

# Package: tangela (via r-universe)

August 25, 2024

**Type** Package

**Title** rJava Interface to Kuromoji

**Version** 0.1.1

**Maintainer** Akiru Kato <paithiov909@gmail.com>

**Description** An rJava wrapper of atilika/kuromoji (v0.7.7).

**License** Apache License (>= 2)

**URL** <https://github.com/paithiov909/tangela>

**BugReports** <https://github.com/paithiov909/tangela/issues>

**Depends** R (>= 3.2.0)

**Imports** audubon (>= 0.3.0), dplyr, purrr, rJava, rlang (>= 0.1.2),  
stringi

**Suggests** knitr, rmarkdown, roxygen2

**Encoding** UTF-8

**LazyData** true

**Roxygen** list(markdown = TRUE)

**RoxygenNote** 7.2.3

**SystemRequirements** Java

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

**RemoteUrl** <https://github.com/paithiov909/tangela>

**RemoteRef** HEAD

**RemoteSha** f53f797d93396cdf3954036ea0e0da62c7d02f0c

## Contents

get_dict_features . . . . .	2
kuromoji . . . . .	2
pack . . . . .	3
prettify . . . . .	4
rebuild_tokenizer . . . . .	4

<b>Index</b>	<b>5</b>
--------------	----------

---

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

---

### Description

Returns dictionary's 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", "sudachi", and "ipa".

### Usage

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

### Arguments

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

### Value

A character vector.

### See Also

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

---

kuromoji	<i>Call kuromoji tokenizer</i>
----------	--------------------------------

---

### Description

Call kuromoji tokenizer

### Usage

```
kuromoji(chr)
```

### Arguments

chr	Character vector to be tokenized.
-----	-----------------------------------

### Value

A data.frame

---

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

---

## Description

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

## Value

A data.frame.

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

A TIF valid data.frame of tokens are 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	<i>Prettify tokenized output</i>
----------	----------------------------------

---

**Description**

Turns a single character column into features separating with delimiter.

**Usage**

```
prettify(df, into = get_dict_features("ipa"), col_select = seq_along(into))
```

**Arguments**

df	A data.frame that has feature column 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.

**Value**

A data.frame.

---

rebuild_tokenizer	<i>Initialize kuromoji tokenizer</i>
-------------------	--------------------------------------

---

**Description**

Initialize kuromoji tokenizer

**Usage**

```
rebuild_tokenizer(user_dic = "")
```

**Arguments**

user_dic	file path to a user dictionary if any.
----------	--

**Value**

The stored kuromoji tokenizer instance is returned invisibly.

# Index

`get_dict_features`, [2](#)

`kuromoji`, [2](#)

`pack`, [3](#)

`prettify`, [4](#)

`rebuild_tokenizer`, [4](#)