Package 'Xplortext'

November 6, 2024

Type Package

Encoding UTF-8

Title Statistical Analysis of Textual Data

Version 1.5.4

Date 2024-11-06

Maintainer Ramón Alvarez-Esteban <ramon.alvarez@unileon.es>

Author Ramón Alvarez-Esteban [aut, cre] (<https://orcid.org/0000-0002-4751-2797>),
Mónica Bécue-Bertaut [aut] (<https://orcid.org/0000-0002-6027-3655>),
Josep-Anton Sánchez-Espigares [ctb] (<https://orcid.org/0000-0001-8195-1913>),
Belchin Adriyanov Kostov [ctb] (<https://orcid.org/0000-0002-2126-3892>)

Description Provides a set of functions devoted to multivariate exploratory statistics on textual data. Classical methods such as correspondence analysis and agglomerative hierarchical clustering are available. Chronologically constrained agglomerative hierarchical clustering enriched with labelled-by-words trees is offered. Given a division of the corpus into parts, their characteristic words and documents are identified. Further, accessing to 'FactoMineR' functions is very easy. Two of them are relevant in textual domain. MFA() addresses multiple lexical table allowing applications such as dealing with multilingual corpora as well as simultaneously analyzing both open-ended and closed questions in surveys. See <http://xplortext.unileon.es> for examples.

License GPL (>=2.0)

Depends R (>= 4.4), FactoMineR(>= 2.11), ggplot2(>= 3.5.1), tm(>= 0.7-14)

Imports ape, cluster, dendextend, flexclust, flashClust, ggdendro, ggforce, ggpubr, ggrepel, graphics, gridExtra,
MASS, methods, patchwork, plotly, stringi, stringr, slam, stats, utils, vegan

URL https://xplortext.unileon.es

RoxygenNote 7.3.2

Contents

Xplortext-package   *Textual Analysis*

Description

Provides a set of functions devoted to multivariate exploratory statistics on textual data. Classical methods such as correspondence analysis and agglomerative hierarchical clustering are available. Chronologically constrained agglomerative hierarchical clustering enriched with labelled-by-words trees is offered. Given a division of the corpus into parts, their characteristic words and documents are identified. Further, accessing to 'FactoMineR' functions is very easy. Two of them are relevant in textual domain. MFA() addresses multiple lexical table allowing applications such as dealing with multilingual corpora as well as simultaneously analyzing both open-ended and closed questions in surveys. See https://xplortext.unileon.es for examples.

Details

Package:	Xplortext
Type:	Package
Version:	1.5.4
Date:	2024-11-06
License: GPL (>=2.0)	

Author(s)

Ramón Alvarez-Esteban
Maintainer: <ramon.alvarez@unileon.es>

References

Bécue, M. (2019). Textual Data Science with R. Chapman & Hall/CRC. doi:10.1201/9781315212661.

Husson F., Lê S., Pagès J. (2011). Exploratory Multivariate Analysis by Example Using R. Chapman & Hall/CRC. doi:10.1201/b10345.

Lebart, L., Salem, A., & Berry, L. (1998). Exploring textual data. (D. Kluwer, Ed.). doi:10.1007/9789401715256.

A website https://xplortext.unileon.es

---

ellipseLexCA	*Confidence ellipses on textual correspondence analysis graphs*

Description

Draws confidence ellipses around documents and/or words on a textual CA graph.

Usage

```
ellipseLexCA(object, selWord="ALL", selDoc="ALL", nbsample=100, level.conf=0.95,
    axes=c(1, 2), ncp=NULL, xlim=NULL, ylim=NULL, title=NULL, col.doc="blue",
    col.word="red", col.doc.ell=col.doc, col.word.ell=col.word, cex=1)
```

Arguments

object	object of LexCA class
selWord	selected words (indexes or names; by default "ALL"); see the details section
selDoc	selected docs (indexes or names; by default "ALL"); see the details section
nbsample	number of samples drawn to evaluate the stability of the points
level.conf	confidence level used to construct the ellipses (by default 0.95)
axes	length 2 vector specifying the dimensions to plot
ncp	maximum number of dimension to draw (by default NULL and ncp is the number of dimensions from LexCA object)
xlim	range for the plotted 'x' values, defaulting to the range of the finite values of 'x' (by default NULL)
ylim	range for the plotted 'y' values, defaulting to the range of the finite values of 'y' (by default NULL)
title	title of the graph (by default NULL and the title is automatically assigned)
col.doc	color for the documents-points (by default "blue")
col.word	color for words-points (by default "red")
col.doc.ell	color for the ellipses around documents-points (by default the same as col.doc)
col.word.ell	color for the ellipses around words-points (by default the same as col.word)
cex	text and symbol size is scaled by cex, in relation to size 1 (by default 1)

Details

The method "multinomial" is used to generate the replicated tables. So, the active lexical table contained in the LexCA object (active table) is taken as a reference.

Then, replicated lexical tables are generated by repeating nbsample times the following process: N (the sum of active table elements) values are drawn from a multinomial distribution with theoretical frequencies equal to the values in the active table cells divided by N. A replicated table is built from each drawing.

The nbsample documents-rows and/or words-columns of the replicated tables are projected as supplementary documents (rows) and/or supplementary words (columns) on the graph computed from the active lexical table. Then, confidence ellipses are drawn around each active element from the nbsample supplementary points.
The replicated samples with empty row-documents and/or word-columns with null frequency are dropped.
If over 10% of the total of replicated samples are dropped, the execution is stopped. Information is given through a stop-message.

The selDoc and selWord arguments allow for selecting the documents and/ or words.
The syntax for these arguments is similar to the one used in plot.LexCA.
However they only concern the active elements and selecting the characteristic words is not allowed.

Some examples follow: selDoc=c(1:5): the documents 1 to 5 are represented.
selDoc=c("doc1","doc5"): documents with labels doc1 or doc5 are represented.
selWord=c("word1","word3"): words with labels word1 or word3 are represented.
selDoc/selWord = "coord 10": the 10 documents/words with the highest coordinates on the 2 chosen axes are selected.
selDoc/selWord="contrib 10": documents/words with a contribution to the inertia of any of both axes over 10% of the axis inertia are selected.
selDoc/selWord="cos2 0.85: the documents/words with cos2 over 0.85 (as summed on the 2 axes) are selected.
selDoc ="meta 3": documents/words with a contribution over 3 times the average document/word contribution on any of both axes are selected.

Value

Returns a LexCA-like map representing the selected points and their confidence ellipses

Author(s)

Monica Bécue-Bertaut, Ramón Alvarez-Esteban <ramon.alvarez@unileon.es>, Josep-Antón Sánchez-Espigares

References

Husson F., Lê S., Pagès J. (2011). Exploratory Multivariate Analysis by Example Using R. Chapman & Hall/CRC. doi:10.1201/b10345.

Lebart, L., Piron, M., & Morineau, A. (2006). Statistique exploratoire multidimensionnelle. (Dunod, Ed.).

Lebart, L., Salem, A., & Berry, L. (1998). Exploring textual data. (Kluwer, Ed.). doi:10.1007/9789401715256.

See Also

LexCA, print.LexCA, plot.LexCA, summary.LexCA

Examples

```
## Not run:
data(open.question)
res.TD<-TextData(open.question, var.text=c(9,10), remov.number=TRUE, Fmin=10, Dmin=10,
  stop.word.tm=TRUE, context.quali=c("Gender","Age_Group","Education"),
  context.quanti=c("Age"))
res.LexCA<-LexCA(res.TD, graph=FALSE,ncp=8)
ellipseLexCA(res.LexCA, selWord="meta 1",selDoc=NULL, col.word="brown")
ellipseLexCA(res.LexCA, selWord="contrib 10",selDoc=NULL, col.word="brown")
ellipseLexCA(res.LexCA, selWord=c("work","job","money","comfortable"), selDoc=NULL,
  col.word="brown")
ellipseLexCA(res.LexCA, selWord="cos2 0.2", selDoc=NULL, col.word="brown")

## End(Not run)
## Not run:
data(open.question)
res.TD<-TextData(open.question, var.text=c(9,10), var.agg="Gen_Age", Fmin=10, Dmin=10,
  remov.number=TRUE, stop.word.tm=TRUE)
res.LexCA<-LexCA(res.TD, graph=FALSE)
ellipseLexCA(res.LexCA, selWord=NULL, col.doc="black")
ellipseLexCA(res.LexCA, selWord="meta 3", selDoc=NULL, col.word="brown")
ellipseLexCA(res.LexCA, selWord="contrib 10", selDoc=NULL, col.word="brown")
ellipseLexCA(res.LexCA, selWord=c("work","job","money","comfortable"), selDoc=NULL,
      col.word="brown")
ellipseLexCA(res.LexCA, selWord="cos2 0.2", selDoc=NULL, col.word="brown")

## End(Not run)
```

LabelTree   *Hierarchical words (LabelTree)*

Description

Extracts the hierarchical characteristic words associated to the nodes of a hierarchical tree; the characteristic words of each node are extracted, then each word is associated to the node that it best characterizes.

Usage

```
LabelTree(object, proba=0.05)
```

Arguments

object	object of LexHCca or LexCHCca class
proba	threshold on the p-value when the characteristic words are computed (by default 0.05)

Value

Returns a list including:

hierWord   list of the characteristic words associated to the nodes of a hierarchical tree; only
                  the non-empty nodes are included

Author(s)

Monica Bécue-Bertaut, Ramón Alvarez-Esteban <ramon.alvarez@unileon.es>, Josep-Anton Sánchez-
Espigares, Belchin Kostov

References

Bécue-Bertaut, M., Kostov, B., Morin, A., & Naro, G. (2014). Rhetorical Strategy in Forensic
Speeches: Multidimensional Statistics-Based Methodology. Journal of Classification,31,85-106.
doi:10.1007/s0035701491489.

Lebart, L., Salem, A., & Berry, L. (1998). Exploring textual data. (D. Kluwer, Ed.). doi:10.1007/
9789401715256.

See Also

LexCA, LexCHCca

Examples

```
data(open.question)
res.TD<-TextData(open.question,var.text=c(9,10), var.agg="Age_Group", Fmin=10, Dmin=10,
       stop.word.tm=TRUE)
 res.LexCA<-LexCA(res.TD, graph=FALSE)
 res.LexCHCca<-LexCHCca(res.LexCA, nb.clust=4, min=3)
 res.LabelTree<-LabelTree(res.LexCHCca)
```

LexCA   *Correspondence Analysis of a Lexical Table from a TextData object*
                             (LexCA)

Description

Performs Correspondence Analysis on the working lexical table contained in TextData object. Sup-
plementary documents, words, segments, contextual quantitative and qualitative variables can be
considered if previously selected in TextData function.

