Package 'textometry'

February 20, 2015

Type Package

Title Textual Data Analysis Package used by the TXM Software

Version 0.1.4

Date 2015-01-08

Author Sylvain Loiseau, Lise Vaudor, Matthieu Decorde, Serge Heiden

Maintainer Matthieu Decorde <matthieu.decorde@ens-lyon.fr>

Description Statistical exploration of textual corpora using several methods
from French 'Textometrie' (new name of 'Lexicometrie') and French 'Data Analysis' schools.
It includes methods for exploring irregularity of distribution of lexicon features across
text sets or parts of texts (Specificity analysis); multi-
dimensional exploration (Factorial analysis), etc.
Those methods are used in the TXM software.

License GPL (>= 3)

Depends R (>= 1.5.0)

NeedsCompilation no

Repository CRAN

Date/Publication 2015-01-09 21:55:58

R topics documented:

textometry-package *Textual Data Analysis Package used by the TXM Software*

Description

Statistical exploration of textual corpora using several methods from French 'Textometrie' (new name of 'Lexicometrie') and French 'Data Analysis' schools. It includes methods for exploring irregularity of distribution of lexicon features across text sets or parts of texts (Specificity analysis); multi-dimensional exploration (Factorial analysis), etc. Those methods are used in the TXM software.

Details

Package:	textometry
Type:	Package
Version:	0.1.3
Date:	2014-06-16
License:	GPLv3
Depends:	R (>= 1.5.0)

Index:

```
specificities    Compute Lexical Specificity of subcorpus
progression    Draw progression graphic
```

Author(s)

Sylvain Loiseau, Lise Vaudor, Matthieu Decorde, Lise Vaudor

Examples

```
data(robespierre);
specificities(robespierre);
```

bfm *adverbs frequency from 5 different domains of the BFM database*

Description

A lexical table containing frequencies of adverbs from the BFM (Base de Francais m\'edi\'eval) database in 5 different domains (literary, historical, didactic, law, religious).

Usage

```
data(bfm)
```

Format

The format is: num [1:2, 1:5] 103000 1370887 23429 413441 15345 ... - attr(*, "dimnames")=List of 2 ..$: chr [1:2] "ADV" "other" ..$: chr [1:5] "literary" "history" "didactic" "juridical" ...

Details

The last line of the table gives the total frequency of all the other part of speech words in each of these domains.

Source

BFM: http://bfm.ens-lyon.fr

References

BFM - Base de Fran\,cais M\'edi\'eval [En ligne]. Lyon : ENS de Lyon, Laboratoire ICAR, 2012, http://bfm.ens-lyon.fr.

---

progression *Draw progression graphic*

Description

Draw the progression graphic of matches of CQL queries in a corpus

Usage

```
progression(positions, names, colors, styles, widths, corpusname, Xmin, T,
 doCumulative, structurepositions, strutnames, graphtitle, bande)
```

Arguments

positions	Vector containing corpus positions of CQL queries matches. A position is an integer from 0 (begining of corpus) to N (end of corpus)
names	String vector containing the CQL queries
colors	Vector containing the line color of each query
styles	Vector containing the line style of each query
widths	Vector containing the line width of each query
corpusname	String: corpus name
Xmin	Integer: corpus starting position of abscissa values
T	Integer: size of the corpus

doCumulative	Boolean: if true draw a cumulative graph, if false draw a density graph
structurepositions	optional Vector containing the structure positions of the corpus
strutnames	optional Vector containing the structures labels to display
graphtitle	String: graph title
bande	Float: density window size factor

Author(s)

Matthieu Decorde

`robespierre` *5 words from Robespierre's discourses*

Description

A lexical table containing frequencies of 5 words from 9 different public discourses of French politician Robespierre (between november 1793 and july 1794).

Usage

```
data(robespierre)
```

Format

The format is: num [1:6, 1:10] 464 45 35 30 6 ... - attr(*, "dimnames")=List of 2 ..$: chr [1:6] "de" "peuple" "republique" "ennemi" ... ..$: chr [1:10] "D1" "D2" "D3" "D4" ...

Details

The last line of the table gives the total frequency of all the other forms in each of these discourses.

Source

Lafon P. (1980) Sur la variabilit\'e de la fr\'e quence des formes dans un corpus, Mots, 1, pp. 127–165.

References

Lafon P. (1980) Sur la variabilit\'e de la fr\'e quence des formes dans un corpus, Mots, 1, pp. 127–165.

Examples

```
data(robespierre)

## See graphic in Lafon, 1980 - page 140

t <- colSums(robespierre)["D9"];     # size of the part
T <- sum(robespierre);               # size of the corpus
f <- rowSums(robespierre)["peuple"]; # total frequency of "peuple"
p <- dhyper(1:15, f, T-f, t)
title <- "Probability of each frequency of 'peuple' in the 'D9' discourse from 1 to 15"
plot(p, type="h", main=title, xlab="k", ylab="Prob(k)");
```

`specificities` *Calculate Lexical Specificity Score*

Description

Calculate the specificity - or association or surprise - score of a word being present `f` times or more in a sub-corpus of `t` words given that it appears a total of `F` times in a whole corpus of `T` words.

