Package ‘deepMOU’

March 4, 2021

Title Clustering of Short Texts by Mixture of Unigrams and Its Deep Extensions

Version 0.1.1

Description Functions providing an easy and intuitive way for fitting and clusters data using the Mixture of Unigrams models by means the Expectation-Maximization algorithm (Nigam, K. et al. (2000). <doi:10.1023/A:1007692713085>), Mixture of Dirichlet-Multinomials estimated by Gradient Descent (Anderlucci, Viroli (2020) <doi:10.1007/s11634-020-00399-3>) and Deep Mixture of Multinomials whose estimates are obtained with Gibbs sampling scheme (Viroli, Anderlucci (2020) <doi:10.1007/s11222-020-09989-9>). There are also functions for graphical representation of clusters obtained.

License GPL-3

Encoding UTF-8

LazyData true

RoxygenNote 7.1.1

Imports skmeans, extraDistr, dplyr, Rfast, entropy, ggplot2, graphics, MASS

NeedsCompilation no

Author Martin D'Ippolito [aut, cre],
Anderlucci Laura [aut],
Cinzia Viroli [aut]

Maintainer Martin D'Ippolito <martinmy69@gmail.com>

Depends R (>= 3.5.0)

Repository CRAN

Date/Publication 2021-03-04 02:30:02 UTC

R topics documented:

`Abstracts` *Abstracts dataset*

Description

Dataset composed by titles and abstracts of articles published in 2015 in five statistical journals.

Usage

```
data(Abstracts)
```

Format

A matrix for the bag-of-words representation of the Abstracts dataset; terms are sorted in alphabetical order.

Details

These are the titles and abstracts of all the articles published in 2015 by the following journals: - Journal of American Statistical Association (JASA) - Journal of the Royal Statistical Society - Series B - Annals of Statistics - Biometrika - Statistical Science

The dataset comprises 379 articles with a vocabulary of 606 words already pre-processed (stemmed, lemmatized, stopwords removal etc.); terms with entropy less than 0.3 were discarded (rule-of-thumb threshold).

Examples

```
x = data(Abstracts)
print(head(x))
```

bubble_clust *Bubble plot*

Description

Bi-dimensional representation (via Multi-Dimensional Scaling) of the clusters, where each bubble corresponds to a cluster, its size is proportional to the relative frequency of documents and color saturation reflects cluster cohesion.

Usage

```
bubble_clust(
  x,
  clusters,
  bubble_size = 1,
  bubble_col = c("red", "white"),
  cex_text = 1,
  main = NULL
)
```

Arguments

`x` Document-term matrix describing the frequency of terms that occur in a collection of documents. Rows correspond to documents in the collection and columns correspond to terms.

`clusters` Integer vector of length of the number of cases, which indicates a clustering. The clusters have to be numbered from 1 to the number of clusters.

`bubble_size` Graphical parameter for bubbles size.

`bubble_col` Choose palette with two colors (default `"red"` and `"white"`). The first (darker) color will denote homogeneous clusters, while the latter (lighter) more heterogeneous ones.

`cex_text` Size of texts inside bubbles.

`main` A title for the plot.

Value

A graphical aid to visualize and to describe the obtained clusters.

Examples

```
# Load the CNAE2 dataset
data("CNAE2")

# Perform parameter estimation and clustering
mou_CNAE2 = mou_EM(x = CNAE2, k = 2)
```

```
# Usage of the function
bubble_clust(mou_CNAE2$x,mou_CNAE2$clusters, bubble_size = 5 )
```

cl_CNAE *Classification labels of the CNAE2 data set*

Description

True classification labels of documents from the CNAE2 data set.

Usage

```
data(CNAE2)
```

Format

A vector containing the true classification labels of documents of CNAE2 data set.

Source

See CNAE2 for further details.

CNAE2 *CNAE dataset on classes 4 and 9*

Description

The data set CNAE2 is a subset of the original CNAE-9 data, that comprises 1080 documents categorized into 9 topics of free text business descriptions of Brazilian companies.

Specifically, CNAE2 contains only the documents belonging to topics "4" and "9". The data set is already pre-processed and provides the bag-of-words representation of the documents; the columns with null counts are removed leading to a matrix with 240 documents on a vocabulary with cardinality 357. This data set is highly sparse (98

Class labels are stored in cl_CNAE

Usage

```
data(CNAE2)
```

Format

A matrix for the bag-of-words representation of the CNAE2 dataset.