Usage

```
LexCA(object, ncp=5, context.sup="ALL", doc.sup=NULL, word.sup=NULL,
  segment=FALSE, graph=TRUE, axes=c(1, 2), lmd=3, lmw=3)
```

Arguments

`object`	object of TextData class
`ncp`	number of dimensions kept in the results (by default 5)
`context.sup`	column index(es) or name(s) of the contextual qualitative or quantitative variables among those selected in TextData function (by default "ALL")
`doc.sup`	vector indicating the index(es) or name(s) of the supplementary documents (rows) (by default NULL)
`word.sup`	vector indicating the index(es) or name(s) of the supplementary words (columns) (by default NULL)
`segment`	if TRUE, the repeated segments identified by TextData function will be considered as supplementary columns (by default FALSE)
`graph`	if TRUE, basic graphs are displayed; use plot.LexCA to obtain more graphs (by default TRUE)
`axes`	length-2 vector indicating the axes to plot (by default axes=c(1,2))
`lmd`	only the documents whose contribution is over lmd times the average-document-contribution are plotted (by default lmd=3)
`lmw`	only the words whose contribution is over lmw times the average-word-contribution are plotted (by default lmw=3)

Details

In the case of a direct CA, DocTerm is a non-aggregate table and:

1. the contextual quantitative variables are considered as supplementary quantitative columns in CA.
2. the categories of the contextual qualitative variables are considered as supplementary columns in CA.

In the case of an aggregate CA, DocTerm is an aggregate table and:

1. the contextual quantitative variables are considered as supplementary quantitative columns in CA; the value of an active aggregate-document for a variable is the mean of the values corresponding to the source-documents belonging to this aggregate-document.
2. the categories of the contextual qualitative variables are threatened as supplementary rows in CA; these rows contain the frequency with which each the set of documents belonging to this category has used the different words.

Value

Returns a list including:

`eig`	matrix with the eigenvalues, the percentages of inertia and the cumulative percentages of inertia
`row`	list of matrices with all the results for the documents (coordinates, square cosines, contributions, inertia)
`col`	list of matrices with all the results for the words (coordinates, square cosines, contributions, inertia)
`row.sup`	if row.sup is non-NULL, list of matrices with all the results for the supplementary documents (coordinates, square cosines)

col.sup	if col.sup is non-NULL, list of matrices with all the results for the supplementary words (coordinates, square cosines)
quanti.sup	if quanti.sup is non-NULL, list of matrices containing the results for the supplementary quantitative variables (coordinates, square cosines)
quali.sup	if quali.sup is non-NULL, list of matrices with all the results for the supplementary categorical variables; see section details
meta	list of the documents/words whose contribution is over lmd/lmw times the average document/word contribution
VCr	Cramer's V coefficient
Inertia	total inertia
info	information about the corpus
segment	if segment is TRUE, list of matrices with the results for the repeated segments (coordinates, square cosines)
var.agg	name of the aggregation variable in the case of an aggregate correspondence analysis
call	a list with some statistics

Author(s)

Ramón Alvarez-Esteban <ramon.alvarez@unileon.es>, Mónica Bécue-Bertaut, Josep-Anton Sánchez-Espigares

References

Benzécri, J, P. (1981). Pratique de l'analyse des donnees. Linguistique & lexicologie (Vol.3). (P. Dunod., Ed).

Husson F., Lê S., Pagès J. (2011). Exploratory Multivariate Analysis by Example Using R. Chapman & Hall/CRC. doi:10.1201/b10345.

Lebart, L., Salem, A., & Berry, L. (1998). Exploring textual data. (D. Kluwer, Ed.). doi:10.1007/9789401715256.

Murtagh F. (2005). Correspondence Analysis and Data Coding with R and Java. Chapman & Hall/CRC.

See Also

TextData, print.LexCA, plot.LexCA, summary.LexCA, ellipseLexCA

Examples

```
data(open.question)
## Not run:
### non-aggregate CA
res.TD<-TextData(open.question, var.text=c(9,10), Fmin=10, Dmin=10,
        remov.number=TRUE, stop.word.tm=TRUE)
res.LexCA<-LexCA(res.TD, lmd=0, lmw=1)


## End(Not run)


### aggregate CA
```

```
res.TD<-TextData(open.question, var.text=c(9,10), var.agg="Age_Group", Fmin=10, Dmin=10,
        remov.number=TRUE, stop.word.tm=TRUE)
res.LexCA<-LexCA(res.TD, lmd=0, lmw=1)
```

LexChar	*Characteristic words and documents (LexChar)*

Description

Measure of the association between vocabulary or words and quantitative or qualitative contextual variables.

Usage

```
LexChar(object, proba=0.05, maxCharDoc=10, maxPrnDoc=100,
            marg.doc="before",  context=NULL, correct=TRUE, nbsample=500,
            seed=12345,...)
```

Arguments

object	TextData, DocumentTermMatrix, dataframe or matrix object
proba	threshold on the p-value used when selecting the characteristic words (by default 0.05)
maxCharDoc	maximum number of characteristic source-documents to extract (by default 10). See details
maxPrnDoc	maximum length to be printed for a characteristic document (by default 100 characters)
marg.doc	if after/before, frequencies after/before TextData selection are used as document weighting (by default "before"); if before.RW all words under threshold in TextData function are included as a new word named RemovedWords
context	name of quantitative or qualitative variables
correct	if TRUE, pvalue correction test is applied for quantitative contextual variables (by default TRUE)
nbsample	number of samples drawn to evaluate the pvalues in quantitative contextual variables
seed	Seed to obtain the same results using permutation tests (by default 12345)
...	further arguments passed to or from other methods

Details

The lexical table provided by TextData can consider either source-documents or aggregate-documents, in accordance with the value of argument "var.agg" in TextData. Context cualitative variables allow to aggregate documents by combining the categories of the qualitative variables and the aggregation variable if any.

Extracting the characteristic words (CharWord) for a too high number of documents is of no interest and time-consuming.

In any case, only the first maxPrnDoc characters of each characteristic document are printed (by default 100).

In the case of the association between words and qualitative variables, the usual characteristic words are provided.

quali$CharWord provides the qualitative variables (including the aggregation variable) and their categories. quali$stats provides association statistics for vocabulary and qualitative variables (including the aggregation variable). quali$CharDoc provides characteristic source-documents for the categories. quanti$CharWord provides characteristic quantitative variables for each word. If there are aggregation variable and/or qualitative contextual variable, from aggregated lexical table. quanti$stats provides statistics for vocabulary and quantitative variables. If there are aggregation variable and/or qualitative contextual variable, from aggregated lexical table.

If the lexical table (object) is not a TextData object, context argument can be columns of the same dataframe. The aggregate lexical table is constructed from the combinations of the categories of the qualitative variables (including the aggregation variable).

Value

Returns a list including:

CharWord	characteristic words of all the documents
stats	association statistics of the lexical table
CharDoc	characteristic source-documents of all the aggregate-documents including qualitative contextual variables
Vocab	characteristic quantitative and qualitative variables of the words. CharWord and stats are provided.

Author(s)

Monica Bécue-Bertaut, Ramón Alvarez-Esteban <ramon.alvarez@unileon.es>, Josep-Antón Sánchez-Espigares, Belchin Kostov

References

Lebart, L., Salem, A., & Berry, L. (1998). Exploring textual data. (D. Kluwer, Ed.). doi:10.1007/9789401715256.

See Also

TextData, print.LexChar, plot.LexChar, summary.LexChar

Examples

```
data(open.question)
 res.TD<-TextData(open.question, var.text=c(9,10), var.agg="Gen_Edu", Fmin=10, Dmin=10,
                  remov.number=TRUE, stop.word.tm=TRUE)
 res.LexChar <-LexChar(res.TD)
 summary(res.LexChar)
```

LexCHCca	*Chronological Constrained Hierarchical Clustering on Correspondence Analysis Components (LexCHCca)*

Description

Chronological constrained agglomerative hierarchical clustering on a corpus of documents

Usage

```
LexCHCca (object, ncp=5, nb.clust=0, min=2, max=NULL, nb.par=5,
 graph=TRUE, proba=0.05, cut.test=FALSE, alpha.test =0.05, description=FALSE,
 nb.desc=5, size.desc=80)
```

Arguments

object	object of LexCA class
ncp	number of dimensions used from LexCA object (by default 5)
nb.clust	number of clusters only if no test (cut.test=FALSE). If 0 (or "click"), the tree is cut at the level the user clicks on. If -1 (or "auto"), the tree is automatically cut at the suggested level. If a (positive) integer, the tree is cut with nb.clust clusters (by default 0)
min	minimum number of clusters. Available only if cut.test=FALSE. (by default 3)
max	maximum number of clusters. Available only if cut.test=FALSE. (by default NULL; then max is computed as the minimum between 10 and the number of documents divided by 2)
nb.par	number of edited paragons (para) and specific documents labels (dist) (by default 5)
graph	if TRUE, graphs are displayed (by default TRUE)
proba	threshold on the p-value used to describe the clusters (by default 0.05)
cut.test	if FALSE (by default), Legendre test is not performed when joining two nodes. This test is used to determine whether two clusters should be joined or not; see details
alpha.test	threshold on the p-value used in selecting aggregation clusters for Legendre test (by default 0.05)
description	if TRUE, description of the clusters by the characteristic words/documents, paragon (para), specific documents (dist) and contextual variables if these latter have been selected in the previous LexCA function (by default FALSE)
nb.desc	number of paragons (para) and specific documents (dist) that are edited when describing the clusters (by default 5)
size.desc	maximum of characters when editing the paragons (para) and specific documents (dist) to describe the clusters (by default 80)

Details

LexCHCca starts from the document coordinates issued from a textual correspondence analysis. The hierarchical tree is built in such a way that only chronological contiguous nodes can be joined. The documents have to be ranked in their chronological order in the source-base (data frame format) before to apply the function (TextData format).

