

Package: vibrtrt (via r-universe)

October 7, 2026

Type Package

Title An R Wrapper for 'vibrato'

Version 0.5.2.3

Maintainer Akiru Kato <paithiov909@gmail.com>

Description An R wrapper for 'vibrato'
<<https://github.com/daac-tools/vibrato>>, a Rust
reimplementation of 'MeCab' for fast tokenization.

License Apache License (>= 2)

URL <https://paithiov909.github.io/vibrtrt/>

BugReports <https://github.com/paithiov909/vibrtrt/issues>

Depends R (>= 4.2)

Imports dplyr, purrr, readr, rlang (>= 0.4.11), stringi

Suggests curl, hfhub, jsonlite, roxygen2, testthat (>= 3.0.0), withr

Config/roxygen2/version 8.1.0

Config/testthat/edition 3

Encoding UTF-8

Roxygen list(markdown = TRUE)

SystemRequirements Cargo (Rust's package manager), rustc

Config/pak/sysreqs libc-dev libx11-dev libclang-dev

Repository <https://paithiov909.r-universe.dev>

Date/Publication 2026-09-07 09:35:04 UTC

RemoteUrl <https://github.com/paithiov909/vibrtrt>

RemoteRef HEAD

RemoteSha cd31472bfbc86ad24599926b2299de940d158dd4

Contents

as_tokens	2
create_tagger	3
get_dict_features	3
is_blank	4
ngram_tokenizer	5
pack	5
prettify	6
tokenize	7
Index	8

as_tokens	Create a list of tokens
-----------	-------------------------

Description

Create a list of tokens

Usage

```
as_tokens(  
  tbl,  
  token_field = "token",  
  pos_field = get_dict_features()[1],  
  nm = NULL  
)
```

Arguments

- | | |
|-------------|---|
| tbl | A tibble of tokens out of tokenize(). |
| token_field | <data-masked> Column containing tokens. |
| pos_field | Column containing features that will be kept as the names of tokens. If you don't need them, give a NULL for this argument. |
| nm | Names of returned list. If left with NULL, "doc_id" field of tbl is used instead. |

Value

A named list of tokens.

create_tagger	<i>Create a tagger function</i>
---------------	---------------------------------

Description

Create a tagger function

Usage

```
create_tagger(sys_dic, user_dic = "", max_grouping_len = 0L, verbose = FALSE)
```

Arguments

sys_dic	Character scalar; path to the system dictionary for 'vibrato'.
user_dic	Character scalar; path to the user dictionary for 'vibrato'.
max_grouping_len	Integer scalar; The maximum grouping length for unknown words. The default value is 0L, indicating the infinity length.
verbose	Logical. If TRUE, returns additional information for debugging.

Value

A function inheriting class `purrr_function_partial`.

get_dict_features	<i>Get dictionary features</i>
-------------------	--------------------------------

Description

Returns names of dictionary features. Currently supports "unidic17" (2.1.2 src schema), "unidic26" (2.1.2 bin schema), "unidic29" (schema used in 2.2.0, 2.3.0), "cc-cedict", "ko-dic" (mecab-ko-dic), "naist11", and "ipa".

Usage

```
get_dict_features(
  dict = c("ipa", "unidic17", "unidic26", "unidic29", "cc-cedict", "ko-dic", "naist11")
)
```

Arguments

dict	Character scalar; one of "ipa", "unidic17", "unidic26", "unidic29", "cc-cedict", "ko-dic", or, "naist11".
------	---

Value

A character vector.

See Also

See also `'CC-CEDICT-MeCab'` and `'mecab-ko-dic'`.

Examples

```
get_dict_features("ipa")
```

is_blank	<i>Check if scalars are blank</i>
----------	-----------------------------------

Description

Check if scalars are blank

Usage

```
is_blank(x, trim = TRUE, ...)
```

Arguments

x	Object to check its emptiness.
trim	Logical.
...	Additional arguments for <code>base::sapply()</code> .

Value

Logicals.

Examples

```
is_blank(list(c(a = "", b = NA_character_), NULL))
```

ngram_tokenizer	<i>Ngrams tokenizer</i>
-----------------	-------------------------

Description

Makes an ngram tokenizer function.

Usage

```
ngram_tokenizer(n = 1L)
```

Arguments

n Integer.

Value

ngram tokenizer function

Examples

```
bigram <- ngram_tokenizer(2)
bigram(letters, sep = "-")
```

pack	<i>Pack a data.frame of tokens</i>
------	------------------------------------

Description

Packs a data.frame of tokens into a new data.frame of corpus, which is compatible with the Text Interchange Formats.

Usage

```
pack(tbl, pull = "token", n = 1L, sep = "-", .collapse = " ")
```

Arguments

tbl	A data.frame of tokens.
pull	<data-masked> Column to be packed into text or ngrams body. Default value is token.
n	Integer internally passed to ngrams tokenizer function created of <code>ngram_tokenizer()</code>
sep	Character scalar internally used as the concatenator of ngrams.
.collapse	This argument is passed to <code>stringi::stri_c()</code> .

Value

A tibble.

Text Interchange Formats (TIF)

The Text Interchange Formats (TIF) is a set of standards that allows R text analysis packages to target defined inputs and outputs for corpora, tokens, and document-term matrices.

Valid data.frame of tokens

The data.frame of tokens here is a data.frame object compatible with the TIF.

A TIF valid data.frame of tokens is expected to have one unique key column (named doc_id) of each text and several feature columns of each tokens. The feature columns must contain at least token itself.

See Also

<https://github.com/ropenscilabs/tif>

```
prettify
```

```
Prettify tokenized output
```

Description

Turns a single character column into features while separating with delimiter.

Usage

```
prettify(
  tbl,
  col = "feature",
  into = get_dict_features("ipa"),
  col_select = seq_along(into),
  delim = ",",
)
```

Arguments

tbl	A data.frame that has feature column to be prettified.
col	<data-masked> Column containing features to be prettified.
into	Character vector that is used as column names of features.
col_select	Character or integer vector that will be kept in prettified features.
delim	Character scalar used to separate fields within a feature.

Value

A data.frame.

tokenize	<i>Tokenize sentences using a tagger</i>
----------	--

Description

Tokenize sentences using a tagger

Usage

```
tokenize(  
  x,  
  text_field = "text",  
  docid_field = "doc_id",  
  split = FALSE,  
  mode = c("parse", "wakati"),  
  tagger  
)
```

Arguments

x	A data.frame like object or a character vector to be tokenized.
text_field	<data-masked> String or symbol; column containing texts to be tokenized.
docid_field	<data-masked> String or symbol; column containing document IDs.
split	split Logical. When passed as TRUE, the function internally splits the sentences into sub-sentences
mode	Character scalar to switch output format.
tagger	A tagger function created by create_tagger() .

Value

A tibble or a named list of tokens.

Index

`as_tokens`, [2](#)

`base::sapply()`, [4](#)

`create_tagger`, [3](#)

`create_tagger()`, [7](#)

`get_dict_features`, [3](#)

`is_blank`, [4](#)

`ngram_tokenizer`, [5](#)

`ngram_tokenizer()`, [5](#)

`pack`, [5](#)

`prettify`, [6](#)

`stringi::stri_c()`, [5](#)

`tokenize`, [7](#)