Package: sudachir (via r-universe)

August 26, 2024

Type Package

Title R Interface to 'Sudachi'

Version 0.1.0.9007

Maintainer Shinya Uryu <suika1127@gmail.com>

Description Interface to 'Sudachi' <https://github.com/WorksApplications/sudachi.rs>, a Japanese morphological analyzer. This is a port of what is available in Python.

License Apache License (>= 2.0)

URL https://github.com/uribo/sudachir

BugReports https://github.com/uribo/sudachir/issues

Imports audubon (>= 0.4.0), cli, dplyr, reticulate, rlang (>= 0.4.11), rstudioapi

Suggests roxygen2, testthat

Encoding UTF-8

Roxygen list(markdown = TRUE)

RoxygenNote 7.2.3

SystemRequirements Python (>= 3.6)

Repository https://uribo.r-universe.dev

RemoteUrl https://github.com/uribo/sudachir

RemoteRef HEAD

RemoteSha fd723a1db7e5f670b3dfd230acd437a2899904d6

Contents

`as_tokens` *Create a list of tokens*

Description

Create a list of tokens

Usage

```
as_tokens(tbl, type, pos = TRUE, ...)
```

Arguments

`tbl`	A data.frame of tokens out of `tokenize_to_df()`.
`type`	Preference for the format of returned tokens. Pick one of "surface", "dictionary", "normalized", or "reading".
`pos`	When passed as `TRUE`, this function uses the part-of-speech information as the name of the returned tokens.
`...`	Passed to `dict_features()`.

Value

A named list of character vectors.

Examples

```
## Not run:
tokenize_to_df("Tokyo, Japan") |>
  as_tokens(type = "surface")

## End(Not run)
```

create_sudachipy_env *Create virtualenv env used by sudachir*

Description

Create virtualenv env used by sudachir

Usage

```
create_sudachipy_env(python_version = "3.9.12")
```

Arguments

python_version	Python version to use within the virtualenv environment created. SudachiPy requires Python 3.6 or higher to install.

dict_features *Get dictionary's features*

Description

Get dictionary's features

Usage

```
dict_features(lang = c("ja", "en"))
```

Arguments

lang	Dictionary features label; one of "ja" or "en".

Examples

```
dict_features("en")
```

form *Create a list of tokens*

Description

This function is a shorthand of `tokenize_to_df() |> as_tokens()`.

Usage

```
form(
  x,
  text_field = "text",
  docid_field = "doc_id",
  instance = rebuild_tokenizer(),
  ...
)
```

Arguments

x	A data.frame like object or a character vector to be tokenized.
text_field	Column name where to get texts to be tokenized.
docid_field	Column name where to get identifiers of texts.
instance	A binding to the instance of <sudachipy.tokenizer.Tokenizer>. If you already have a tokenizer instance, you can improve performance by providing a predefined instance.
...	Passed to `as_tokens()`.

Value

A named list of character vectors.

Examples

```
## Not run:
form(
  "Tokyo, Japan",
  type = "surface"
)

## End(Not run)
```

install_sudachipy *Install SudachiPy*

Description

Install SudachiPy to virtualenv virtual environment. As a one-time setup step, you can run `install_sudachipy()` to install all dependencies.

Usage

```
install_sudachipy()
```

Details

`install_sudachipy()` requires Python and virtualenv installed. See https://www.python.org/getit/.

Examples

```
## Not run:
install_sudachipy()

## End(Not run)
```

rebuild_tokenizer *Rebuild 'Sudachi' tokenizer*

Description

Rebuild 'Sudachi' tokenizer

Usage

```
rebuild_tokenizer(
  mode = c("C", "B", "A"),
  dict_type = c("core", "small", "full"),
  config_path = NULL
)
```

Arguments

`mode`	Split mode (A, B, C)
`dict_type`	Dictionary type.
`config_path`	Absolute path to `sudachi.json`.

Value

Returns a binding to the instance of <sudachipy.tokenizer.Tokenizer>.

Examples

```
## Not run:
tokenizer <- rebuild_tokenizer()
tokenize_to_df("Tokyo, Japan", instance = tokenizer)

## End(Not run)
```

`remove_sudachipy` *Remove SudachiPy*

Description

Uninstalls SudachiPy by removing the virtualenv environment.

Usage

```
remove_sudachipy()
```

Examples

```
## Not run:
install_sudachipy()
remove_sudachipy()

## End(Not run)
```

`sudachir-defunct` *'Sudachi' tokenizer*

Description

The old `tokenizer()` function was removed.

Usage

```
tokenizer(...)
```

Arguments

...	Not used.

Details

In general, users should not directly touch the <sudachipy.tokenizer.Tokenizer> and its MorphemeList objects. If you must access those objects, use the return value of the rebuild_tokenizer() function.

---

tokenize_to_df *Create a data.frame of tokens*

Description

Create a data.frame of tokens

Usage

```
tokenize_to_df(
  x,
  text_field = "text",
  docid_field = "doc_id",
  into = dict_features(),
  col_select = seq_along(into),
  instance = rebuild_tokenizer(),
  ...
)
```

Arguments

x	A data.frame like object or a character vector to be tokenized.
text_field	Column name where to get texts to be tokenized.
docid_field	Column name where to get identifiers of texts.
into	Column names of features.
col_select	Character or integer vector of column names that kept in the return value. When passed as NULL, returns comma-separated features as is.
instance	A binding to the instance of <sudachipy.tokenizer.Tokenizer>. If you already have a tokenizer instance, you can improve performance by providing a predefined instance.
...	Currently not used.

Value

A tibble.

Examples

```
## Not run:
tokenize_to_df(
  "Tokyo, Japan",
  into = dict_features("en"),
  col_select = c("pos1", "pos2")
)

## End(Not run)
```

Index