Legendre test allows to determine whether the fusion between two nodes based on their contiguity lead to a heterogenous new node (no homogeneity-between-clusters). If Legendre test is applied (cut.test=TRUE), the number of clusters is the number obtained by the test and nb.clust has not effects.

If no Legendre test is applied (cut.test= FALSE), the number of clusters is determined either a priori or from the constrained hierarchical tree structure.

The object $para contains the distance between each document and the centroid of its class.

The object $dist contains the distance between each document and the centroid of the farthest cluster.

The results of the description of the clusters and graphs are provided.

Value

Returns a list including:

data.clust	the active lexical table used in LexCA plus a new column called Clust_ containing the partition
coord.clust	coordinates table issued from CA plus a new column called weigths and another column called Clust_, corresponds to the partition
centers	coordinates of the gravity centers of the clusters
description	$des.word for description of the clusters of documents by their characteristic words, the paragons (des.doc$para) and specific documents (des.doc$dist) of each cluster; see details
call	list of internal objects. call$t giving the results for the hierarchical tree
dendro	hclust object. This allows for using the dendrogram in other packages
phases	details of the tracking of the agglomerative hierarchical process. In particular, the cut points (joining documents not allowed) can be identified
sum.squares	sum of squares decomposition for documents and clusters

Author(s)

Monica Bécue-Bertaut, Ramón Alvarez-Esteban <ramon.alvarez@unileon.es>, Josep-Antón Sánchez-Espigares, Belchin Kostov

References

Bécue-Bertaut, M., Kostov, B., Morin, A., & Naro, G. (2014). Rhetorical Strategy in Forensic Speeches: Multidimensional Statistics-Based Methodology. Journal of Classification,31, 85-106. doi:10.1007/s0035701491489.

Husson F., Lê S., Pagès J. (2017). Exploratory Multivariate Analysis by Example Using R. Chapman & Hall/CRC. doi:10.1201/b21874.

Lebart L. (1978). Programme d'agrégation avec contraintes. Les Cahiers de l'Analyse des Données, 3, pp. 275–288.

Legendre, P. & Legendre, L. (1998), Numerical Ecology (2nd ed.), Amsterdam: Elsevier Science.

Murtagh F. (1985). Multidimensional Clustering Algorithms. Vienna: Physica-Verlag, COMP-STAT Lectures.

See Also

plot.LexCHCca, LexCA

Examples

```
data(open.question)
res.TD<-TextData(open.question,var.text=c(9,10), var.agg="Age_Group", Fmin=10, Dmin=10,
        stop.word.tm=TRUE)
res.LexCA<-LexCA(res.TD, graph=FALSE)
res.ccah<-LexCHCca(res.LexCA, nb.clust=4, min=3)
```

LexGalt	*Correspondence Analysis on a Simple or Multiple Generalized Aggregate Lexical Table (LexGalt)*

Description

Performs an extension of correspondence analysis on either a simple or a multiple generalized aggregated lexical table. In the case of a multiple table, a multiple factor analysis approach is used

Usage

```
LexGalt(object, context="ALL", conf.ellip =FALSE, nb.ellip = 100, graph=TRUE,
        axes = c(1, 2), label.group=NULL)
```

Arguments

object	object or list of objects (s) of TextData class (see details)
context	column index(es) or name(s) of the contextual variables (either qualitative or quantitative) used to build the generalized aggregated lexical table(s). These variables must have been previously selected in TextData function (by default "ALL")
conf.ellip	computing confidence ellipses (available only in the case of a simple table) (by default FALSE)
nb.ellip	number of samples drawn to evaluate the stability of the points (by default 100) only if conf.ellip= TRUE
graph	if TRUE, all several graphs are displayed; use plot.LexGalt to obtain detailed graphs (by default TRUE)
axes	length-2 vector indicating the axes to plot (by default axes=c(1,2))
label.group	In the case of analyzing a multiple generalized aggregated lexical table, vector containing the name of the groups (by default, NULL and the group are named GROUP.1, GROUP.2 and so on)

Details

The default "context" argument is "ALL" and may contain qualitative and/or quantitative variables (names or indexes). If both types of variables are included, two independent LexGalt analyses are performed, saving the results for the qualitative analysis into an object named SQL (or MQL in the multiple case) and for the quantitative analysis into the SQN object (or MQN in the multiple case).

In the multiple case, each TextData object must be created from as many executions of the function TextData as there are tables. They are joined in a list in the call to LexGalt function:

LexGalt(list(object1,object2,object3),...).

The variable names of each object in the list must be the same as the name of the variables selected in object1.

Value

Returns a list including an object named SQL if the simple qualitative analysis is performed, SQN for simple quantitative analysis, MQL for multiple qualitative analysis or MQN for multiple quantitative analysis (see details):

eig	eigenvalues, percentages of inertia and cumulative percentages of inertia
word	the results for the words (coordinates, square cosine, contributions)
quali.var	results for the categorical variables (coordinates of each categories of each variables, square cosines)
quanti.var	results for the quantitative variables (coordinates, correlation between variables and axes, square cosines)
ellip	coordinates for confidence ellipses (words and categories) are drawn
group	in the case of multiple analysis, results for the groups (coordinates, contributions and square cosines) (MQL or MQN)

Returns the factor maps. The plots may be improved using the plot.LexGalt function.

Author(s)

Belchin Kostov, Monica Bécue-Bertaut, Ramón Alvarez-Esteban <ramon.alvarez@unileon.es>, Josep-Antón Sánchez-Espigares

References

Bécue-Bertaut M. and Pagès J. (2015). Correspondence analysis of textual data involving contextual information: CA-GALT on principal components. Advances in Data Analysis and Classification, vol.(9) 2: 125-142. doi:10.1007/s1163401401719

Bécue-Bertaut M., Pagès J. and Kostov B. (2014). Untangling the influence of several contextual variables on the respondents' lexical choices. A statistical approach. SORT - Statistics and Operations Research Transactions, vol.(38) 2: 285-302.

Kostov B. A. (2015). A principal component method to analyse disconnected frequency tables by means of contextual information. (Doctoral dissertation). Retrieved from https://upcommons.upc.edu/handle/2117/95759.

Kostov, B., Bécue-Bertaut, M., & Husson, F. (2015). Correspondence Analysis on Generalised Aggregated Lexical Tables (CA-GALT) in the FactoMineR Package. The R Journal, Vol.7, Num.1, 109-117. doi:10.32614/RJ2015010

See Also

plot.LexGalt

Examples

```
data(open.question)

res.TD<-TextData(open.question,var.text=c(9,10), Fmin=10, Dmin=10,
 context.quali=c("Gender", "Age_Group", "Education"),
 remov.number=TRUE, stop.word.tm=TRUE)

# res.LexGalt <- LexGalt(res.TD, graph=FALSE, conf.ellip =FALSE)
# plot(res.LexGalt, selQualiVar="ALL")
```

LexHCca	*Hierarchical Clustering on Textual Correspondence Analysis Coordinates (LexHCca)*

Description

Agglomerative hierarchical clustering of documents or words issued from correspondence analysis coordinates

Usage

```
LexHCca(x, cluster.CA="docs",  type="agnes", ncp=5, nb.clust="click", min=2,
    max=NULL, kk=Inf, consol=FALSE, iter.max=500, graph=TRUE, description=TRUE,
    proba=0.05, nb.desc=5, size.desc=80, seed=12345,...)
```

Arguments

x	object of LexCA class
cluster.CA	if "rows" or "docs" cluster analysis is performed on documents; if "columns" or "words", cluster analysis is performed on words (by default "docs")
'	
type	type of cluster; "agnes" (Agglomerative), "diana" (Divisive) (by default agnes)
ncp	number of dimensions used from LexCA object (by default 5)
nb.clust	number of clusters. If 0 (or "click"), the tree is cut at the level the user clicks on. If -1 (or "auto"), the tree is automatically cut at the suggested level. If a (positive) integer, the tree is cut with nb.clust clusters (by default "click")
min	minimum number of clusters (by default 2)
max	maximum number of clusters (by default NULL, then max is computed as the minimum between 10 and the number of documents divided by 2)
kk	in case the user wants to perform a Kmeans clustering previously to the hierarchical clustering (preprocessing step), kk is an integer corresponding to the number of clusters of this previous partition. Further, the hierarchical tree is constructed starting from the nodes of this partition as terminal elements. This is very useful when the number of elements to be classified is very large. By default, the value is Inf and no Kmeans preprocessing is performed

consol	if TRUE, a Kmeans consolidation step is performed after the hierarchical clustering (consolidation cannot be performed if kk is used and equals a number) (by default FALSE)
iter.max	maximum number of iterations in the consolidation step (by default 500)
graph	if TRUE, graphs are displayed (by default TRUE)
description	if TRUE, description of the clusters of documents or words by the axes, the characteristic words in the case of clustering documents or the characteristic documents in the case of clustering words. The documents or words considered as paragon (para) or specific (dist) are identified. In the case of clustering documents, contextual variables also characterize the clusters. These variables have to be selected in LexCA (by default TRUE)
proba	threshold on the p-value used in selecting the elements characterizing significantly the clusters (by default 0.05)
nb.desc	Maximum of characters when editing the paragons (para) and specific documents (dist) to describe the clusters (by default 80))
size.desc	text size of edited paragons (para) and specific documents (dist) when describing the clusters of documents (by default 80)
seed	Seed to obtain the same results in successive Kmeans (by default 12345)
...	other arguments from other methods

Details

LexHCca starts from the documents/words coordinates issued from correspondence analysis axes. Euclidean metric and Ward method are used.

If the agglomerative clustering starts from many elements (documents or words), it is possible to previously perform a Kmeans partition with kk clusters to further build the tree from these (weighted) kk clusters.

The object $para contains the distance between each document and the centroid of its class.

The object $dist contains the distance between each document and the centroid of the farthest cluster.

The results include a thorough description of the clusters. Graphs are provided.