Source

Original CNAE9 dataset

Examples

```
x = data(CNAE2)
print(head(x))
```

deep_mou_gibbs *Deep Mixture of Unigrams*

Description

Performs parameter estimation by means of Gibbs sampling and cluster allocation for the Deep Mixture of Unigrams.

Usage

```
deep_mou_gibbs(x, k, g, n_it = 500, seed_choice = 1, burn_in = 200)
```

Arguments

`x` Document-term matrix describing the frequency of terms that occur in a collection of documents. Rows correspond to documents in the collection and columns correspond to terms.

`k` Number of clusters/groups at the top layer.

`g` Number of clusters at the bottom layer.

`n_it` Number of Gibbs steps.

`seed_choice` Set seed for reproducible results.

`burn_in` Number of initial Gibbs samples to be discarded and not included in the computation of final estimates.

Details

Starting from the data matrix `x`, the Deep Mixture of Unigrams is fitted and `k` clusters are obtained. The algorithm for the estimation of the parameters is the Gibbs sampling. In particular, the function assigns initial values to all the parameters to be estimated. Then `n_it` samples for the parameters are obtained using conditional distributions on all the other parameters. The final estimates are obtained by averaging the samples given that initial `burn_in` samples are discarded. Clustering is eventually performed by maximizing the posterior distribution of the latent variables. For further details see the references.

Value

A list containing the following elements:

`x` The data matrix.

`clusters` the clustering labels.

`k` the number of clusters at the top layer.

g the number of clusters at the bottom layer.

numobs the sample size.

p the vocabulary size.

z1 the allocation variables at the top layer.

z2 the allocation variables at the bottom layer.

Alpha the estimates of Alpha parameters.

Beta the estimates of the Beta parameters.

pi_hat estimated probabilities of belonging to the k clusters at the top layer conditional to the g clusters at the bottom layer.

pi_hat_2 estimated probabilities of belonging to the g clusters at the bottom layer.

References

Viroli C, Anderlucci L (2020). "Deep mixtures of Unigrams for uncovering topics in textual data." *Statistics and Computing*, pp. 1-18. doi: 10.1007/s11222020099899.

Examples

```
# Load the CNAE2 dataset
data("CNAE2")

# Perform parameter estimation and clustering, very few iterations used for this example
deep_CNAE2 = deep_mou_gibbs(x = CNAE2, k = 2, g = 2, n_it = 5, burn_in = 2)

# Shows cluster labels to documents
deep_CNAE2$clusters
```

---

dir_mult_GD *Dirichlet-Multinomial mixture model by Gradient Descend algorithm*

Description

Performs parameter estimation by means of a Gradient Descend algorithm and cluster allocation for the Dirichlet-Multinomial mixture model.

Usage

```
dir_mult_GD(
  x,
  k,
  n_it = 100,
  eps = 1e-05,
  seed_choice = 1,
  KK = 20,
  min_iter = 2,
  init = NULL
)
```

Arguments

`x` Document-term matrix describing the frequency of terms that occur in a collection of documents. Rows correspond to documents in the collection and columns correspond to terms.

`k` Number of clusters/groups.

`n_it` Number of Gradient Descend steps.

`eps` Tolerance level for the convergence of the algorithm. Default is `1e-05`.

`seed_choice` Set seed for reproducible results.

`KK` Maximum number of iterations allowed for the nlminb function (see below).

`min_iter` Minimum number of Gradient Descend steps.

`init` Vector containing the initial document allocations for the initialization of the algorithm. If NULL (default) initialization is carried out via spherical k-means (skmeans).

Details

Starting from the data given by `x` the Dirichlet-Multinomial mixture model is fitted and `k` clusters are obtained. The algorithm for the parameter estimation is the Gradiend Descend. In particular, the function assigns initial values to weights of the Dirichlet-Multinomial distribution for each cluster and inital weights for the elements of the mixture. The estimates are obtained with maximum `n_it` steps of the Descent Algorithm algorithm or until a tolerance level `eps` is reached; by using the posterior distribution of the latent variable z, the documents are allocated to the cluster which maximizes the posterior distribution. For further details see the references.

Value

A list containing the following elements:

`x` The data matrix.

`clusters` the clustering labels.

`k` the number of clusters.

`numobs` the sample size.

`p` the vocabulary size.

`likelihood` vector containing the likelihood values at each iteration.

