Package 'PersianStemmer'

June 28, 2019

Type Package

Title Persian Stemmer for Text Analysis

Version 1.0

Date 2019-06-20

Author Roozbeh Safshekan and Rich Nielsen

Maintainer Roozbeh Safshekan <rse@mit.edu>

Description Allows users to stem Persian texts for text analysis.

License GPL (>= 2)

NeedsCompilation no

Repository CRAN

Date/Publication 2019-06-28 16:00:03 UTC

R topics documented:

PersianStemmer-package *A package for stemming Persian for text analysis.*

Description

This package is a Persian Stemmer for Text Analysis.

Details

Use the PerStem function.

Author(s)

Roozbeh Safshekan <rse@mit.edu> and Rich Nielsen <rnielsen@mit.edu>

See Also

PerStem

Examples

```
# Load data
data(UniversityofTehran)

# Stem and transliterate the text
PerStem(UniversityofTehran,NoEnglish=TRUE, NoNumbers= TRUE,
                    NoStopwords=TRUE, NoPunctuation= TRUE,
                    StemVerbs = TRUE, NoPreSuffix= TRUE, Context = TRUE,
                    StemBrokenPlurals=TRUE,Transliteration= TRUE)
```

FixBrokenPlurals *Stems Arabic broken plurals*

Description

Stems Arabic broken plurals and returns singulars.

Usage

```
FixBrokenPlurals(texts)
```

Arguments

texts	A string with Arabic broken plurals that should be stemmed.

Value

FixBrokenPlurals returns a string with Arabic broken plurals stemmed.

Author(s)

Safshekan, Nielsen

Examples

```
# Create string with Arabic broken plurals
x <- '\u0645\u0635\u0627\u062F\u06CC\u0642
\u0648\u0632\u0631\u0627
\u062D\u062F\u0648\u062F'

# Remove new line characters and fixe half-spaces from a string.
x <- RemNewlineHalfspace(x)

# Remove all characters that are not Latin, Persian or punctuation,
# and standardize Persian characters.
x <- RefineChars(x)

# Stem Arabic broken plurals
FixBrokenPlurals(x)
```

FixVerbs *Stemms verbs*

Description

Stems verbs and returns past and present roots.

Usage

```
FixVerbs(texts, Context)
```

Arguments

texts	A Persian string in unicode.
Context	If TRUE, the function stems past-root+'he' only if other verbs with the same past-root exist in text. If FALSE, the function stems verbs without considering other words in text.

Value

FixVerbs returns a string with verbs stemmed.

Author(s)

Safshekan, Nielsen

Examples

```
# Create string with Persian verbs
x <- '\u0646\u0648\u0634\u062A\u0647 \u0634\u062F\u0647
\u0628\u0648\u062F\u0647 \u0627\u0633\u062A - \u0646\u0648\u0634\u062A\u0645 -
\u062F\u0627\u0631\u06CC\u0645 \u0645\u06CC\u0631\u0648\u06CC\u0645 -
\u062E\u0648\u0627\u0646\u062F\u0647 \u0645\u06CC\u0634\u0648\u0646\u062F -
\u062E\u0648\u0627\u0647\u062F \u06AF\u0641\u062A -
\u0628\u0631\u062F\u0647 \u0627\u0633\u062A -
\u0645\u06CC\u06AF\u0648\u06CC\u06CC\u0645'

# Remove new line characters and fixe half-spaces from a string.
x <- RemNewlineHalfspace(x)

# Remove all characters that are not Latin, Persian or punctuation,
# and standardize Persian characters.
x <- RefineChars(x)

# Stems verbs
y <- FixVerbs(x, Context = TRUE)
z <- FixVerbs(x, Context = FALSE)

# Remove the numeric signifiers which are used in PerStem function.
gsub("0|1|2|3|4|5","",y)
gsub("0|1|2|3|4|5","",z)
```

`PerStem` *Persian Stemmer for Text Analysis*

Description

Stems Persian texts for text analysis.

Usage

```
PerStem(dat, NoEnglish = TRUE, NoNumbers = TRUE,
NoStopwords = TRUE, NoPunctuation = TRUE,
StemVerbs = TRUE, NoPreSuffix = TRUE,
Context = TRUE, StemBrokenPlurals = TRUE,
Transliteration = TRUE)
```

Arguments

`dat`	The original data.
`NoEnglish`	Removes English characters.
`NoNumbers`	Removes numbers.
`NoStopwords`	Removes stopwords by using the default stopword list.
`NoPunctuation`	If TRUE the function removes punctuation. If FALSE, it fixes punctuation for text analysis.