Value

Returns a list including:

data.clust	the active lexical table used in LexCA plus a new column called Clust_ containing the partition
coord.clust	coordinates table issued from CA plus a new column called Clust_ containing the partition
centers	coordinates of the gravity centers of the clusters
clust.count	counts of documents/words belonging to each cluster and contribution of the clusters to the variability decomposition
clust.content	list of the document/word labels according to the cluster they belong to
call	list of internal objects. call$t giving the results for the hierarchical tree. See the second reference for more details

description	$desc.axes for description of the clusters by the characteristic axes ($axes) and eta-squared between axes and clusters ($quanti.var).
	$des.cluster.doc for description of the clusters by their characteristic words ($word), supplementary words ($wordsup) and, if contextual variables were considered in LexCA, description of the partition/clusters by qualitative ($qualisup) and quantitative ($quantisup) variables, paragons ($para) and specific words ($dist) of each cluster.
	$des.word.doc description of the clusters of words by their characteristic documents ($docs), paragons ($para) and specific documents ($dist) of each cluster.
type	Type of cluster used (by default agnes).
coef.hclust	Agglomerative coefficient (Divisive coefficient for diana), measuring the clustering structure of the dataset.

Returns the hierarchical tree and the first CA map of the documents/words. The labels are colored according to the cluster.

Author(s)

Ramón Alvarez-Esteban <ramon.alvarez@unileon.es>, Monica Bécue-Bertaut, Josep-Anton Sánchez-Espigares

References

Bécue-Bertaut M. Textual Data Science with R. Chapman & Hall/CRC. doi:10.1201/9781315212661.

Husson F., Lê S., Pagès J. (2017). Exploratory Multivariate Analysis by Example Using R. Chapman & Hall/CRC. doi:10.1201/b21874.

Lebart, L., Salem, A., & Berry, L. (1998). Exploring textual data. (D. Kluwer, Ed.). doi:10.1007/9789401715256.

See Also

LexCA, plot.LexHCca

Examples

```
data(open.question)
res.TD<-TextData(open.question, var.text=c(9,10), Fmin=10, Dmin=10, stop.word.tm=TRUE,
        context.quali=c("Gender","Age_Group","Education"), context.quanti=c("Age"))
res.LexCA<-LexCA(res.TD, graph=FALSE, ncp=8)
res.hcca<-LexHCca(res.LexCA, graph=FALSE, nb.clust=5)
```

---

open.question	*Open.question (data)*

Description

Extract of the answers provided in a survey designed to better know opinions about what is most important in life.

Two open-ended questions are included in the questionnaire "What is most important to you in life?" and "What are other very important things to you? (relaunch of the first question).

Usage

```
data(open.question)
```

Format

Data frame with 300 rows and 10 columns. The rows correspond to the respondents. The first 8 columns correspond to socio-demographic variables collected through closed questions: Gender, Age_Group, Age, Education level, Genre crossed with Age, Genre crossed with Education level, Age crossed with Education level and, finally Genre crossed with Education level and Age. Age is a quantitative variable while the other variables are qualitative. The last two columns contain the answers to the open-ended questions.

---

plot.LexCA   *Plot of LexCA objects*

Description

Plots textual correspondence analysis (CA) graphs from a LexCA object.

Usage

```
## S3 method for class 'LexCA'
plot(x, selDoc="ALL", selWord="ALL", selSeg=NULL, selDocSup=NULL,
  selWordSup=NULL, quanti.sup=NULL, quali.sup=NULL, maxDocs=20, eigen=FALSE,
 title=NULL, axes=c(1,2), col.doc="blue", col.word="red", col.doc.sup="darkblue",
  col.word.sup="darkred", col.quanti.sup = "blue", col.quali.sup="darkgreen",
  col.seg="cyan4", col="grey", cex=1, xlim=NULL, ylim=NULL, shadowtext=FALSE,
  habillage="none", unselect=1, label="ALL", autoLab=c("auto", "yes", "no"),
  new.plot=TRUE, graph.type = c("classic", "ggplot"),...)
```

Arguments

x	object of LexCA class
selDoc	vector with the active documents to plot (indexes, names or rules; see details; by default "ALL")
selWord	vector with the active words to plot (indexes, names or rules; see details; by default "ALL")
selSeg	vector with the supplementary repeated segments to plot (indexes, names or rules; see details; by default NULL)
selDocSup	vector with the supplementary documents to plot (indexes, names or rules; see details; by default NULL)
selWordSup	vector of the supplementary words to plot (indexes, names or rules; see details; by default NULL)
quanti.sup	vector of the supplementary quantitative variables to plot (indexes, names or rules; see details; by default NULL)
quali.sup	vector with the supplementary categorical variables/categories to plot (indexes, names or rules; see details; by default NULL). The selected categories (through the variables or directly) are plotted

maxDocs	limit to the number of active documents in the lexical table when selecting the words to be plotted for being characteristic of the selected documents (by default 20)
eigen	if TRUE, the eigenvalues barplot is drawn (by default FALSE); no other elements can be simultaneously selected
title	title of the graph (by default NULL and the title is automatically assigned)
axes	length-2 vector indicating the axes considered in the graph (by default c(1,2))
col.doc	color for the point-documents (by default "blue")
col.word	color for the point-words (by default "red")
col.doc.sup	color for the supplementary point-documents (by default "darkblue")
col.word.sup	color for the supplementary point-words (by default "darkred")
col.quanti.sup	color for the quanti.sup variables (by default "blue")
col.quali.sup	color for the categorical supplementary point-categories, (by default "darkgreen")
col.seg	color for the supplementary point-repeated segments, (by default "cyan4")
col	color for the bars in the eigenvalues barplot (by default "grey")
cex	text and symbol size is scaled by cex, in relation to size 1 (by default 1)
xlim	range for 'x' values on the graph, defaulting to the finite values of 'x' range (by default NULL)
ylim	range for the 'y' values on the graph, defaulting to the the finite values of 'y' range (by default NULL)
shadowtext	if TRUE, shadow on the labels (rectangles are written under the labels which may lead to difficulties to modify the graph with another program) (by default FALSE)
habillage	index or name of the categorical variable used to differentiate the documents by colors given according to the category; by default "none")
unselect	either a value between 0 and 1 or a color. In the first case, transparency level of the unselected objects (if unselect=1 the transparency is total and the elements are not represented; if unselect=0 the elements are represented as usual but without any label); in the case of a color (e.g. unselect="grey60"), the non-selected points are given this color (by default 1)
label	a list of character for the variables which are labelled (by default ALL and all the drawn variables are labelled). You can label all the active variables by putting "var" and/or all the supplementary variables by putting "quanti.sup" and/or a list with the names of the variables which should be labelled. Value should be one of "all", "none", "row", "row.sup", "col", "col.sup", "quali.sup" or NULL.
autoLab	if autoLab="auto", autoLab turns to be equal to "yes" if there are less than 50 elements and equal to "no" otherwise; if "yes", the labels are moved, as little as possible, to avoid overlapping (time-consuming if many elements); if "no" the labels are placed quickly but may overlap
new.plot	if TRUE, a new graphical device is created (by default FALSE)
graph.type	a string that gives the type of graph used: "ggplot" or "classic" (by default classic)
...	further arguments passed from other methods...

Details

The argument autoLab = "yes" is time-consuming if many overlapping labels. Furthermore, the visualization of the words cloud can result distorted because of the apparent greater dispersion of the words labels. An alternative would be reducing the character size of the words labels to reduce overlapping (e.g. cex=0.7).

selDoc, selWord, selSeg, selDocSup, selWordSup, quanti.sup and quali.sup allow for selecting all or part of the elements of the corresponding type, using either labels, indexes or rules.

The syntax is the same for all types.

1. Using labels:

```
selDoc = c("doc1","doc5"): only the documents with labels doc1 and doc5 are plotted.
quali.sup=c("varcateg1","category12"): only the categories (all of them) of
   categorical variable labeled "varcateg1" and the category labeled "category12"
   are plotted.
```

2.- Using indexes:

```
selDoc = c(1:5): documents 1 to 5 are plotted.
quali.sup=c(1:5,7): categories 1 to 5 and 7 are plotted. The numbering of the
   categories have to be consulted in the LexCA numerical results.
```

3.- Using rules: Rules are based on the coordinates (coord), the contribution (contrib or meta; concerning only active elements) or the square cosine (cos2).
Somes examples are given hereafter:

```
selDoc="coord 10": only the 10 documents with the highest coordinates, as globally
   computed on the 2 axes, are plotted.
selWord="contrib 10": the words with a contribution to the inertia, of any of
   the 2 axes.
selWord="meta 3": the words with a contribution over 3 times the average word
   contribution on any of the two axes are plotted. Only active words or documents
   can be selected.
selDocSup="cos2 .85": the supplementary documents with a cos2 over 0.85, as summed
   on the 2 axes, are plotted.
selWord="char 0.05": only the characteristic words of the documents selected in
   SelDoc are plotted. The selection of the words follow the rationale used in
   function LexChar using as limit for the p-value the value given, here.0.05.
```

Author(s)

Ramón Alvarez-Esteban <ramon.alvarez@unileon.es>, Mónica Bécue-Bertaut, Josep-Antón Sánchez-Espigares

References

Husson F., Lê S., Pagés J. (2011). Exploratory Multivariate Analysis by Example Using R. Chapman & Hall/CRC. doi:10.1201/b10345.

See Also

LexCA, print.LexCA, summary.LexCA

Examples

```
data(open.question)
res.TD<-TextData(open.question,var.text=c(9,10), var.agg="Age_Group", Fmin=10, Dmin=10,
        remov.number=TRUE, stop.word.tm=TRUE)
res.CA <- LexCA(res.TD, graph=FALSE)
plot(res.CA, selDoc="contrib 30", selWord="coord 20")
```

---

plot.LexChar *Plot LexChar objects*

Description

Draws the characteristic and anti-characteristic words of documents from a LexChar object.

Usage

```
## S3 method for class 'LexChar'
plot(x, char.negat=TRUE, col.char.posit="blue", col.char.negat="red",
col.lines="black", theme=theme_bw(), text.size=12, numr=1, numc=2, top=NULL,
max.posit=15, max.negat=15, type=c("CharWord","quanti","quali"),sel.var.cat="ALL",
txt.var.cat=NULL, sel.words="ALL",...)
```

Arguments

x	object of LexChar class
char.negat	if TRUE, the anti-characteristic words are plotted (by default TRUE)
col.char.posit	color for the characteristic words (by default "blue")
col.char.negat	color for the anti-characteristic words (by default "red")
col.lines	color for the lines of barplot (by default "black")
theme	used to modify the theme settings by ggplot2 package (by default theme_bw())
text.size	size of the font (by default 12)
numr	number of rows in each multiple graph (by default 1 row)
numc	number of columns in each multiple graph (by default 2 columns)
top	title of the graph (by default NULL)
max.posit	maximum number of characteristic words (by default 15)
max.negat	maximum number of anti-characteristic words (by default 15)
type	CharWord and draws the characteristic and anti-characteristic words; quanti draws characteristic words for all quantitative variables; quali draws only the words for one qualitative variable (by default CharWord)
sel.var.cat	name of contextual quantitative and/or qualitative contextual variables
txt.var.cat	new names of each category or quantitative variable (by default NULL
sel.words	words selected to plot if its p-value is less than prob (by default ALL
...	further arguments passed to or from other methods...

