

# Package ‘uniqtag’

August 29, 2016

**Type** Package

**Version** 1.0

**Title** 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.

**URL** <https://github.com/sjackman/uniqtag>

**BugReports** <https://github.com/sjackman/uniqtag/issues>

**Suggests** testthat

**License** MIT + file LICENSE

**Author** Shaun Jackman [cre]

**Maintainer** Shaun Jackman <sjackman@gmail.com>

**NeedsCompilation** no

**Repository** CRAN

**Date/Publication** 2015-04-29 01:17:59

## R topics documented:

uniqtag-package . . . . .	2
cumcount . . . . .	2
kmers_of . . . . .	3
make_unique . . . . .	3
uniqtag . . . . .	4

<b>Index</b>	<b>6</b>
--------------	----------

---

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

---

**Description**

Abbreviate strings to short, unique identifiers.

**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	<i>Return the k-mers of a string.</i>
----------	---------------------------------------

---

**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	<i>Make character strings unique.</i>
-------------	---------------------------------------

---

**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**

<code>xs</code>	a character vector
<code>sep</code>	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**

<code>xs</code>	a character vector
<code>k</code>	the size of the identifier, an integer
<code>uniq</code>	a function to make the abbreviations unique, such as <code>make_unique</code> , <code>make_unique_duplicates</code> , <code>make_unique_all_or_none</code> , <code>make_unique_all</code> , <code>make.unique</code> , or to disable this function, <code>identity</code> or <code>NULL</code>
<code>sep</code>	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("-", unqtags3x)]
```

# Index

`cumcount`, [2](#)

`kmers_of`, [3](#)

`make_unique`, [3](#)

`make_unique_all` (`make_unique`), [3](#)

`make_unique_all_or_none` (`make_unique`), [3](#)

`make_unique_duplicates` (`make_unique`), [3](#)

`uniqtag`, [4](#)

`uniqtag-package`, [2](#)

`vkmers_of` (`kmers_of`), [3](#)