Package 'Xplortext'

July 5, 2019

Type Package

Encoding UTF-8

Title Statistical Analysis of Textual Data

Version 1.2.1

Date 2019-07-04

Author Mónica Bécue-Bertaut, Ramón Alvarez-Esteban, Josep-Anton Sánchez-Espigares

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

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://www.xplortext.org> for examples.

License GPL (>= 2.0)

Depends R (>= 3.4.0), FactoMineR(>= 1.36), ggplot2(>= 2.2.1), tm(>= 0.7-3)

Imports ggdendro, graphics, gridExtra, MASS, methods, stringi, stringr, slam, stats, utils, flexclust, flashClust

URL http://www.xplortext.org

RoxygenNote 6.1.1

NeedsCompilation no

Repository CRAN

Date/Publication 2019-07-05 07:40:03 UTC

R topics documented:

`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 `http://www.xplortext.org` for examples.

Details

Package:	Xplortext
Type:	Package
Version:	1.2.1
Date:	2019-07-04
License:	GPL (>=2.0)

Author(s)

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

References

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 http://www.xplortext.org

`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), 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
`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, 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. Supplementary 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. doi: 10.1201/9781420034943.

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

Characteristic words of documents from TextData objects.

Usage

```
LexChar(object, proba=0.05, maxDocs=20, maxCharDoc=10, maxPrnDoc=100)
```

Arguments

object	TextData object
proba	threshold on the p-value used when selecting the characteristic words (by default 0.05)
maxDocs	maximum number of documents in the working lexical table (by default 20). See details
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)

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. Extracting the characteristic words for a too high number of documents is of no interest and time-consuming. So that, this function can be applied only when the number of documents in the lexical table is under or equal to maxDocs (by default 20). In the case of aggregate documents, extracting the characteristic source-documents is possible but of interest only if the source-documents are not too long. In any case, only the first maxPrnDoc characters of each characteristic document are printed (by default 100).

Value

Returns a list including:

`CharWord`	characteristic words of all the documents
`CharDoc`	characteristic source-documents of all the aggregate-documents

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`

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)
LexChar(res.TD)
```

`LexCHCca` *Chronogically Constrained Agglomerative Hierarchical Clustering on Correspondence Analysis Components (LexCHCca)*

Description

Chronogically constrained agglomerative hierarchical clustering on a corpus of documents.

Usage

```
LexCHCca (object, nb.clust=0, min=3, max=NULL, nb.par=5, graph=TRUE, proba=0.05)
```

Arguments

object	object of LexCA class
nb.clust	number of clusters (see details). If 0, the tree is cut at the level the user clicks on. If -1, 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 (by default 3)
max	maximum number of clusters (by default NULL and 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 in selecting the characteristic words of the clusters and in selecting the axes when describing the clusters by the axes (by default 0.05)

Details

LexCHCca starts from the documents coordinates on textual correspondence analysis axes. The hierarchical tree is built taking into account that only chronological contiguous nodes can be grouped. The documents have to be ranked in the lexical table in the chronological order. Euclidean metric and complete linkage method are used.

The number of clusters is determined either a priori or from the constrained hierarchical tree structure. If nb.clust=0, a level for cutting the tree is automatically suggested. This is computed in the following way, reading the tree downward. At a given step, the tree could be cut into Q clusters (Q varying between min and max). The distance between the two nodes that are no longer grouped together using complete linkage method when passing from Q-1 to Q clusters and the distance between the two nodes that are no longer grouped together when passing from Q to Q+1 are computed. The suggested level corresponds to the maximum value of the ratio between the former and the latter of these values. These distances correspond to the criterion value when building the tree bottom up. The user can choose to cut the tree at this level or at another one.

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

The tree is plotted jointly with a barchart of the successive values of the aggregation criterion.

Value

Returns a list including:

data.clust	the original active lexical table with a supplementary column called clust containing the partition
desc.word	description of the clusters by their characteristic words
desc.axes	description of the clusters by the characteristic axes
call	list or parameters and internal objects
desc.doc	labels of the paragon (para) and specific documents (dist) of each cluster
dendro	list with the succession of nodes that are found when reading the tree downward

Returns the graphs with the tree and the correspondence analysis map where the documents are colored according to the cluster they belong to (2D).

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.

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, COMPSTAT Lectures.

See Also

plot.LexCHCca, LabelTree, 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)
```

LexHCca — *Hierarchical Clustering of Documents on Textual Correspondence Analysis Coordinates (LexHCca)*

Description

Agglomerative hierarchical clustering on a corpus of documents.

Usage

```
LexHCca(object, nb.clust=0, consol=FALSE, iter.max=10, min=3, max=NULL,
  kk=Inf, order=TRUE, graph=TRUE, proba=0.05,cluster.CA="rows",description=TRUE,
   nb.par=0,size.par=80,marg.doc=FALSE,seed=12345,...)
```