`pi_hat` estimated probabilities of belonging to the `k` clusters.

`Theta` matrix containing the estimates of the Theta parameters for each cluster.

`f_z_x` matrix containing the posterior probabilities of belonging to each cluster.

`AIC` Akaike Information Criterion (AIC) value of the fitted model.

`BIC` Bayesian Information Criterion (BIC) value of the fitted model.

References

Anderlucci L, Viroli C (2020). "Mixtures of Dirichlet-Multinomial distributions for supervised and unsupervised classification of short text data." *Advances in Data Analysis and Classification*, **14**, 759-770. doi: 10.1007/s11634020003993.

Examples

```
# Load the CNAE2 dataset
data("CNAE2")

# Perform parameter estimation and clustering, very
# few iterations are used for this example
dir_CNAE2 = dir_mult_GD(x = CNAE2, k = 2, n_it = 2)

# Shows cluster labels to documents
dir_CNAE2$clusters
```

---

heatmap_words *Heatmap of word frequencies by cluster*

Description

Displays the heatmap of the cluster frequency distributions of the most frequent terms sorted by the most informative ones.

Usage

```
heatmap_words(
  x,
  clusters,
  n_words = 50,
  legend_position = "bottom",
  font_size = 12,
  low_color = "grey92",
  top_color = "red",
  main = "Row frequencies of terms distribution",
  xlabel = NULL,
  ylabel = NULL,
  legend_title = "Entropy"
)
```

Arguments

x
: Document-term matrix describing the frequency of terms that occur in a collection of documents. Rows correspond to documents in the collection and columns correspond to terms.

clusters
: Integer vector of length of the number of cases, which indicates a clustering. The clusters have to be numbered from 1 to the number of clusters.

n_words
: Number of words to include in the heatmap (default is 50).

legend_position
: Position of the legend ("none", "left", "right", "bottom", "top", or two-element numeric vector as in theme). Default is "bottom".

`font_size` Text size in pts (default is 12).

`low_color` Base color for terms with no occurrence in a cluster (default is `"grey92"`)

`top_color` Base color for terms concentrated in a single cluster (default is `"red"`)

`main` A title for the plot. Default is `"Row frequencies of terms distribution"`.

`xlabel` A title for the x-axis. Default is `NULL`.

`ylabel` A title for the y-axis. Default is `NULL`.

`legend_title` A title for the legend. Default is `"Entropy"`.

Details

Takes as input the bag-of-words matrix and returns a heatmap displaying the row frequency distribution of terms according to the clusters. Words are sorted by entropy.

Value

A graphical aid to describe the clusters according to the most informative words.

Examples

```
# Load the CNAE2 dataset
data("CNAE2")

# Get document labels by clustering using mou_EM
mou_CNAE2 = mou_EM(x = CNAE2, k = 2)

# Usage of the function
heatmap_words(x = mou_CNAE2$x, clusters = mou_CNAE2$clusters)
```

`mou_EM` *Mixture of Unigrams by Expectation-Maximization algorithm*

Description

Performs parameter estimation by means of the Expectation-Maximization (EM) algorithm and cluster allocation for the Mixture of Unigrams.

Usage

```
mou_EM(x, k, n_it = 100, eps = 1e-07, seed_choice = 1, init = NULL)
```

Arguments

`x` Document-term matrix describing the frequency of terms that occur in a collection of documents. Rows correspond to documents in the collection and columns correspond to terms.

`k` Number of clusters/groups.

`n_it` Number of iterations for the Expectation-Maximization algorithm.

`eps` Tolerance level for the convergence of the algorithm. Default is `1e-07`.

`seed_choice` Set seed for reproducible results

`init` Vector containing the initial document allocations for the initialization of the algorithm. If NULL (default) initialization is carried out via spherical k-means (skmeans).

Details

Starting from the data given by `x` the Mixture of Unigrams is fitted and `k` clusters are obtained. The algorithm for the parameter estimation is the Expectation-Maximization (EM). In particular, the function assigns initial values to weights of the Multinomial distribution for each cluster and inital weights for the components of the mixture. The estimates are obtained with maximum `n_it` steps of the EM algorithm or until the tolerance level `eps` is reached; by using the posterior distribution of the latent variable z, the documents are allocated to the cluster which maximizes the posterior distribution. For further details see the references.

Value

A list containing the following elements:

`x` The data matrix.

`clusters` the clustering labels.