Author(s)

Ramón Alvarez-Esteban <ramon.alvarez@unileon.es>, Monica Bécue-Bertaut, Josep-Anton Sánchez-Espigares

See Also

[LexChar](), [print.LexChar](), [summary.LexChar]()

Examples

```
data(open.question)
res.TD<-TextData(open.question, var.text=c(9,10), var.agg="Gen_Edu", Fmin=10, Dmin=10,
        remov.number=TRUE, stop.word.tm=TRUE)
LD<-LexChar(res.TD,maxCharDoc = 0)
plot(LD)
```

| plot.LexCHCca | *Plots for Chronological Constrained Hierarchical Clustering from LexCHCca Objects* |

Description

Plots graphs from LexCHCca results: tree, barplot of the aggregation criterion values and first CA map with the documents colored in accordance with the cluster.

Usage

```
## S3 method for class 'LexCHCca'
plot(x, axes=c(1, 2), type=c("tree","map","bar"), rect=TRUE, title=NULL,
  ind.names=TRUE, new.plot=FALSE, max.plot=15, tree.barplot=TRUE,...)
```

Arguments

x	object of LexCHCca class
axes	length-2 vector defining the axes of the CA map to plot (by default (1,2))
type	type of graph. "tree" plots the tree; "bar" plots the barplot of the successive values of the aggregation criterion (downward reading of the tree); "map" plots the CA map where the individuals are colored in accordances with the cluster of belonging (by default "tree")
rect	if TRUE, when choice="tree" rectangles are drawn around the clusters (by default TRUE)
title	title of the graph. If NULL, a title is automatically defined (by default NULL)
ind.names	if TRUE, the document labels are written on the CA map (by default TRUE)
new.plot	if TRUE, a new window is opened (by default FALSE)
max.plot	maximum of bars in the bar plot of the aggregation criterion (by default 15)
tree.barplot	if TRUE, the barplot of intra inertia losses is added on the tree graph (by default TRUE)
...	further arguments passed from other methods...

Value

Returns the chosen plot

Author(s)

Mónica Bécue-Bertaut, Ramón Alvarez-Esteban <ramon.alvarez@unileon.es>, Josep-Anton Sánchez-Espigares

See Also

LexCHCca

Examples

```
## Not run:
data(open.question)
res.TD<-TextData(open.question,var.text=c(9,10), var.agg="Age_Group", Fmin=10, Dmin=10,
        stop.word.tm=TRUE)
res.LexCA<-LexCA(res.TD, graph=FALSE)
res.chcca<-LexCHCca(res.LexCA, nb.clust=4, min=3, graph=FALSE)
plot(res.chcca, type="tree")
plot(res.chcca, type="map")
plot(res.chcca, type="bar", max.plot=5)

## End(Not run)
```

---

plot.LexGalt   *Plot LexGalt objects*

Description

Plots Generalised Aggregate Lexical Tables (LexGalt) graphs from a LexGalt object

Usage

```
## S3 method for class 'LexGalt'
plot(x,type="QL", selDoc=NULL, selWord=NULL, selQualiVar=NULL,
 selQuantiVar=NULL, conf.ellip=FALSE, selWordEllip=NULL, selQualiVarEllip=NULL,
 selQuantiVarEllip=NULL, level.conf=0.95, eigen=FALSE, title = NULL, axes = c(1, 2),
 xlim = NULL, ylim = NULL, col.eig="grey", col.doc = "black", col.word = NULL,
 col.quali = "blue", col.quanti = "blue", col="grey", pch = 20, label = TRUE,
 autoLab = c("auto", "yes", "no"), palette = NULL, unselect = 1,
 selCov=FALSE, selGroup="ALL", partial=FALSE, plot.group=FALSE,
 col.group=NULL, label.group=NULL, legend=TRUE, pos.legend="topleft",
 new.plot = TRUE, cex=1,...)
```

Arguments

x	object of LexGalt class
type	results from a qualitative analysis (type="QL") or quantitative analysis (type="QN"); see details; by default Q)
selDoc	vector with the documents to plot (indexes, names or rules; see details; by default NULL)
selWord	vector with the words to plot (indexes, names or rules (indexes, names or rules; see details; by default NULL)

selQualiVar	vector with the categories of categorical variables to plot (indexes, names or rules; see details; by default NULL)
selQuantiVar	vector with the numerical variables to plot (indexes, names or rules; see details; by default NULL)
conf.ellip	to drawn confidence ellipses, by default FALSE
selWordEllip	vector with the words that defines which ellipses are drawn (indexes, names or rules; see details; by default NULL)
selQualiVarEllip	
	vector with the categories of categorical variables which ellipses are drawn (indexes, names or rules; see details; by default NULL)
selQuantiVarEllip	
	vector with the numerical variables which ellipses are drawn(indexes, names or rules; see details; by default NULL)
level.conf	level of confidence used to construct the ellipses; by default 0.95
eigen	if TRUE, the eigenvalues barplot is drawn (by default FALSE); other elements can be simultaneously selected
title	title of the graph (by default NULL and the title is automatically assigned)
axes	length-2 vector indicating the axes considered in the graph; by default c(1,2)
xlim	range for 'x' values on the graph, defaulting to the finite values of 'x' range (by default NULL)
ylim	range for the 'y' values on the graph, defaulting to the the finite values of 'y' range (by default NULL)
col.eig	value or vector with colors for the bars of eigenvalues (by default "grey")
col.doc	color for the point-documents(by default "black")
col.word	color for the point-words (by default NULL is darkred in simple analysis; see details)
col.quali	color for the categories of categorical variables (by default "blue")
col.quanti	color for the numerical variables (by default "blue")
col	color for the bars in the eigenvalues barplot (by default "grey")
pch	plotting character for coordinates, cf. [points](#) function in the graphics package
label	a list of character for the elements which are labelled (by default TRUE and all the drawn elements are labelled).
autoLab	if autoLab="auto", autoLab turns to be equal to "yes" if there are less than 50 elements and equal to "no" otherwise; if "yes", the labels are moved, as little as possible, to avoid overlapping (time-consuming if many elements); if "no" the labels are placed quickly but may overlap
palette	the color palette used to draw the points. By default colors are chosen. If you want to define the colors : palette=c("black", "red", "blue"); or you can use: palette=rainbow(10), or in black and white for example: palette=gray(seq(0,.9,len=3))
unselect	may be either a value between 0 and 1 that gives the transparency of the unselected objects (if unselect=1 the transparceny is total and the elements are not drawn, if unselect=0 the elements are drawn as usual but without any label) or may be a color (for example unselect="grey60")
selCov	a boolean, if TRUE then data are scaled to unit variance (by default TRUE)
selGroup	vector with the groups to plot if multiple analysis was performed (indexes, names or rules; see details; by default NULL)

`partial`	if TRUE partial elements (results for the groups) are shown, if ALL results for the conjoint analysis are superimposed; by default FALSE
`plot.group`	draw a plot comparing the groups in multiple case (by default TRUE)
`col.group`	color for the groups if multiple analysis was performed (by default NULL and they are selected from palette)
`label.group`	a vector containing the new name of the groups. If "BLANK" no labels with the group are added at the end of the drawn elements (by default, NULL and the name of each group is added)
`legend`	show the legend of labels of groups. See [legend](#) from graphics package (by default TRUE
`pos.legend`	position of the legend of labels of groups. See [legend](#) from graphics package (by default "topleft")
`new.plot`	if TRUE, a new graphical device is created (by default TRUE)
`cex`	text and symbol size is scaled by cex, in relation to size 1 (by default 1)
`...`	further arguments passed from other methods...

Details

The argument autoLab = "yes" is time-consuming if many overlapping labels. Furthermore, the visualization of the words cloud can result distorted because of the apparent greater dispersion of the words labels. An alternative would be reducing the character size of the words labels to reduce overlapping (e.g. cex=0.7).

selDoc, selWord, selQualiVar, selQuantiVar, selWordEllip, selQualiVarEllip, selQuantiVarEllip allow for selecting all or part of the elements of the corresponding type, using either labels, indexes or rules.

The syntax is the same for all types.

1. Using labels:

```
selDoc = c("doc1","doc5"): only the documents with labels doc1 and doc5 are plotted.
selQualiVar=c("category1","category2"): only the categories labeled category1 and
 category2 are plotted.
```

2.- Using indexes:

```
selDoc = c(1:5): documents 1 to 5 are plotted.
quali.sup=c(1:5,7): categories 1 to 5 and 7 are plotted. The numbering of the
   categories have to be consulted in the LexGalt numerical results.
```

3.- Using rules: Rules are based on the coordinates (coord), the contribution (contrib or meta) or the square cosine (cos2).
Somes examples are given hereafter:

```
selDoc="coord 10": only the 10 documents with the highest coordinates, as globally
   computed on the 2 axes, are plotted.
selWord="contrib 10": the words with a contribution to the inertia, of any of
   the 2 axes.
selWord="meta 3": the words with a contribution over 3 times the average word
   contribution on any of the two axes are plotted.
selWord="cos2 .85": the words with a cos2 over 0.85, as summed
   on the 2 axes, are plotted.
```

```
col.word by default NULL is "darkred" for simple analysis, if it is null takes
the colors from col.group
i.e. col.group=c("red","blue"). To select the colors for some words in object res,
we can use:
str.col.words <- rep("darkred",nrow(res$MQL$word$coord))
str.col.words[which(rownames(res$MQL$word$coord) == "kids")] <- "red"
str.col.words[which(rownames(res$MQL$word$coord) == "friends")] <- "green"
str.col.words[which(rownames(res$MQL$word$coord) == "job")] <- "pink"
plot(res, selGroup=1, selWord=c("friends", "job", "kids", "at"),new.plot=FALSE,
col.group=c("darkred","blue"), autoLab = "yes", col.word=str.col.words)
```

Author(s)

Belchin Kostov, Monica Bécue-Bertaut, Ramón Alvarez-Esteban <ramon.alvarez@unileon.es>,
Josep-Antón Sánchez-Espigares

References

Bécue-Bertaut M. and Pagès J. (2015). Correspondence analysis of textual data involving contextual
information: CA-GALT on principal components. Advances in Data Analysis and Classification,
vol.(9) 2: 125-142.

