

Package ‘proustr’

April 3, 2019

Title Tools for Natural Language Processing in French

Version 0.4.0

Date 2019-02-05

Description

Tools for Natural Language Processing in French and texts from Marcel Proust's collection ``A La Recherche Du Temps Perdu''. The novels contained in this collection are ``Du cote de chez Swann'', ``A l'ombre des jeunes filles en fleurs'', ``Le Cote de Guer-
mantes'',
``Sodome et Gomorrhe I et II'', ``La Prisonniere'', ``Albertine dis-
parue'', and ``Le Temps retrouve''.

URL <https://github.com/ColinFay/proustr>

BugReports <https://github.com/ColinFay/proustr/issues>

Depends R (\geq 2.10)

License MIT + file LICENSE

Imports stringr, rlang, tidyr, tokenizers, SnowballC, attempt

LazyData true

RoxygenNote 6.1.0

Encoding UTF-8

Suggests testthat, knitr, rmarkdown, covr, dplyr

VignetteBuilder knitr

NeedsCompilation no

Author Colin Fay [aut, cre] (<https://orcid.org/0000-0001-7343-1846>)

Maintainer Colin Fay <contact@colinfay.me>

Repository CRAN

Date/Publication 2019-02-05 14:50:02 UTC

R topics documented:

albertinedisparue	2
alombredesjeunesfillesenfleurs	2
ducotedecheszwann	3
laprisonniere	3
lecotedeguermantes	4

letempretrouve	4
proust_books	5
proust_char	5
proust_characters	6
proust_random	6
proust_sentiments	7
proust_stopwords	7
pr_detect_days	8
pr_detect_months	8
pr_detect_pro	9
pr_keep_only_alnum	10
pr_normalize_punc	10
pr_stem_sentences	11
pr_stem_words	11
pr_unacent	12
sodomeetgomorrhe	12
stop_words	13

albertinedisparue	<i>Marcel Proust's novel "Albertine disparue"</i>
--------------------------	---

Description

A dataset containing Marcel Proust's "Albertine disparue". This text has been downloaded from WikiSource.

Usage

```
albertinedisparue
```

Format

A tibble with text, book, volume, and year

Source

```
!https://fr.wikisource.org/wiki/Albertine_disparue!
```

alombredesjeunesfillesenfleurs	<i>Marcel Proust's novel "À l'ombre des jeunes filles en fleurs"</i>
---------------------------------------	--

Description

A dataset containing Marcel Proust's "À l'ombre des jeunes filles en fleurs". This text has been downloaded from WikiSource.

Usage

```
alombredesjeunesfillesenfleurs
```

Format

A tibble with text, book, volume, and year

Source

`https://fr.wikisource.org/wiki/`

<code>ducotedechezswann</code>	<i>Marcel Proust's novel "Du côté de chez Swann"</i>
--------------------------------	--

Description

A dataset containing Marcel Proust's "Du côté de chez Swann". This text has been downloaded from WikiSource.

Usage

```
ducotedechezswann
```

Format

A tibble with text, book, volume, and year

Source

`https://fr.wikisource.org/wiki/Du_c`

<code>laprisonniere</code>	<i>Marcel Proust's novel "La Prisonnière"</i>
----------------------------	---

Description

A dataset containing Marcel Proust's "La prisonnière". This text has been downloaded from WikiSource.

Usage

```
laprisonniere
```

Format

A tibble with text, book, volume, and year

Source

`https://fr.wikisource.org/wiki/La_Prisonni`

lecotedeguermant	Marcel Proust's novel "Le côté de Guermantes"
------------------	---

Description

A dataset containing Marcel Proust's "À l'ombre des jeunes filles en fleurs". This text has been downloaded from WikiSource.

Usage

lecotedeguermant

Format

A tibble with text, book, volume, and year

Source

`https://fr.wikisource.org/wiki/Le_C`

letempretrouve	Marcel Proust's novel "Le temps retrouvé"
----------------	---

Description

A dataset containing Marcel Proust's "Le temps retrouvé". This text has been downloaded from WikiSource.

Usage

letempretrouve

Format

A tibble with text, book, volume, and year.

Source

`https://fr.wikisource.org/wiki/Le_Temps_retrouv`

proust_books*Tidy data frame of Marcel Proust's 7 novels from La Recherche*

Description

Returns a tidy tibble of Marcel Proust's 7 novels from *À la recherche du temps perdu*. The tibble contains four columns: `text`, `book`, `volume` and `year`.

Usage

```
proust_books()
```

Value

A tibble with four columns: `text`, `book`, `volume` and `year`.

Examples

```
#Create the tibble  
proust <- proust_books()
```

proust_char*Characters from "À la recherche du temps perdu"*

Description

A dataset containing Marcel Proust's characters from *"À la recherche du temps perdu"* and their frequency in each book. This dataset has been downloaded from proust-personnages.

Usage

```
proust_char
```

Format

A tibble with their name

Source

```
http://proust-personnages.fr/?page\_id=10254
```

<code>proust_characters</code>	<i>Characters from Proust Books</i>
--------------------------------	-------------------------------------

Description

Returns a tidy data frame of Marcel Proust's characters.

Usage

```
proust_characters()
```

Value

A tibble

Source

<http://proust-personnages.fr/>

Examples

```
#Creates the tibble
proust <- proust_characters()
```

<code>proust_random</code>	<i>Create a Random Proust extract</i>
----------------------------	---------------------------------------

Description

Create your own flavor of Proust with this random extractor.

Usage

```
proust_random(count = 1, collapse = TRUE)
```

Arguments

<code>count</code>	the number of line you want to randomly extract and paste.
<code>collapse</code>	if FALSE, the output will be a tibble. Default is TRUE, a character vector.