`StemVerbs` Performs stemming on verbs and returns past or present root of the verb.

`NoPreSuffix` Performs stemming by removing prefixes and suffixes.

`Context` If TRUE, the function performs stemming on a word only if its stem exists in text. If FALSE, the function performs stemming without considering other words in text.

`StemBrokenPlurals`
Performs stemming on Arabic broken plurals and return singulars by using the default Arabic broken plurals list.

`Transliteration`
Transliterates Persian unicode characters into Latin characters using a transliteration system developed by Roozbeh Safshekan and Rich Nielsen.

Details

`PerStem` prepares texts in Persian for text analysis by stemming.

Value

`PerStem` returns the stemmed Persian text.

Author(s)

Roozbeh Safshekan, Richard Nielsen

Examples

```
# Load data
data(UniversityofTehran)

# Stem and transliterate the text
PerStem(UniversityofTehran,NoEnglish=TRUE, NoNumbers= TRUE,
                    NoStopwords=TRUE, NoPunctuation= TRUE,
                    StemVerbs = TRUE, NoPreSuffix= TRUE, Context = TRUE,
                    StemBrokenPlurals=TRUE,Transliteration= TRUE)
```

RefineChars *Removes all characters that are not Latin, Persian or punctuation, and standardizes Persian characters.*

Description

Removes all unicode characters except Latin, Persian or General Punctuation characters and standardizes Persian characters.

Usage

```
RefineChars(texts)
```

Arguments

texts	A string from which all characters that are not Latin, Persian or punctuation should be removed, or in which Persian characters should be standardized.

Value

RefineChars returns a string with only Latin, standardized Persian or general punctuation characters.

Author(s)

Safshekan, Nielsen

Examples

```
# Create string with Latin, Persian, Japanese, non-standardized Persian and punctuation characters.
x <- '\u062F\u0627\u0646\u0634\u06AF\u0627\u0647\u064A \u060C
\u0641\u06CC\u0632\u06CC\u0643 university
\u65E5\u672C \u0664\u0665\u0666'

# Remove new line characters and fixe half-spaces from a string.
x <- RemNewlineHalfspace(x)

# Remove all characters that are not Latin, Persian or punctuation,
# and standardize Persian characters.
RefineChars(x)
```

RemNewlineHalfspace *Removes new line characters and fixes half-spaces*

Description

Removes new line characters and fixes half-spaces in a string.

Usage

```
RemNewlineHalfspace(texts)
```

Arguments

texts	A string which its new line characters and half-spaces should be removed or fixed.

Value

RemNewlineHalfspace returns a string with new line characters and half-spaces removed or fixed.

Author(s)

Safshekan, Nielsen

Examples

```
# Create string with Persian string with new line characters and half-spaces

x <- '\u062F\u0627\u0646\u0634\u06AF\u0627\u0647\u200C\u0647\u0627\u06CC
\u062A\u0647\u0631\u0627\u0646'

# Remove newline characters (eg.\n\r\t\f\v) and fix half-spaces
RemNewlineHalfspace(x)
```

RemoveEnglish *Remove English characters*

Description

Removes English characters from a string.

Usage

```
RemoveEnglish(texts)
```

Arguments

texts	A string from which English characters should be removed.

Value

RemoveEnglish returns a string with English characters removed.

Author(s)

Safshekan, Nielsen

Examples

```
# Create string with English characters
x <- '\u062F\u0627\u0646\u0634\u06AF\u0627\u0647 University'

# Remove English characters
RemoveEnglish(x)
```

RemoveNumbers *Remove numerals.*

Description

Removes numerals from a string.

Usage

```
RemoveNumbers(texts)
```

Arguments

texts	A string from which numerals should be removed.

Value

RemoveNumbers returns a string with numerals removed.

Author(s)

Safshekan, Nielsen

Examples

```
# Create string with Persian characters and number
x <- '\u0633\u0627\u0644 \u06F1\u06F3\u06F9\u06F8'

# Remove Numbers
RemoveNumbers(x)
```

RemovePreSuffix *Remove Persian prefixes and suffixes.*

Description

Removes Persian prefixes and suffixes from a unicode string using the default list of Persian prefixes and suffixes.