`k` the number of clusters.

`numobs` the sample size.

`p` the vocabulary size.

`likelihood` vector containing the likelihood values at each iteration.

`pi_hat` estimated probabilities of belonging to the `k` clusters.

`omega` matrix containing the estimates of the omega parameters for each cluster.

`f_z_x` matrix containing the posterior probabilities of belonging to each cluster.

`AIC` Akaike Information Criterion (AIC) value of the fitted model.

`BIC` Bayesian Information Criterion (BIC) value of the fitted model.

References

Nigam, K., McCallum, A., Thrun, S., Mitchell, T.: Text classification from labeled and unlabeled documents using EM. Machine learning 39, 103-134 (2000).

Examples

```
# Load the CNAE2 dataset
data("CNAE2")

# Perform parameter estimation and clustering
mou_CNAE2 = mou_EM(x = CNAE2, k = 2)

# Shows cluster labels to documents
mou_CNAE2$clusters
```

---

plot.deepMOU *Plotting method for "shallow" and deep mixtures of Unigrams and mixtures of Dirichlet-Multinomials*

Description

Bi-dimensional representation (via Multi-Dimensional Scaling) of the clusters, where each bubble corresponds to a cluster, its size is proportional to the relative frequency of documents and color saturation reflects cluster cohesion.

Usage

```
## S3 method for class 'deepMOU'
plot(
  x,
  y,
  bubble_size = 1,
  bubble_col = c("red", "white"),
  cex_text = 1,
  main = NULL,
  ...
)
```

Arguments

`x`	Output from `mou_EM`, `dir_mult_GD` or `deep_mou_gibbs`.
`y`	Parameter not used
`bubble_size`	Graphical parameter for bubbles size.
`bubble_col`	Choose palette with two colors (default `"red"` and `"white"`). The first (darker) color will denote homogeneous clusters, while the latter (lighter) more heterogeneous ones.
`cex_text`	Size of texts inside bubbles.
`main`	A main title for the plot.
`...`	Parameter not used

Details

The default graphical representation of `mou_EM`, `dir_mult_GD` and `deep_mou_gibbs` is the bubble plot. Namely, a bi-dimensional representation (via Multi-Dimensional Scaling) of the clusters, each bubble corresponds to a cluster, its size is proportional to the relative frequency of documents and color saturation reflects cluster cohesion.

Value

A graphical aid to visualize and to describe the obtained clusters.

Examples

```
# Load the CNAE2 dataset
data("CNAE2")

# Perform parameter estimation and clustering
mou_CNAE2 = mou_EM(x = CNAE2, k = 2)

# Usage of the function
plot(mou_CNAE2, bubble_size = 5 )
```

words_freq_plot *Graph of most frequent words of each cluster*

Description

Graphical plot of the most frequent words of each cluster

Usage

```
words_freq_plot(
  x,
  clusters,
  clust_label = NULL,
  n_words = 5,
  words_size = 2,
  axis_size = 1,
  set_colors = NA,
  main = "Most frequent words for each cluster",
  xlabel = "",
  ylabel = ""
)
```

Arguments

x — Document-term matrix describing the frequency of terms that occur in a collection of documents. Rows correspond to documents in the collection and columns correspond to terms.

clusters — Integer vector of length of the number of cases, which indicates a clustering. The clusters have to be numbered from 1 to the number of clusters.

clust_label — Vector of length of the number of cluster containing the cluster names to be displayed (by default "Cluster_1", "Cluster_2", ...).

n_words — Number of words to display.

words_size — A numerical value giving the amount by which plotting words should be magnified with respect to the default setting.

axis_size — Magnification to be used for axis annotation with respect to the default setting.

set_colors — Choose palette for word colors.

main — A title for the plot. Default is "Most frequent words for each cluster".

xlabel — A title for the x-axis. Default is empty.

ylabel — A title for the y-axis. Default is empty.

Details

The number of most frequent words to be shown can be set by n_words and also clusters names can be passed beforehand as a character vector to clust_label

Value

A graphical aid for visualizing the most frequent terms for each cluster.

Examples

```
# Load the CNAE2 dataset
data("CNAE2")

# Perform parameter estimation and clustering
mou_CNAE2 = mou_EM(x = CNAE2, k = 2)

# Usage of the function
words_freq_plot(mou_CNAE2$x, mou_CNAE2$clusters,n_words = 4, words_size = 2, main = "Example" )
```

Index