

Package ‘openNLP’

September 20, 2011

Title openNLP Interface

Version 0.0-8

Date 2010-06-26

Author Ingo Feinerer, Kurt Hornik

Maintainer Kurt Hornik <Kurt.Hornik@R-project.org>

Imports methods, rJava (>= 0.6-3), tm

Enhances tm

Suggests openNLPmodels.en, openNLPmodels.es

SystemRequirements Java (>= 5.0)

Description An interface to openNLP (<http://opennlp.sourceforge.net/>), a collection of natural language processing tools including a sentence detector, tokenizer, pos-tagger, shallow and full syntactic parser, and named-entity detector, using the Maxent Java package for training and using maximum entropy models.

License LGPL-2.1

Encoding UTF-8

Repository CRAN

Date/Publication 2010-06-26 20:04:13

R topics documented:

sentDetect	2
tagPOS	3
tmSentDetect-methods	4
tmTagPOS-methods	4
tmTokenize-methods	5
tokenize	5

Index	7
--------------	----------

sentDetect	<i>Detect sentences</i>
------------	-------------------------

Description

Detect sentences.

Usage

```
sentDetect(s, language = "en", model = NULL)
```

Arguments

s	A character vector with texts from which sentences should be detected.
language	A character string giving the language of s. This argument is only used if model is NULL for selecting a default model. At the moment, languages 'en' (English), 'es' (Spanish), 'de' (German) and 'th' (Thai) are supported, provided that the corresponding openNLP model language packages (openNLPmodels.en , ...) are available.
model	A model.

Details

If model is NULL then a default model for sentence detection is loaded from the corresponding openNLP models language package.

Value

A character vector with all the sentences found in the elements of s.

Author(s)

Ingo Feinerer

References

OpenNLP <http://opennlp.sourceforge.net/>

Examples

```
s <- "This is a sentence. This another---but with dash-like structures, and some commas.  
Maybe another with question marks? Sure!"  
sentDetect(s, language = "en")  
s <- "¿Como se llama usted? El castellano es la lengua española oficial del Estado."  
sentDetect(s, language = "es")
```

tagPOS	<i>Tag part-of-speech</i>
--------	---------------------------

Description

Tag part-of-speech.

Usage

```
tagPOS(sentence, language = "en", model = NULL, tagdict = NULL)
```

Arguments

sentence	A character vector with sentences to get POS-tagged.
model	A model.
language	A character string giving the language of sentence. This argument is only used if model is NULL for selecting a default model. At the moment, languages 'en' (English), 'es' (Spanish), 'de' (German) and 'th' (Thai) are supported, provided that the corresponding openNLP model language packages (openNLP-models.en , ...) are available.
tagdict	A tag dictionary.

Details

If model is NULL then a default model for sentence detection is loaded from the corresponding openNLP models language package. If tagdict is NULL a default English tag dictionary is taken.

Value

A character vector holding POS-tagged sentence.

Author(s)

Ingo Feinerer

References

OpenNLP <http://opennlp.sourceforge.net/>

Examples

```
sentence <- "This is a short sentence consisting of some nouns, verbs, and adjectives."  
tagPOS(sentence, language = "en")  
s <- "el castellano es la lengua española oficial del Estado."  
tagPOS(s, language = "es")
```

tmSentDetect-methods *Methods for Function tmSentDetect in Package 'openNLP'*

Description

Methods for function tmSentDetect in package **openNLP**.

Methods

object = "PlainTextDocument" Detect sentences in object and return the object.

Examples

```
if(!require("tm")) stop("could not load tm package")
data("crude")
crude[[1L]]
tmSentDetect(crude[[1L]])
```

tmTagPOS-methods *Methods for Function tmTagPOS in Package 'openNLP'*

Description

Methods for function tmTagPOS in package **openNLP**.

Methods

object = "PlainTextDocument" Tag part-of-speech in object and return the object.

Examples

```
if(!require("tm")) stop("could not load tm package")
data("crude")
crude[[1L]]
tmTagPOS(crude[[1L]])
```

tmTokenize-methods *Methods for Function tmTokenize in Package 'openNLP'*

Description

Methods for function tmTokenize in package **openNLP**.

Methods

object = "PlainTextDocument" Tokenize object and return the object.

Examples

```
if(!require("tm")) stop("could not load tm package")
data("crude")
crude[[1L]]
tmTokenize(crude[[1L]])
```

tokenize *Tokenizer*

Description

Tokenizes the input.

Usage

```
tokenize(s, language = "en", model = NULL)
```

Arguments

s	A character vector of texts to be tokenized.
language	A character string giving the language of s. This argument is only used if model is NULL for selecting a default model. At the moment, languages 'en' (English), 'es' (Spanish), 'de' (German) and 'th' (Thai) are supported, provided that the corresponding openNLP model language packages (openNLPmodels.en , ...) are available.
model	A model.

Details

If model is NULL then a default model for sentence detection is loaded from the corresponding openNLP models language package.

Value

A character vector with all the tokens found in the elements of `s`.

Author(s)

Ingo Feinerer

References

OpenNLP <http://opennlp.sourceforge.net/>

Examples

```
s <- "This is a sentence."  
tokenize(s, language = "en")  
s <- "¿Como se llama usted? El castellano es la lengua española oficial  
del Estado."  
tokenize(s, language = "es")
```

Index

*Topic **file**

sentDetect, [2](#)

tagPOS, [3](#)

tokenize, [5](#)

*Topic **methods**

tmSentDetect-methods, [4](#)

tmTagPOS-methods, [4](#)

tmTokenize-methods, [5](#)

sentDetect, [2](#)

tagPOS, [3](#)

tmSentDetect (tmSentDetect-methods), [4](#)

tmSentDetect, PlainTextDocument-method
(tmSentDetect-methods), [4](#)

tmSentDetect-methods, [4](#)

tmTagPOS (tmTagPOS-methods), [4](#)

tmTagPOS, PlainTextDocument-method
(tmTagPOS-methods), [4](#)

tmTagPOS-methods, [4](#)

tmTokenize (tmTokenize-methods), [5](#)

tmTokenize, PlainTextDocument-method
(tmTokenize-methods), [5](#)

tmTokenize-methods, [5](#)

tokenize, [5](#)