Bécue-Bertaut M., Pagès J. and Kostov B. (2014). Untangling the influence of several contextual
variables on the respondents' lexical choices. A statistical approach. SORT - Statistics and Opera-
tions Research Transactions, vol.(38) 2: 285-302.

Kostov B. A. (2015). A principal component method to analyse disconnected frequency tables
by means of contextual information. (Doctoral dissertation). Retrieved from http://upcommons.
upc.edu/handle/2117/95759.

See Also

LexGalt

Examples

```
## Not run:
data(open.question)

res.TD<-TextData(open.question,var.text=c(9,10),  Fmin=10, Dmin=10,
 context.quali=c("Gender", "Age_Group", "Education"),
 remov.number=TRUE, stop.word.tm=TRUE)

res.LexGalt <- LexGalt(res.TD, graph=FALSE, nb.ellip =0)
plot(res.LexGalt, selQualiVar="ALL")

## End(Not run)
```

---

plot.LexHCca   *Plots for Hierarchical Clustering from LexHCca Objects*

Description

Plots graphs from LexHCca results: tree and CA maps with the documents or words colored in
accordance with the cluster.

Usage

```
## S3 method for class 'LexHCca'
plot(x, type=c("map", "tree", "phylo", "clado", "radial", "fan"),
     plot=c("points", "labels", "centers"), selClust="ALL",
     selInd="ALL",axes=c(1, 2), theme=theme_bw(), palette=NULL, title=NULL,
     axis.title=NULL, axis.text=NULL, xlim=NULL, ylim=NULL, hvline=NULL,
     points=NULL, labels=NULL,centers=NULL, traject=NULL, hull=NULL,
     rotate=FALSE, branches=NULL,...)
```

Arguments

x	object of LexHCca class
type	type of graph. "map" plots the CA map where the individuals are colored in accordance with the cluster of belonging (by default); "tree" plots the dendrogram if hierarchical method without consolidation is performed from LexHCca; other options are "phylo", "clado", "radial", "fan". See details
plot	elements to plot for map graph: points, labels, centers, hull, hvline or traject; by default "ALL" and points, labels and centers are plotted. Also combinations are allowed, i.e: plot=c(points,centers); For no maps plot elements are: branches, labels, hull and hvline. See details
selClust	vector indexes with the numbers of the clusters to plot (by default "ALL")
selInd	vector with the active documents/words to plot (indexes, names or rules; see details; by default "ALL"). You can also use the "transparent"" option defining the color for clusters and/or cases
axes	length-2 vector indicating the axes of the CA map to plot; by default (1,2)
theme	used to modify the theme settings by ggplot2 package of the CA map (by default theme_bw())
palette	color palette used to draw the clusters. As many numbers as clusters. See details
title	title of the map graph. If NULL or FALSE, a title is automatically defined (by default NULL). Other parameters can be chosem using for map in a list: text, color, size, family, face, just; For "tree" only "text" argument can be used. See details
axis.title	axis titles parameters can be used por map plots: text.x, text.y, color, size, family, face, just; If text.x and text.y are NULL automatic texts are plotted (by default NULL). ; For tree only FALSE are allowed and height are removed. See details
axis.text	For maps, format of numbers can be chosen: color, size, family, face
xlim	For map, pair of values xlim=c(xmin,xmax). If a NA value, this limit is automatically calculated
ylim	For map, pair of values ylim=c(ymin,ymax). If a NA value, this limit is automatically calculated
hvline	For map, horizontal (intercept.y) and vertical line (intercept.x) added by default at (0,0) position in map. Parameters: intercept.y, intercept.x, linetype (by default "dashed"), color, linesize, alpha.t. For tree draws a line at level of the height chosen by the clusters selected. Parameters pos (position), linesize, linetype and color
points	For maps: format of points. Parameters: size (if size=0 the points are no plotted), shape (by default 21), fill (if a color, the same for all the points, if color is NULL palette colors used for the clusters are applied; if more than one color use

	palette argument; only for shapes from 21 to 25 to fill the point), stroke (controls the edge of the point (by default 0 no edge), border (color of the border, same specifications than fill), alpha.t (by default 1). See geom_point() in ggplot2 library. See details
labels	format of labels. For no maps: cex (value or vector with the length of cases, if 0 transparent) and color. For map plots, parameters: cex, size (if size=0 the labels are not plotted; by default 4), family, face, hjust, vjust, color.text, alpha.t.text, numbers(if TRUE the label will be replaced by the number of the cluster to which it belongs, by default FALSE), color.fill (color into the rectangle, by default FALSE is transparent), alpha.t.fill, groupLabels and labels will be added to each cluster in tree plots. For map: force (to do repulsive textual annotations and make it easier to read), max.overlaps (maximum number of overlapped points, by default 10, can be Inf), set.seed (by default a new seed for each plot draws different positions, for the same seed i.e: set.seed=1234)
traject	for map: draws trajectory arrows in accordance with the order of clusters or in the selInd order. Parameters: color (by default blue), linetype (by default 1 solid), space (by default 0 and no space is added from point to arrow, be careful with this value), size (width,by default 1), arrow.length (of the arrow, by default .3), arrow.type (by default "closed"), arrow.angle (by default 30), alpha.t. See geom_segment for details
centers	draws the barycenter of the clusters. Parameters: size (by default 5), family, face, color (of the border, only one), fill, alpha.t, labels (string vector with the names of the clusters)
hull	draws a hull containing all the elements of each cluster. Parameters: type (ellipse, by default, hull), alpha.t, color, linetype (by default "dotted")). For tree, no null value, rect for example, draws a rectangle. See details.
rotate	rotation degrees, TRUE or FALSE. Not allowed for map. By default 0 or FALSE.
branches	color, linesize and linetype (integer (0-6), a name (0 = blank, 1 = solid, 2 = dashed, 3 = dotted, 4 = dotdash, 5 = longdash, 6 = twodash)
...	other arguments from other methods

Details

Parameter type="tree" shows the dendrogram

```
- if hierarchical cluster without consolidation is performed.
- if hierarchical cluster with consolidation before the consolidation.
- if kmeans the hierarchical tree with the output of kmeans.
```

You can make customer dendrograms by accessing the hclust format object located inside the object in hclust format from object$call$t$tree

Selection of individuals (documents or words) to plot:

1. Using labels:

```
selInd = c("doc1","doc5"): only the documents with labels doc1 and doc5 are plotted.
```

2. Using indexes:

```
selInd = c(1:5): cases 1 to 5 are plotted.
```

3. Using rules:

```
Rules are based on the coordinates (coord), the contribution (contrib or meta;
concerning only active elements) or the square cosine (cos2).
```

Somes examples hereafter:

```
selInd="coord 10": only the 10 cases with the highest coordinates, as globally
   computed on the 2 axes, are plotted.
selInd="contrib 10": the cases with a contribution to the inertia, of any of
   the 2 axes over 10 percent.
selInd="meta 3": the cases with a contribution over 3 times the average word/document
   contribution on any of the two axes are plotted.
selInd="cos2 .85": the documents with a cos2 over 0.85, as summed on the 2 axes,
   are plotted.
```

Parameters can be used in combination, e.g.: title=list("text"="CA", "color"="red").

See grDevices package (The R Graphics Devices and Support for Colours and Fonts).

palette, the color of the palette used to draw the points. By default colors are chosen. If you want to define the colors for three clusters : palette=c("black","red","blue"); or you can use: palette= palette(rainbow(30)); or in black and white for example: palette=palette(gray(seq(0,.9,len=25))).

Family Fonts (family). Also see the extrafont package for a much better support of fonts: library(extrafont); font_import(). By default "family"='serif'.

Face fonts (face). Can be 'plain', 'bold', 'italic', 'bold.italic', 'symbol'. By default 'plain'.

alpha.t is the level of transparency for some objects. 0 value means full transparency and 1 opacity. By default 1.

Values for horizontal justification hjust, vertical vjust and both hvjust can be (c,centered or 0.5 if centered; l,left or 0 if left; r, right or 1 if right).

groupLabels, only for tree, can be NULL or FALSE and no labels are added to each cluster, TRUE for all the clusters numbers are used, "as.roman", "letters" or "LETTERS" for capital letters. For several lines in the same cluster or no labels:labels=list("groupLabels"=c(paste0("FirstLine,"\n","SecondLine", "b", "").

```
By default in:
* title: text="Clusters on the CA map"; color=black; size=18; familiy=serif; face=plain;
      hjust=0.5.

* axis titles:  text.x=Dim x (%), text.y=Dim y (%), color=black, size=12, family=serif,
      face=plain, just=centered.

* axis.text: color=black, size=8, family=serif, face=plain.

* hvline: intercept.x=0, intercept.y=0, linetype=dashed, color=gray, size=0.5, alpha.t=1.

* points: size=2, shape=21, border:automatic cluster color, fill:automatic cluster color,
      stroke=0, border: automatic cluster color, alpha.t=1.

* labels: size=4, family=serif, face=plain, hjust=1, vjust=1, color.text=same of points,
     alpha.t.text=1, numbers=FALSE, rect=FALSE, color.fill=transparent, alpha.t.fill=1,
      force=1, max.overlaps=10.
```

* traject: color=blue, linetype=solid, space=1, arrow.length=.3, arrow.type= closed,
        arrow.angle=30, alpha.t=1.

* centers: size=5, family=serif, face=italic, color, fill=automatic cluster color,
        alpha.t=1, labels=automatic strig vector with the names of the clusters.

* hull: type=ellipse, alpha.t=0.1, color=black, linetype=dotted .
    For rectangles in tree, you can use some dendextend::rect.dendrogram arguments as
    which for select the cluster, border for the color, prop_k_height (value between 0
    to 1, indicating what proportion of the height our rect will be between the height
    needed for k and k+1 clustering), lower_rect (value of how low should the lower
    part of the rect be), upper_rect (value to add (default is 0) to how high should
        the upper part of the rect be).

