Package: ldccr (via r-universe)

July 6, 2024

Title Utilities for Various Japanese Corpora

Version 2024.07.07

Maintainer Akiru Kato <paithiov909@gmail.com>

Description The goal of ldccr package is to make easy to use Japanese language resources. This package provides parsers for several Japanese corpora that are free or open licensed and a downloader of zipped text files published on Aozora Bunko.

License MIT + file LICENSE

URL https://github.com/paithiov909/ldccr

BugReports https://github.com/paithiov909/ldccr/issues

Depends R (>= 2.10)

Imports dplyr, memoise, purrr (>= 1.0.0), RcppSimdJson, readr, rlang, stringi, tibble, utils, yesno

Suggests testthat (>= 3.0.0)

Config/testthat/edition 3

Encoding UTF-8

LazyData true

LazyDataCompression xz

Roxygen list(markdown = TRUE)

RoxygenNote 7.3.2

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

RemoteUrl https://github.com/paithiov909/ldccr

RemoteRef HEAD

RemoteSha 52c44cee7de0bd60227813e2d15d841bda8e1818

Contents

`AozoraBunkoSnapshot` *Meta data of text files published on Aozora Bunko*

Description

Meta data of text files published on Aozora Bunko

Usage

```
AozoraBunkoSnapshot
```

Format

An object of class `tbl_df` (inherits from `tbl`, `data.frame`) with 19168 rows and 55 columns.

Source

http://www.aozora.gr.jp/index_pages/list_person_all_extended_utf8.zip

See Also

The structure of the data is described here.

clean_emoji *Remove emojis*

Description

Remove emojis

Usage

```
clean_emoji(text, replacement = "")
```

Arguments

text	A character vector.
replacement	String.

Value

A character vector.

clean_url *Remove URLs*

Description

Remove URLs

Usage

```
clean_url(text, replacement = "")
```

Arguments

text	A character vector.
replacement	String.

Value

A character vector.

download_unidic *Download and unzip 'UniDic'*

Description

Download 'UniDic' of specified version into `dirname`. This function is partial port of polm/unidic-py. Note that to unzip dictionary will take up 770MB on disk after downloading.

Usage

```
download_unidic(version = "latest", dirname = "unidic")
```

Arguments

`version`	String; version of 'UniDic'.
`dirname`	String; directory where unzip the dictionary.

Value

Full path to `dirname` is returned invisibly.

is_within_era *Check if dates are within Japanese era*

Description

Check if dates are within Japanese era

Usage

```
is_within_era(date, era)
```

Arguments

`date`	Dates.
`era`	String.

Value

Logicals.

jrte_rte_files *Data for Textual Entailment*

Description

Data for Textual Entailment

Usage

```
jrte_rte_files(
  keep = c("rte.nlp2020_base", "rte.nlp2020_append", "rte.lrec2020_surf",
    "rte.lrec2020_sem_short", "rte.lrec2020_sem_long", "rte.lrec2020_me")
)
```

Arguments

keep	Character vector. File names to parse.

Value

tsv file names.

ldnws_categories *List of categories of the Livedoor News Corpus*

Description

List of categories of the Livedoor News Corpus

Usage

```
ldnws_categories(
  keep = c("dokujo-tsushin", "it-life-hack", "kaden-channel", "livedoor-homme",
    "movie-enter", "peachy", "smax", "sports-watch", "topic-news")
)
```

Arguments

keep	Character vector. File names to parse.

Value

A list.

NekoText — *Whole text of 'Wagahai Wa Neko Dearu' written by Natsume Souseki from Aozora Bunko*

Description

Whole text of 'Wagahai Wa Neko Dearu' written by Natsume Souseki from Aozora Bunko

Usage

```
NekoText
```

Format

An object of class `character` of length 2258.

Source

https://www.aozora.gr.jp/cards/000148/files/789_ruby_5639.zip

parse_jrte_reasoning — *Parse reasoning column of 'rte.*.tsv'*

Description

Parse reasoning column of 'rte.*.tsv'

Usage

```
parse_jrte_reasoning(tbl)
```

Arguments

`tbl`	A tibble returned from `read_jrte` which name is `rte.*.tsv`.

Value

A tibble.

parse_to_jdate *Parse dates to Japanese dates*

Description

Parse dates to Japanese dates

Usage

```
parse_to_jdate(date, format)
```

Arguments

date	Dates.
format	String.

Value

A chacter vector.

read_aozora *Download text file from Aozora Bunko*

Description

Download a file from specified URL, unzip and convert it to UTF-8.

Usage

```
read_aozora(
  url = "https://www.aozora.gr.jp/cards/000081/files/472_ruby_654.zip",
  txtname = NULL,
  directory = file.path(getwd(), "cache")
)
```

Arguments

url	URL of text download link.
txtname	New file name as which text is saved. If NULL provided, keeps name of the source file.
directory	Path where new file is saved.

Value

The path to the file downloaded.

read_ja_text8 *Read the ja.text8 corpus*

Description

Download and read the ja.text8 corpus as a tibble.

Usage

```
read_ja_text8(
  url =
   "https://s3-ap-northeast-1.amazonaws.com/dev.tech-sketch.jp/chakki/public/ja.text8.zip",
  size = NULL
)
```

Arguments

url	String.
size	Integer. If supplied, samples rows by this argument.

Details

By default, this function reads the ja.text8 corpus as a tibble by splitting it into sentences. The ja.text8 as whole corpus consists of over 582,000 sentences, 16,900,026 tokens, and 290,811 vocabularies.

Value

A tibble.

read_jrte *Read the JRTE Corpus*

Description

Download and read the Japanese Realistic Textual Entailment Corpus. The result of this function is memoised with `memoise::memoise` internally.

Usage

```
read_jrte(
 url = "https://github.com/megagonlabs/jrte-corpus/archive/refs/heads/master.zip",
  exdir = tempdir(),
  keep = jrte_rte_files(),
  keep_rhr = FALSE,
  keep_pn = FALSE
)
```

Arguments

url	String.
exdir	String. Path to tempolarily unzip text files.
keep	List. File names to parse and keep in returned value.
keep_rhr	Logical. If supplied TRUE, keeps rhr.tsv.
keep_pn	Logical. If supplied TRUE, keeps pn.tsv.

Value

A list of tibbles.

---

read_ldnws *Read the Livedoor News Corpus*

Description

Download and read the Livedoor News Corpus. The result of this function is memoised with memoise::memoise internally.

Usage

```
read_ldnws(
  url = "https://www.rondhuit.com/download/ldcc-20140209.tar.gz",
  exdir = tempdir(),
  keep = ldnws_categories(),
  collapse = "\n\n"
)
```

Arguments

url	String.
exdir	String. Path to tempolarily untar text files.
keep	List. Categories to parse and keep in data.frame.
collapse	String with which base::paste collapses lines.

Details

This function downloads the Livedoor News Corpus and parses it to a tibble. For details about the Livedoor News Corpus, please see this page.

Value

A tibble.

`unidic_availables` *List of available 'UniDic'*

Description

List of available 'UniDic'

Usage

```
unidic_availables()
```

Index