Package: malaytextr (via r-universe)

September 16, 2024

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

Roxygen list(markdown = 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/testthat/edition 3

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

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

RemoteRef HEAD

RemoteSha 217e687dff84381c5191ec344e15827a317f8079

Contents

malayrootwords 2
malaysia_politic_sentiment 2
malaystopwords 3
normalized 3
remove_url 4
sentiment_general 4
stem_malay 5

Index **7**

`malayrootwords` *Data of Malay root words*

Description

Data of Malay root words

Usage

```
malayrootwords
```

Format

A tibble with 4310 rows and 2 variables:

`Col Word` dbl Malay Word

`Root Word` dbl Malay Root Word

`malaysia_politic_sentiment` *Malaysia Politic Tweets Sentiment Dataset (Positive, Negative or Neutral)*

Description

Malaysia Politic Tweets Sentiment Dataset (Positive, Negative or Neutral)

Usage

```
malaysia_politic_sentiment
```

Format

A tibble with 71 rows and 3 variables:

`id` dbl Represents a unique identifier assigned to each tweet

`text` dbl Tweet related to Malaysia politics

`Sentiment` dbl The sentiment classification assigned to each tweet

`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

`normalized` *Data of Malay normalized words*

Description

Data of Malay normalized words

Usage

```
normalized
```

Format

A tibble with 153 rows and 2 variables:

`Col Word` dbl Word

`Normalized Word` dbl Normalized Word

remove_url *Remove URL links*

Description

Remove URL links

Usage

```
remove_url(string)
```

Arguments

string	String to change

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/fkQC2dXwnc", "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 1428 rows and 2 variables:

`Word` dbl Sentiment Word

`Sentiment` 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	Depreciated. Please use word instead

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

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. "Malay Language Stemmer."

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
malaysia_politic_sentiment, 2
malaystopwords, 3
normalized, 3
sentiment_general, 4
stem_malay, 5

malayrootwords, 2
malaysia_politic_sentiment, 2
malaystopwords, 3

normalized, 3

remove_url, 4

sentiment_general, 4
stem_malay, 5