Author(s)

Ramón Alvarez-Esteban <ramon.alvarez@unileon.es>, Mónica Bécue-Bertaut, Josep-Anton Sánchez-Espigares

References

The Xplortext web site provides several examples at <https://xplortext.unileon.es/?page_id=766>.

See Also

[LexHCca](#)

Examples

```
data(open.question)
res.TD<-TextData(open.question,var.text=c(9,10), var.agg="Age_Group", Fmin=10, Dmin=10,
        stop.word.tm=TRUE)
res.LexCA<-LexCA(res.TD, graph=FALSE)
res.hcca<-LexHCca(res.LexCA, nb.clust=4, min=3, graph=FALSE)
plot(res.hcca, type="tree")
plot(res.hcca, type="map")
```

---

plot.TextData   *Plot TextData objects*

Description

Draws the barcharts of the longest documents, most frequent words and segments from a TextData
object.

Usage

```
## S3 method for class 'TextData'
plot(x, ndoc=25, nword=25, nseg=25, sel=NULL, ordFreq=TRUE,
 stop.word.tm=FALSE, idiom="en", stop.word.user=NULL, theme=theme_bw(), title=NULL,
 xtitle=NULL, col.fill="grey", col.lines="black", text.size=12, freq=NULL, vline=NULL,
 interact=FALSE, round.dec = 4,...)
```

Arguments

x	object of TextData class
ndoc	number of documents in the barchart (by default 25)
nword	number of words in the barchart (by default 25)
nseg	number of segments in the barchart (by default 25)
sel	type of barchart (doc, word or seg for documents, words or repeated segments) (by default NULL and all he graphs are drawn)), see details
ordFreq	if ordFreq=TRUE, glossaries of words and repeated segments, are drawn in frequency order; if ordFreq=FALSE, glossaries are drown in alphabetic order (by default TRUE)
stop.word.tm	if TRUE, the tm stopwords (if the words are selected in TextData object) are not considered for the barchart (by default FALSE)
idiom	declared idiom for the textual column(s) (by default English "en", see IETF language in package NLP)
stop.word.user	the user's stopwords (if the words are selected in TextData object) are not considered for the barchart (by default NULL)
theme	theme settings (see ggplot2 package; by default theme_bw())
title	title of the graph (by default NULL and the title is automatically assigned)
xtitle	x title of the graph (by default NULL and the x title is automatically assigned)
col.fill	background color for the barChart bars (by default grey)
col.lines	lines color for the barChart bars (by default black)
text.size	text font size (by default 12)
freq	add frequencies to word and document barplots, see details (by default NULL)
vline	if "YES" or TRUE add vertical line to barplot, see details (by default NULL)
interact	if FALSE a ggplot graph, if TRUE an interactive plotly graph, see details (by default FALSE)
round.dec	number of decimals (by default 4)
...	further arguments passed to or from other methods...

Details

freq adds frequencies to barplot (by default NULL). If "YES" or TRUE displays the frequencies at the right of the bars at +5 position. Numerical values display the frequencies at the right positions (positive values) or at the left (negative values).

vline adds two vertical line to word and document barplot (by default NULL). If TRUE a first vertical row line is added at mean level computed from the selected items from TextData, and a second vertical blue line with the frequency mean of words/documents selected to plot in plot.TextData. If row and blue lines are the same, only blue line is shown. If vline is a number, a line is show with this value.

Barchart selected in sel argument (doc, word and/or repeated segments) is in ggplot format. Barchart is used with geom_bar function of ggplot package. If it is only one element in sel argument the plot can be saved in ggplot format: newobject <- plot(TextDataObject,sel="word")

Selection of docs, words or segments can be done by numbers sel=list(type="doc", select=c(1,2:4,6)) or names sel= list(type="doc", select=c("M31_55", "M>55")).

If interact, rank for words/docs/segments from TextData selection are shown.

Author(s)

Ramón Alvarez-Esteban <ramon.alvarez@unileon.es>, Mónica Bécue-Bertaut, Josep-Antón Sánchez-Espigares

See Also

[TextData](), [print.TextData](), [summary.TextData]()

Examples

```
# Non aggregate analysis

 data(open.question)
 res.TD<-TextData(open.question, var.text=c(9,10), remov.number=TRUE, Fmin=10, Dmin=10,
 stop.word.tm=TRUE, context.quali=c("Gender","Age_Group","Education"), context.quanti=c("Age"))
 plot(res.TD)

# Aggregate analysis
 data(open.question)
 res.TD<-TextData(open.question, var.text=c(9,10), var.agg="Gen_Age", remov.number=TRUE,
 Fmin=10, Dmin=10, stop.word.tm=TRUE, context.quali=c("Gender","Age_Group","Education"),
 context.quanti=c("Age"), segment=TRUE)
 plot(res.TD)
```

---

print.LexCA   *Print LexCA objects*

Description

Prints the Textual Correspondence Analysis (CA) results from a LexCA object

Usage

```
## S3 method for class 'LexCA'
print(x, file = NULL, sep=";", ...)
```

Arguments

x	object of LexCA class
file	a connection, or a character string giving the name of the file to print to (in csv format). If NULL (the default), the results are not printed in a file
sep	character to insert between the objects to print (if the argument file is non-NULL) (by default ";")
...	further arguments passed to or from other methods

Author(s)

Ramón Alvarez-Esteban <ramon.alvarez@unileon.es>, Mónica Bécue-Bertaut, Josep-Antón Sánchez-Espigares

See Also

LexCA, plot.LexCA, summary.LexCA, TextData

Examples

```
data(open.question)
res.TD<-TextData(open.question,var.text=c(9,10), var.agg="Age_Group", Fmin=10, Dmin=10,
        remov.number=TRUE, stop.word.tm=TRUE)
res.LexCA<-LexCA(res.TD,lmd=0,lmw=1)
print(res.LexCA)
```

---

print.LexChar	*Print LexChar objects*

Description

Prints characteristic words and documents from LexChar objects

Usage

```
## S3 method for class 'LexChar'
print(x, file = NULL, sep=";", dec=".",  ...)
```

Arguments

x	object of LexChar class
file	a connection, or a character string giving the name of the file to print to (in csv format). If NULL (the default), the results are not printed in a file
sep	character to insert between the objects to print (if the argument file is non-NULL) (by default ";")
dec	decimal point (by default ".")
...	further arguments passed to or from other methods

Author(s)

Ramón Alvarez-Esteban <ramon.alvarez@unileon.es>, Mónica Bécue-Bertaut, Josep-Antón Sánchez-Espigares

See Also

LexChar, plot.LexChar

Examples

```
data(open.question)
res.TD<-TextData(open.question, var.text=c(9,10), var.agg="Gen_Edu", Fmin=10, Dmin=10,
        stop.word.tm=TRUE)
LD<-LexChar(res.TD, maxCharDoc = 0)
print(LD)
```

---

print.TextData   *Print TextData objects*

Description

Print statistical results for documents, words and segments from TextData objects, in alphabetical and frequency order.

Usage

```
## S3 method for class 'TextData'
print(x, file = NULL, sep=";", ...)
```

Arguments

x	object of TextData class
file	connection, or character string giving the name of the file to print to (in csv format). If NULL (by default value), the results are not printed in a file
sep	character inserted between the objects to print (if file argument is non-NULL) (by default ";")
...	further arguments passed to or from other methods

Author(s)

Ramón Alvarez-Esteban <ramon.alvarez@unileon.es>, Monica Bécue-Bertaut, Josep-Antón Sánchez-Espigares

See Also

[TextData](), [plot.TextData](), [summary.TextData]()

Examples

```
data(open.question)
res.TD<-TextData(open.question, var.text=c(9,10), remov.number=TRUE, Fmin=10, Dmin=10,
stop.word.tm=TRUE, context.quali=c("Gender","Age_Group","Education"),
   context.quanti=c("Age"))
print(res.TD)
```

---

summary.LexCA   *Summary LexCA object*

Description

Summarizes LexCA objects

Usage

```
## S3 method for class 'LexCA'
summary(object, ncp=5, nb.dec = 3, ndoc=10, nword=10, nseg=10,
 nsup=10, metaDocs=FALSE, metaWords=FALSE, file = NULL, ...)
```

Arguments

object	object of LexCA class
ncp	number of dimensions to be printed (by default 5)
nb.dec	number of decimal digits to be printed (by default 3)
ndoc	number of documents whose coordinates are listed (by default 10). Use ndoc="ALL" to have the results for all the documents. Use ndoc=0 or ndoc=NULL if the results for documents are not wanted.
nword	number of words whose coordinates are listed (by default 10). Use nword="ALL" to have the results for all the words. Use nword=0 or nword=NULL if the results for words are not wanted
nseg	number of repeated segments whose coordinates are listed (by default 10). Use nseg="ALL" to have the results for all the segments. Use nseg=0 or nseg=NULL if the results for segments are not wanted
nsup	number of supplementary elements whose coordinates are listed (by default 10). Use nsup="ALL" to have the results for all the elements. Use nsup=0 or nsup=NULL if the results for the supplementary elements are not wanted
metaDocs	axis by axis, the highest contributive documents are listed, separately for negative-part and positive-part documents; these documents have been identified in LexCA, taking into account lmd value (by default FALSE)
metaWords	axis by axis, the highest contributive words are listed, separately for negative-part and positive-part words; these words have been identified in LexCA, taking into account lmw value (by default FALSE)
file	a connection, or a character string naming the file to print to (csv format). If NULL (the default), the results are not printed in a file
...	further arguments passed from other methods

Author(s)

Ramón Alvarez-Esteban <ramon.alvarez@unileon.es>, Monica Bécue-Bertaut, Josep-Antón Sánchez-Espigares

See Also

[LexCA](), [print.LexCA](), [plot.LexCA]()

Examples

```
data(open.question)
res.TD<-TextData(open.question, var.text=c(9,10), Fmin=10, Dmin=10, stop.word.tm=TRUE)
res.LexCA<-LexCA(res.TD, lmd=1, lmw=1)
summary(res.LexCA)
```

summary.LexChar   *Summary LexChar object*

Description

Summarizes LexChar objects

Usage

```
## S3 method for class 'LexChar'
summary(object, CharWord=TRUE, stats=TRUE, CharDoc=TRUE, Vocab=TRUE,
    file = NULL, ...)
```

Arguments

object	object of TextData class
CharWord	if TRUE characteristic words of all the documents are shown (by default TRUE)
stats	if TRUE association statistics of lexical table are shown (by default TRUE)
CharDoc	if TRUE characteristic source-documents of all the aggregate-documents are shown (by default TRUE)
Vocab	if TRUE characteristic quantitative and qualitative variables of the words. CharWord and stats are provide
file	a connection, or a character string naming the file to print to in csv format. If NULL (the default), the results are not printed in a file
...	further arguments passed to or from other methods,...