Value

a character vector

Examples

```
proust_random(4)
```

<code>proust_sentiments</code>	<i>Old sentiment lexicon This function has been deprecated, and will be in next proustr version. See the rfeel package now: http://github.com/ColinFay/rfeel</i>
--------------------------------	---

Description

Old sentiment lexicon This function has been deprecated, and will be in next proustr version. See the rfeel package now: <http://github.com/ColinFay/rfeel>

Usage

```
proust_sentiments(type = c("polarity", "score"))
```

Arguments

<code>type</code>	For backward compatibility
-------------------	----------------------------

Value

a tibble

<code>proust_stopwords</code>	<i>Stop Words</i>
-------------------------------	-------------------

Description

Stop words concatenated from various web sources.

Usage

```
proust_stopwords()
```

Value

a tibble with stopwords

Source

```
https://raw.githubusercontent.com/stopwords-iso/stopwords-fr/master/stopwords-fr.txt
```

Examples

```
proust_stopwords()
```

pr_detect_days	<i>Detect french days</i>
----------------	---------------------------

Description

Detect the name of the days (in French)

Usage

```
pr_detect_days(df, col)
```

Arguments

df	a dataframe
col	the column containing the text

Value

a tibble with the number of days detected by the algo

Examples

```
a <- data.frame(jours = c("C'est lundi 1er mars et mardi 2",  
  "Et mercredi 3", "Il est revenu jeudi."))  
pr_detect_days(a, jours)
```

pr_detect_months	<i>Detect french months</i>
------------------	-----------------------------

Description

Detect the name of the months (in French)

Usage

```
pr_detect_months(df, col)
```

Arguments

df	a dataframe
col	the column containing the text

Value

a tibble with the number of days detected by the algo

Examples

```
a <- data.frame(month = c("C'est lundi 1er mars et mardi 2",  
  "Et mercredi 3", "Il est revenu en juin."))  
pr_detect_months(a, month)
```

pr_detect_pro	<i>Detect French pronouns</i>
---------------	-------------------------------

Description

Detect the pronouns from a text (in French)

Usage

```
pr_detect_pro(df, col, verbose = FALSE)
```

Arguments

df	a dataframe
col	the column containing the text
verbose	wether or not to return the list of pronouns. Defaults is FALSE

Details

The shortcuts in the pronoun col stand for:

pps: first person singular (première personne du singulier)

dps: second person singular (deuxième personne du singulier)

tps: third person singular (troisième personne du singulier)

ppp: first person plural (première personne du pluriel)

dpp: second person singular (deuxième personne du pluriel)

tpp: third person singular (troisième personne du pluriel)

Value

a tibble with the detected pronouns

Examples

```
library(proustr)
a <- proust_books()[1,]
pr_detect_pro(a, text, verbose = TRUE)
pr_detect_pro(a, text)
```

<code>pr_keep_only_alnum</code>	<i>Remove non alnum elements</i>
---------------------------------	----------------------------------

Description

Remove non alnum elements

Usage

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

Arguments

<code>text</code>	a vector
<code>replacement</code>	what to replace the non alnum with. Default is " ".

Value

a vector

Examples

```
pr_keep_only_alnum("neuilly-en-thelle")
```

<code>pr_normalize_punc</code>	<i>Normalize punctuation</i>
--------------------------------	------------------------------

Description

Normalize a text written with usual french punctuation

Usage

```
pr_normalize_punc(df, col)
```

Arguments

<code>df</code>	a dataframe
<code>col</code>	the column to normalize

Value

a tibble with normalized text

Examples

```
a <- proustr::albertinedisparue[1:20,]
pr_normalize_punc(albertinedisparue, text)
```

pr_stem_sentences	<i>Stem a dataframe containing a column with sentences</i>
-------------------	--

Description

Implementation of the SnowballC stemmer. Note that punctuation and capital letters are removed when processing.

Usage

```
pr_stem_sentences(df, col, language = "french")
```

Arguments

df	the data.frame containing the text
col	the column with the text
language	the language of the text. Default is french. See SnowballC::getStemLanguages() function for a list of supported languages.

Value

a tibble

Examples

```
a <- proustr::laprisonniere[1:10,]  
pr_stem_sentences(a, text)
```

pr_stem_words	<i>Stem a dataframe containing a column with words</i>
---------------	--

Description

Implementation of the SnowballC stemmer. Note that punctuation and capitals letters are also removed.

Usage

```
pr_stem_words(df, col, language = "french")
```

Arguments

df	the data.frame containing the sentences
col	the column with the sentences
language	the language of the words Default is french. See SnowballC::getStemLanguages() function for a list of supported languages.

Value

a tibble

Examples

```
a <- data.frame(words = c("matin", "heure", "fatigué", "sonné", "lois", "tests", "fusionner"))
pr_stem_words(a, words)
```

<code>pr_unacent</code>	<i>Remove accents</i>
-------------------------	-----------------------

Description

Remove accents from a character vector

Usage

```
pr_unacent(text)
```

Arguments

`text` a vector

Value

a vector

Examples

```
pr_unacent("du chêne")
```

<code>sodomeetgomorrhe</code>	<i>Marcel Proust's novel "Sodome et Gomorrhe"</i>
-------------------------------	---

Description

A dataset containing Marcel Proust's "Sodom et Gomorrhe". This text has been downloaded from WikiSource.

Usage

```
sodomeetgomorrhe
```

Format

A tibble with text, book, volume, and year

Source

```
https://fr.wikisource.org/wiki/Sodome_et_Gomorrhe
```

stop_words	<i>Stopwords</i>
------------	------------------

Description

ISO stopwords

Usage

stop_words

Format

A tibble

Source

<https://raw.githubusercontent.com/stopwords-iso/stopwords-iso/master/stopwords-iso.json>