Usage

```
RemovePreSuffix(texts, Context)
```

Arguments

texts	A Persian string in unicode
Context	If TRUE, the function removes prefixes and suffixes of a word only if its stem exists in text. If FALSE, the function removes prefixes and suffixes without considering other words in text.

Value

RemovePreSuffix returns a string with Persian prefixes and suffixes removed.

Author(s)

Safshekan, Nielsen

Examples

```
# Create string with Persian characters
x <- '\u0627\u0628\u0631\u0642\u062F\u0631\u062A\u0647\u0627\u06CC\u06CC
\u06A9\u062A\u0627\u0628\u0647\u0627\u06CC\u0645 \u06A9\u062A\u0627\u0628'

# Remove new line characters and fixe half-spaces from a string.
x <- RemNewlineHalfspace(x)

# Remove all characters that are not Latin, Persian or punctuation,
# and standardize Persian characters.
x <- RefineChars(x)

# Remove Prefixes and Suffixes
RemovePreSuffix(x, Context = TRUE)
RemovePreSuffix(x, Context = FALSE)
```

RemoveStopwords *Remove Persian stop-words.*

Description

Defines a list of Persian stopwords and removes them from a string.

Usage

```
RemoveStopwords(texts)
```

Arguments

texts	A string from which Persian stopwords should be removed.

Value

RemoveStopwords returns a string with Persian stopwords removed.

Author(s)

Safshekan, Nielsen

Examples

```
# Create Persian string with stopwords
x <- '\u0627\u0632
\u062F\u0627\u0646\u0634\u06AF\u0627\u0647
\u0622\u0645\u062F'

# Remove new line characters and fixe half-spaces from a string.
x <- RemNewlineHalfspace(x)

# Remove all characters that are not Latin, Persian or punctuation,
# and standardize Persian characters.
x <- RefineChars(x)

# Remove stopwords
RemoveStopwords(x)
```

ReverseTransliterate *Transliterate Latin characters into Persian unicode characters*

Description

Transliterates Latin characters into Persian unicode characters using a transliteration system developed by Roozbeh Safshekan and Rich Nielsen.

Usage

```
ReverseTransliterate(texts)
```

Arguments

texts	A string in Latin characters to be transliterated into Persian characters.

Value

ReverseTransliterate returns a string in Persian characters.

Author(s)

Safshekan, Nielsen

Examples

```
# Create Latin string
x <- 'danWGah thran'

# Converts Latin characters into Persian unicode characters
ReverseTransliterate(x)
```

RFPunctuation *Remove or fix punctuation.*

Description

Removes punctuation characters or inserts spaces before and after them so that they can be used in text analysis as separate units.

Usage

```
RFPunctuation(texts, NoPunctuation)
```

Arguments

texts	A string with punctuation which should be removed or fixed.
NoPunctuation	If TRUE, the function removes punctuation. If FALSE, the function inserts spaces before and after punctuation.

Value

RFPunctuation returns a string with punctuation removed or fixed for text analysis.

Author(s)

Safshekan, Nielsen

Examples

```
# Create string with Persian characters and punctuation
x <- '\u062F\u0627\u0646\u0634\u06AF\u0627\u0647\u060C \u062A\u0647\u0631\u0627\u0646\u061F'

# Remove punctuation
RFPunctuation(x, NoPunctuation = TRUE)

# Fix punctuation
RFPunctuation(x, NoPunctuation = FALSE)
```

Transliterate *Transliterate Persian unicode characters into Latin characters*

Description

Transliterates Persian unicode characters into Latin characters using a transliteration system developed by Roozbeh Safshekan Rich Nielsen.

Usage

```
Transliterate(texts)
```

Arguments

texts	A string in Persian characters to be transliterated into Latin characters.

Value

Transliterate returns a string in Latin characters.

Author(s)

Safshekan, Nielsen

Examples

```
# Create Persian string
x <- '\u062F\u0627\u0646\u0634\u06AF\u0627\u0647 \u062A\u0647\u0631\u0627\u0646'

# Performs transliteration of Persian into Latin characters
Transliterate(x)
```

UniversityofTehran *Persian texts*

Description

Persian text from the University of Tehran website

Usage

```
data("UniversityofTehran")
```

Format

Persian text data

Source

https://ut.ac.ir/fa/page/200

Examples

```
# Load data
data(UniversityofTehran)
```

Index