Arguments

object	object of LexCA class
nb.clust	number of clusters (see details). If 0, the tree is cut at the level the user clicks on. If -1, the tree is automatically cut at the suggested level. If a (positive) integer, the tree is cut with nb.clust clusters (by default 0)
consol	if TRUE, consolidation is performed after hierachical clustering (by default FALSE)
iter.max	maximum number of iterations in the consolidation step (by default 10)
min	minimum number of clusters (by default 3)
max	maximum number of clusters (by default NULL and then max is computed as the minimum between 10 and the number of documents divided by 2)
kk	An integer corresponding to the number of clusters used in a Kmeans preprocessing before the hierarchical clustering; the top of the hierarchical tree is then constructed from this partition. This is very useful if the number of individuals is high. Note that consolidation cannot be performed if kk is different from Inf and some graphics are not drawn. Inf is used by default and no preprocessing is done, all the graphical outputs are then given.
order	if TRUE, the clusters are numbered depending on the coordinate of their centroid on the first axis (by default TRUE)
graph	if TRUE, graphs are displayed (by default TRUE)
proba	threshold on the p-value used in selecting words, documents, axes and contextual variables when describing the clusters (by default 0.05)
cluster.CA	if 'rows' or 'docs' cluster is performed with documents; 'columns' or 'words' with words (by default 'rows')
description	if TRUE, description of the clusters by their characteristic words/documents, by the characteristic axes and by contextual variables if considered in LexCA (by default TRUE)
nb.par	number of edited paragons (para) and specific documents (dist) (by default 0)
size.par	text size of edited paragons (para) and specific documents (dist) (by default 80)
marg.doc	if FALSE, frequencies before TextData selection are the marginal frequencies for documents in description analysis, TRUE if frequencies after TextData selection (by default FALSE)
seed	Seed to obtain the same results using k-means (by default 12345)
...	other arguments from other methods

Details

LexHCca starts from the documents coordinates on textual correspondence analysis axes. Euclidean metric and Ward method are used.

The number of clusters is determined either a priori or from the hierarchical tree structure. If nb.clust=0, a level for cutting the tree is automatically suggested. This is computed in the following way, reading the tree downward. At a given step, the tree could be cut into Q clusters (Q varying between min and max). The between-inertia gain when passing from Q-1 to Q clusters and the

between-inertia gain when passing from Q to Q+1 clusters are computed. The suggested level corresponds to the maximum value of the ratio between the former and the latter of these inertia-gains. Note that the between-inertia gain when passing from Q to Q+1 clusters is equal to the value of the Ward criterion when passing from Q+1 to Q clusters when building the tree bottom up. In this latter case, a level where to cut the tree is suggested. The user can choose to cut the tree at this level or at another one.

The results include a thorough description of the clusters, taking into account contextual variables. Graphs are provided.

Value

Returns a list including:

`data.clust`	the original active lexical table used in LexCA plus a new column called clust containing the partition
`centers`	coordinates of centers from LexCA results for each cluster
`clust.count`	count of documents/words belonging to each cluster and some statistics
`clust.content`	list of the document/word labels according to the cluster they belong to
`ss`	total sum of squares
`call`	list of internal objects. `call$t` giving the results for the hierarchical tree; See the first reference for more details
`desc.axes`	description of the clusters by the characteristic axes
`desc.wordvar`	if description=TRUE, description of the clusters by their characteristic words, supplementary words and, if contextual variables were considered in LexCA, description of the partition/clusters by these variables
`desc.doc`	if description=TRUE, description of the clusters by their characteristic documents
`wordslabels`	labels of the paragon (para) and specific words (dist) of each cluster
`docslabels`	labels of the paragon (para) and specific documents (dist) of each cluster
`docspara`	if nb.par>0, description of the clusters by the nb.par "para" documents writing the first size.par characters of the literal text
`docsdist`	if nb.par>0, description of the clusters by the nb.par "dist" documents writing the first size.par characters of the literal text

Returns the hierarchical tree with a barplot of the successive inertia gains, the CA map of the documents enriched by the tree (3D), the CA map with the document labels colored according to their cluster (2D).

Author(s)

Monica Bécue-Bertaut, Ramón Alvarez-Esteban <ramon.alvarez@unileon.es>, Josep-Anton 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., Salem, A., & Berry, L. (1998). Exploring textual data. (D. Kluwer, Ed.). doi: 10.1007/9789401715256.

See Also

LexCA

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=TRUE, nb.clust=5, order=TRUE)
```

---

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, ...)
```

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 NULL 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 TRUE)
...	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, ...)
```

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)
`...`	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

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), choice="tree", 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))
choice	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, choice="tree")
plot(res.chcca, choice="map")
plot(res.chcca, choice="bar", max.plot=5)

## End(Not run)
```

---

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, stop.word.tm=FALSE,
  stop.word.user=NULL, theme=theme_bw(), title=NULL, xtitle=NULL, col.fill="grey",
  col.lines="black", text.size=12, freq=NULL, vline=NULL,...)
```

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 the barchart are draw)
stop.word.tm	the tm stopwords are not considered for the barchart (by default FALSE)
stop.word.user	the user's stopwords 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)
...	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 a vertical line to barplot (by default NULL). If TRUE a vertical line is added at mean level. If "median" a vertical line is added at median level.

Author(s)

Ramón Alvarez-Esteban <ramon.alvarez@unileon.es>, Monica 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=";", ...)
```

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 ";")
`...`	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.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=";",
   ...)
```

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 ";")
...	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="([']|[[:punct:]]|[[:space:]]|[[:cntrl:]])+",
 sep.strong="\u005B()\u00BF?./:\u00A1!=+;{}-\u005D", 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. For example:
sep.weak = "[[space]]|¡|!|;|\,|\.|\-|\(|\)|#|:|¿|\'|%|_|\u0022"

Some special characters can be introduced as unicode characters.

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

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