Package ‘malaytextr’

January 17, 2023

Title  Text Mining for Bahasa Malaysia

Version  0.1.3

Description  It is designed to work with text written in Bahasa Malaysia. We provide
functions and data sets that will make working with Bahasa Malaysia text much easier.
For word stemming in particular, we will look up the Malay words in a dictionary and
then proceed to remove “extra suffix” as explained in Khan, Rehman Ullah,
Fitri Suraya Mohamad, Muh Inam UlHaq, Shahren Ahmad Zadi Adruce,
Philip Nuli Anding, Sajjad Nawaz Khan, and Abdulrazak Yahya Saleh Al-Hababi
(2017) <https://ijrest.net/vol-4-issue-12.html>. This package includes
a dictionary of Malay words that may be used to perform word stemming,
a dataset of Malay stop words, a dataset of sentiment words
and a dataset of normalized words.

License  MIT + file LICENSE

Encoding  UTF-8

LazyData  true

RoxygenNote  7.2.3

URL  https://github.com/zahiernasrudin/malaytextr

BugReports  https://github.com/zahiernasrudin/malaytextr/issues

Imports  dplyr, magrittr, rlang, stringr

Depends  R (>= 2.10)

Suggests  rmarkdown, knitr, testthat (>= 3.0.0)

VignetteBuilder  knitr

Config тесттthat/edition  3

NeedsCompilation  no

Author  Zahier Nasrudin [aut, cre] (<https://orcid.org/0000-0002-7060-776X>)

Maintainer  Zahier Nasrudin <zahiernasrudin@gmail.com>

Repository  CRAN

Date/Publication  2023-01-17 11:50:02 UTC
R topics documented:

malayrootwords .................................................. 2
malaystopwords .................................................. 2
normalized ......................................................... 3
remove_url .......................................................... 3
sentiment_general .................................................. 4
stem_malay .......................................................... 4

Index

malayrootwords Data of Malay root words

Description
Data of Malay root words

Usage
malayrootwords

Format
A tibble with 4295 rows and 2 variables:
Col Word  dbl Malay Word
Root Word  dbl Malay Root Word

malaystopwords Data of Malay stop words

Description
Data of Malay stop words

Usage
malaystopwords

Format
A tibble with 512 rows and 1 variable:
stopwords  dbl Malay stop words
Data of Malay root words

Description
Data of Malay root words

Usage
normalized

Format
A tibble with 153 rows and 2 variables:

<table>
<thead>
<tr>
<th>Col Word</th>
<th>dbl Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normalized Word</td>
<td>dbl Normalized Word</td>
</tr>
</tbody>
</table>

remove_url
Remove URL links

Description
Remove URL links

Usage
remove_url(string)

Arguments

<table>
<thead>
<tr>
<th>string</th>
</tr>
</thead>
<tbody>
<tr>
<td>String to change</td>
</tr>
</tbody>
</table>

Details
remove_url() is an approach to remove link(s) from a string

Value
Returns a string with URL links removed

Examples
x <- c("test https://t.co/fkQC2dXwmc", "another one https://www.google.com/ to try")
remove_url(x)
sentiment_general  

*Data of Sentiment Words (Positive or Negative)*

**Description**

Data of Sentiment Words (Positive or Negative)

**Usage**

`sentiment_general`

**Format**

A tibble with 1424 rows and 2 variables:

- Word  dbl  Sentiment Word
- Root  Word  dbl  Sentiment

---

stem_malay  

*Stemming Malay words*

**Description**

Malaytextr function to stem Malay words

**Usage**

`stem_malay(word, dictionary, col_feature1, col_dict1, col_dict2, Word)`

**Arguments**

- `word`  A data frame, or a character vector
- `dictionary`  A data frame with a column of words to be stemmed and a column of root words
- `col_feature1`  Column that contains words to be stemmed from `word`
- `col_dict1`  Column that will be used to match with `col_feature1` from `word`
- `col_dict2`  Column that contains the root words from `dictionary`
- `Word`  Deprecated. Please use `word` instead
stem_malay

Format
An object of class function of length 1.

Details
stem_malay() is an approach to find the Malay words in a dictionary and then proceed to remove "extra suffix" as explained by Khan et al. (2017), and then "prefix" and lastly, "suffix".

Value
Returns a data frame with the following properties:
• Col Word: Renamed input from word
• Root Word: An additional column which contains the word(s) after being stemmed.

References

Examples

#Specifying a character vector &
#use a dictionary from malaytextr package
stem_malay(word = "banyaknya", dictionary = malayrootwords)

#A data frame,
#Use a dictionary from malaytextr package,
#With a dataframe, you will need to specify the column to be stemmed
x <- data.frame(text = c("banyaknya", "sangat", "terkedu", "pengetahuan"))
stem_malay(word = x, dictionary = malayrootwords, col_feature1 = "text")
Index

* datasets
  malayrootwords, 2
  malaystopwords, 2
  normalized, 3
  sentiment_general, 4
  stem_malay, 4

malayrootwords, 2
malaystopwords, 2
normalized, 3
remove_url, 3
sentiment_general, 4
stem_malay, 4