicographically smallest substring of size k is called the ``UniqTag'' of that string.

License MIT + file LICENSE

Encoding UTF-8

RoxygenNote 7.1.2

URL https://github.com/sjackman/uniqtag

BugReports https://github.com/sjackman/uniqtag/issues

Suggests testthat

NeedsCompilation no

Author Shaun Jackman [aut, cph, cre]

Maintainer Shaun Jackman <sjackman@gmail.com>

Repository CRAN

Date/Publication 2022-06-10 06:10:02 UTC

R topics documented:

uniqtag-package *Abbreviate strings to short, unique identifiers.*

Description

For each string in a set of strings, determine a unique tag that is a substring of fixed size k unique to that string, if it has one. If no such unique substring exists, the least frequent substring is used. If multiple unique substrings exist, the lexicographically smallest substring is used. This lexicographically smallest substring of size k is called the "UniqTag" of that string.

Author(s)

Shaun Jackman <sjackman@gmail.com>

cumcount *Cumulative count of strings.*

Description

Return an integer vector counting the number of occurrences of each string up to that position in the vector.

Usage

```
cumcount(xs)
```

Arguments

xs	a character vector

Value

an integer vector of the cumulative string counts

Examples

```
cumcount(abbreviate(state.name, 3, strict = TRUE))
```

kmers_of *Return the k-mers of a string.*

Description

Return the k-mers (substrings of size k) of the string x, or return the string x itself if it is shorter than k.

Usage

```
kmers_of(x, k)

vkmers_of(xs, k)
```

Arguments

x	a character string
k	the size of the substrings, an integer
xs	a character vector

Value

kmers_of: a character vector of the k-mers of x

vkmers_of: a list of character vectors of the k-mers of xs

Functions

- kmers_of: Return the k-mers of the string x.
- vkmers_of: Return the k-mers of the strings xs.

make_unique *Make character strings unique.*

Description

Append sequence numbers to duplicate elements to make all elements of a character vector unique.

Usage

```
make_unique(xs, sep = "-")

make_unique_duplicates(xs, sep = "-")

make_unique_all(xs, sep = "-")

make_unique_all_or_none(xs, sep = "-")
```

Arguments

xs	a character vector
sep	a character string used to separate a duplicate string from its sequence number

Functions

- make_unique: Append a sequence number to duplicated elements, including the first occurrence.
- make_unique_duplicates: Append a sequence number to duplicated elements, except the first occurrence.

  This function behaves similarly to make.unique
- make_unique_all: Append a sequence number to every element.
- make_unique_all_or_none: Append a sequence number to every element or no elements.

  Return xs unchanged if the elements of the character vector xs are already unique. Otherwise append a sequence number to every element.

See Also

make.unique

Examples

```
abcb <- c("a", "b", "c", "b")
make_unique(abcb)
make_unique_duplicates(abcb)
make_unique_all(abcb)
make_unique_all_or_none(abcb)
make_unique_all_or_none(c("a", "b", "c"))
x <- make_unique(abbreviate(state.name, 3, strict = TRUE))
x[grep("-", x)]
```

---

uniqtag *Abbreviate strings to short, unique identifiers.*

Description

Abbreviate strings to unique substrings of k characters.

Usage

```
uniqtag(xs, k = 9, uniq = make_unique_all_or_none, sep = "-")
```

Arguments

xs	a character vector
k	the size of the identifier, an integer
uniq	a function to make the abbreviations unique, such as make_unique, make_unique_duplicates, make_unique_all_or_none, make_unique_all, make.unique, or to disable this function, identity or NULL
sep	a character string used to separate a duplicate string from its sequence number

Details

For each string in a set of strings, determine a unique tag that is a substring of fixed size k unique to that string, if it has one. If no such unique substring exists, the least frequent substring is used. If multiple unique substrings exist, the lexicographically smallest substring is used. This lexicographically smallest substring of size k is called the UniqTag of that string.

The lexicographically smallest substring depend on the locale's sort order. You may wish to first call `Sys.setlocale("LC_COLLATE", "C")`

Value

a character vector of the UniqTags of the strings x

See Also

abbreviate, locales, make.unique

Examples

```
Sys.setlocale("LC_COLLATE", "C")
states <- sub(" ", "", state.name)
uniqtags <- uniqtag(states)
uniqtags4 <- uniqtag(states, k = 4)
uniqtags3 <- uniqtag(states, k = 3)
uniqtags3x <- uniqtag(states, k = 3, uniq = make_unique)
table(nchar(states))
table(nchar(uniqtags))
table(nchar(uniqtags4))
table(nchar(uniqtags3))
table(nchar(uniqtags3x))
uniqtags3[grep("-", uniqtags3x)]
```

Index