Usage

```
specificities(lexicaltable, types=NULL, parts=NULL)
```

Arguments

`lexicaltable`	a complete lexical table, i.e. a numeric matrix where each line represents a word and each column a part of the corpus. Each cell gives the frequency of the given word in the corresponding part of the corpus.
`types`	list of rows (words) for which the specificity score must be calculated. If `NULL`, the specificity score is calculated for every row; If `types` is a character vector, it indicates the row names for which the specificity score is to be calculated (an error is thrown if `lexicaltable` has no row names); If `types` is an integer vector, it indicates the index of rows for which the specificity score is to be calculated.
`parts`	list of columns (parts) for which the specificity score must be calculated. If `NULL`, the specificity index is calculated for every part; If `parts` is a character vector, it indicates the column names for which the specificity score is to be calculated (an error is thrown if `lexicaltable` has no column names); If `parts` is an integer vector, it indicates the index of columns for which the specificity score is to be calculated.

Value

Returns a matrix of `nrow(lexicaltable) * ncol(lexicaltable)` (the number of rows and columns may be reduced using `types` or `parts`), each cell giving the specificity score.

Author(s)

Matthieu Decorde, Serge Heiden, Sylvain Loiseau, Lise Vaudor

References

Lafon P. (1980) Sur la variabilit\'e de la fr\'e quence des formes dans un corpus, Mots, 1, pp. 127–165. http://www.persee.fr/web/revues/home/prescript/article/mots_0243-6450_1980_num_1_1_1008

See Also

specificities.probabilities, specificities.lexicon

Examples

```
data(robespierre);
spe <- specificities(robespierre);
string <- paste("The word %s appears f=%d times in a sub-corpus of t=%d words,",
" given a total frequency of F=%d in the robespierre corpus made",
" of T=%d words. The corresponding specificity score is %f", sep="");
print(sprintf(string,
'peuple',
robespierre['peuple','D4'],
colSums(robespierre)['D4'],
rowSums(robespierre)['peuple'],
sum(robespierre),
spe['peuple', 'D4']));
```

---

specificities.distribution.plot

Display specificities probability

Description

Display specificities probability distribution (call dhyper and specificities.probabilities.vector)

Usage

```
specificities.distribution.plot(x, F, t, T)
```

Arguments

x	observed number of A words
F	total number of A
t	size of part
T	size of corpus

Value

nothing

Author(s)

Matthieu Decorde, Serge Heiden

`specificities.lexicon` **OBSOLETE FUNCTION (see 'specificities.lexicon.new')* specificities association score with two frequency lists.*

Description

Compute specificities association score between a lexicon and a sub-lexicon

Usage

```
specificities.lexicon(lexicon, sublexicon)
```

Arguments

`lexicon`	a frequency list (named vector)
`sublexicon`	a frequency list (named vector)

Value

specificities index as a named vector.

See Also

`specificities` for specificities score and references

`specificities.lexicon.new` *specificities association score with two frequency list.*

Description

Compute specificities association score between a lexicon and a sub-lexicon. A new version of the "specificities.lexicon" function

Usage

```
specificities.lexicon.new(lexicon, sublexicon)
```

Arguments

lexicon	a frequency list (named vector)
sublexicon	a frequency list (named vector)

Value

specificities index as a named vector.

See Also

specificities for specificities score and references

---

specificities.probabilities

Calculate specificity probabilities

Description

Utility function computing specificity probabilities for the specificities function.

Usage

```
specificities.probabilities(lexicaltable, types = NULL, parts = NULL)
```

Arguments

lexicaltable	see specificities
types	see specificities
parts	see specificities

Value

Returns a matrix of signed specificity probabilities (between -1.0 and 1.0). By convention:

sign	The sign indicates if the observed frequency is lower (minus) or higher (plus) than the mode of the specificity model
.Machine$double.xmin limit	-10.0 and 10.0 values are used to hold the sign when the zero/.Machine$double.xmin boundary line has been crossed (the phyper function always returns 0.0)

See Also

see specificities.

`specificities.probabilities.vector` *Vector raw hypergeometric probabilities*

Description

Calculate specificity probabilities on vector (call phyper and phyper_right)

Usage

```
specificities.probabilities.vector(v_f, v_F, T, t)
```

Arguments

`v_f`	vector of lexicon ferquencies
`v_F`	vector of corpus frequencies
`T`	corpus size
`t`	sub-corpus size

Value

Hypergeometric probabilities. See `specificities.lexicon`.

Author(s)

Matthieu Decorde, Serge Heiden

Index