Details

Vocab$quali$CharWord provides the qualitative variables and their categories. Vocab$quali$stats provides association statistics for vocabulary and qualitative variables. Vocab$quanti$CharWord provides characteristic quantitative variables for each word. This summary.LexChart function provides the characteristic words for each quantitative variable. Vocab$quali$stats provides statistics for vocabulary and quantitative variables.

Author(s)

Ramón Alvarez-Esteban <ramon.alvarez@unileon.es>, Monica Bécue-Bertaut, Josep-Antón Sánchez-Espigares

See Also

LexChar, print.LexChar, plot.LexChar

Examples

```
data(open.question)
res.TD<-TextData(open.question, var.text=c(9,10), var.agg="Gen_Edu", Fmin=10, Dmin=10,
        remov.number=TRUE, stop.word.tm=TRUE)
res.LexChar <- LexChar(res.TD)
summary(res.LexChar)
```

---

summary.TextData *Summary of TextData objects*

Description

Summarizes TextData objects.

Usage

```
## S3 method for class 'TextData'
summary(object, ndoc=10, nword=50, nseg=50, ordFreq = TRUE, file = NULL, sep=";",
    info=TRUE,...)
```

Arguments

object	object of TextData class
ndoc	statistical report on the first ndoc documents (by default 10). Use ndoc="ALL" to have the results for all the documents. Use ndoc=0 or ndoc=NULL if the results on the documents are not wanted
nword	index of the nword first words (by default 50). Use nword="ALL" to have the complete index. Use nword=0 or nword=NULL if the results on the words are not wanted
nseg	index of the nfirst nseg repeated segments (by default 50). Use nseg="ALL" to have the complete list of segments. Use nseg=0 or nseg=NULL if the results on the segments are not wanted
ordFreq	if ordFreq=TRUE, glossaries of words and repeated segments, are listed in frequency order; if ordFreq=FALSE, glossaries are listed in alphabetic order (by default TRUE)
file	a connection, or a character string naming the file to print to in csv format. If NULL (the default), the results are not printed in a file
sep	character string to insert between the objects to print (if the argument file is not NULL) (by default ";")
info	if TRUE the selection criteria of the words are shown(by default TRUE)
...	further arguments passed to or from other methods,...

Author(s)

Ramón Alvarez-Esteban <ramon.alvarez@unileon.es>, Monica Bécue-Bertaut, Josep-Antón Sánchez-Espigares

See Also

[TextData](), [print.TextData](), [plot.TextData]()

Examples

```
# Non aggregate analysis
data(open.question)
res.TD<-TextData(open.question, var.text=c(9,10), remov.number=TRUE, Fmin=10, Dmin=10,
 stop.word.tm=TRUE, context.quali=c("Gender","Age_Group","Education"), context.quanti=c("Age"))
summary(res.TD)

# Aggregate analysis and repeated segments
data(open.question)
res.TD<-TextData(open.question, var.text=c(9,10), var.agg="Gen_Age", remov.number=TRUE,
 Fmin=10, Dmin=10, stop.word.tm=TRUE, context.quali=c("Gender","Age_Group","Education"),
 context.quanti=c("Age"), segment=TRUE)
summary(res.TD)
```

TextData   *Building textual and contextual tables (TextData)*

Description

Creates a textual and contextual working-base (TextData format) from a source-base (data frame format).

Usage

```
TextData(base, var.text=NULL, var.agg=NULL, context.quali=NULL, context.quanti= NULL,
 selDoc="ALL", lower=TRUE, remov.number=TRUE,lminword=1, Fmin=Dmin,Dmin=1, Fmax=Inf,
 stop.word.tm=FALSE, idiom="en", stop.word.user=NULL, segment=FALSE,
 sep.weak="default",
 sep.strong="\u005B()\u00BF?./:\u00A1!=;{}\u005D\u2026", seg.nfreq=10, seg.nfreq2=10,
 seg.nfreq3=10, graph=FALSE)
```

Arguments

base	source data frame with at least one textual column
var.text	vector with index(es) or name(s) of the selected textual column(s) (by default NULL)
var.agg	index or name of the aggregation categorical variable (by default NULL)
context.quali	vector with index(es) or name(s) of the selected categorical variable(s) (by default NULL)
context.quanti	vector with index(es) or name(s) of the selected quantitative variable(s) (by default NULL)
selDoc	vector with index(es) or name(s) of the selected source-documents (rows of the source-base) (by default "ALL")
lower	if TRUE, the corpus is converted into lowercase (by default TRUE)
remov.number	if TRUE, numbers are removed (by default TRUE)
lminword	minimum length of a word to be selected (by default 1)
Fmin	minimum frequency of a word to be selected (by default Dmin)
Dmin	a word has to be used in at least Dmin source-documents to be selected (by default 1)

Fmax	maximum frequency of a word to be selected (by default Inf)
stop.word.tm	if TRUE, stoplist automatically provided in accordance with the idiom (by default FALSE)
idiom	declared idiom for the textual column(s) (by default English "en", see IETF language in package NLP)
stop.word.user	stoplist provided by the user
segment	if TRUE, the repeated segments are identified (by default FALSE)
sep.weak	string with the characters marking out the terms (by default punctuation characters, space and control). See details
sep.strong	string with the characters marking out the repeated segments (by default "[()??./:?!=+;-]\")
seg.nfreq	minimum frequency of a more-than-three-words-long repeated segment (by default 10)
seg.nfreq2	minimum frequency of a two-words-long repeated segment (by default 10)
seg.nfreq3	minimum frequency of a three-words-long repeated segment (by default 10)
graph	if TRUE, documents, words and repeated segments barcharts are displayed; use plot.TextData to use more options (by default FALSE)

Details

Each row of the source-base is considered as a source-document. TextData function builds the working-documents-by-words table, submitted to the analysis.

sep.weak contains the string with the characters marking out the terms (by default punctuation characters, space and control). Backslash or double backslash are used to start an escape sequence defining special characters. Each special character must by separated the symbol | (or) in sep.weak and sep.strong. The default is:
sep.weak = ("[%`:*$&#/^|<=>;'+@.,~?(){}|[[:space:]]|
\u2014|\u002D|\u00A1|\u0021|\u00BF|\u00AB|\u00BB|\u2026|\u0022|\u005D|\u0097")

Some special characters can be introduced as unicode characters. Back slash (escape contol) is not allowed.

Information related to context.quanti and context.quali arguments:

1. If numeric, contextual variables can be included in both vectors. The function TextData converts the numeric variable into factor to include it in context.quali vector. This possibility is interesting in some cases. For example, when treating open-ended questions, we can be interested in computing the correlation between the contextual variable "Age" and the axes and, at the same time, to draw the trajectory of the different values of "Age" (year by year) on the CA maps.

2. In the case of one or several columns with textual data not selected in vector var.text, if the argument context.quali is equal to "ALL", these columns will be considered as categorical variables.

Non-aggregate table versus aggregate table.

If var.agg=NULL:

1. The work-documents are the non-empty-source-documents.

2. DocTerm: non-aggregate lexical table with:

           as many rows as non-empty source-documents
           as many columns as words are selected.

3. context$quali: data frame crossing the non-empty source-documents (rows) and the categorical contextual-variables (columns).

4. context$quanti: data frame crossing the non-empty source-documents (rows) and the quantitative contextual-variables (columns). Both contextual tables can be juxtaposed row-wise to DocTerm table.

If var.agg is NON-NULL:

1. The work-documents are aggregate-documents, issued from aggregating the source-documents depending on the categories of the aggregation variable; the aggregate-documents inherit the names of the corresponding categories.

2. DocTerm is an aggregate table with:

           as many rows as as categories the aggregation variable has
           as many columns as words are selected.

3. context$quali$qualitable: juxtaposes as many supplementary aggregate tables as categorical contextual variables. Each table has:

        as many rows as categories the contextual categorical variable has
        as many columns as selected words, i.e. as many columns as DocTerm has.

4. context$quali$qualivar: names of categories of the supplementary categorical variables.

5. context$quanti: data frame crossing the working aggregate-documents (rows) and the quantitative contextual-variables (columns). The value for an active aggregate-document is the mean-value of the source-documents belonging to this aggregate-document.

Value

A list including:

summGen	general summary
summDoc	document summary
indexW	index of words
DocTerm	working lexical table (non-aggregate or aggregate table depending on var.agg value); working-documents by words table in slam package compressed format
context	contextual variables if context.quali or context.quanti are non-NULL; the structure greatly differs in accordance with the nature of DocTerm table (non-aggregate/ aggregate), see details
info	information about the selection of words
var.agg	a one-column data frame with the values of the aggregation variable; NULL if non-aggregate analysis
SourceTerm	in the case of DocTerm being an aggregate analysis, the source-documents by words table is kept in this data structure, in slam package compressed format
indexS	working-documents by repeated-segments table, in slam package compressed format
remov.docs	vector with the names of the removed empty source-documents
VCr	Cramer's V coefficient of document x term matrix
Inertia	total inertia of document x term matrix

Author(s)

Ramón Alvarez-Esteban <ramon.alvarez@unileon.es>, Monica Bécue-Bertaut, Josep-Antón Sánchez-Espigares

References

Lebart, L., Salem, A., & Berry, L. (1998). Exploring textual data. (D. Kluwer, Ed.). doi:10.1007/9789401715256.

See Also

print.TextData, summary.TextData, plot.TextData

Examples

```
# Non aggregate analysis
data(open.question)
res.TD<-TextData(open.question, var.text=c(9,10), remov.number=TRUE, Fmin=10, Dmin=10,
 stop.word.tm=TRUE, context.quali=c("Gender","Age_Group","Education"), context.quanti=c("Age"))


# Aggregate analysis and repeated segments
data(open.question)
res.TD<-TextData(open.question, var.text=c(9,10), var.agg="Gen_Age", remov.number=TRUE,
 Fmin=10, Dmin=10, stop.word.tm=TRUE, context.quali=c("Gender","Age_Group","Education"),
 context.quanti=c("Age"), segment=TRUE)
```

Index