Package ‘textmatch’

April 11, 2019

Title Toolkit for Matching Textual Data and Evaluating Textual Similarity

Version 0.0.0.9000

Description What the package does (one paragraph).

Depends R (>= 3.5.2)

License What license is it under?

Encoding UTF-8

LazyData false

RoxygenNote 6.1.1.9000

Imports dplyr,
data.table,
quanteda,
stats,
progress

Suggests knitr,
rmarkdown

VignetteBuilder knitr

R topics documented:

calculate_pair_distances

Create a data frame of pairs of documents obtained through coarsened exact matching (CEM) within a specified number of bins and return indices for matched sets

Description

Create a data frame of pairs of documents obtained through coarsened exact matching (CEM) within a specified number of bins and return indices for matched sets

Usage

```
calculate_pair_distances(rep.list, Z, propensity.method = NULL,
  include = c("cosine", "euclidean", "mahalanobis", "propensity"),
  exclude = c("jaccard"))
```

Arguments

`rep.list` a named list of representations

`Z` A logical or binary vector indicating treatment and control for each unit in the study. TRUE or 1 represents a treatment unit, FALSE of 0 represents a control unit.

`propensity.method` Either GLM or MNIR for propensity score estimation

Value

A data.frame of indices for matched pairs of documents

FoxCNNcorpus

Corpus with 1,565 articles from CNN and 1,796 articles from Fox News.

Description

Corpus with 1,565 articles from CNN and 1,796 articles from Fox News.

Usage

```
FoxCNNcorpus
```

Format

A data.frame with 3,361 observations of 5 articles

Details

Corpus of front-page news articles published online by CNN or Fox News from 12/20/2014 to 05/09/2015 containing 1,565 articles from CNN and 1,796 articles from Fox News. Data contains article identifiers corresponding to data in FoxCNNmeta as well as raw and cleaned text data.

References

Mozer et al. (2019) Matching with Text Data: An Experimental Evaluation of Methods for Matching Documents and of Measuring Match Quality. *Political Analysis*, Forthcoming.

`FoxCNNmeta` *Dataset containing metadata for FoxCNN corpus with 1,565 articles from CNN and 1,796 articles from Fox News.*

Description

Dataset containing metadata for FoxCNN corpus with 1,565 articles from CNN and 1,796 articles from Fox News.

Usage

```
FoxCNNmeta
```

Format

A data.frame with 3,361 observations of 5 articles

Details

Metadata for the FoxCNN corpus including article names, original URLs, and dates of publication.

References

Mozer et al. (2019) Matching with Text Data: An Experimental Evaluation of Methods for Matching Documents and of Measuring Match Quality. *Political Analysis*, Forthcoming.

`FoxCNNsurvey` *Dataset containing distance measurements and average quality scores for sample of 505 pairs of matched Fox and CNN articles evaluated by human coders on Mechanical Turk. Raw quality scores are regressed on distance measurements to fit a predictive model for match quality as a function of the 117 metrics considered.*

Description

Dataset containing distance measurements and average quality scores for sample of 505 pairs of matched Fox and CNN articles evaluated by human coders on Mechanical Turk. Raw quality scores are regressed on distance measurements to fit a predictive model for match quality as a function of the 117 metrics considered.

Usage

```
FoxCNNsurvey
```

Format

A data.frame with 505 pairs and 108 features.

Details

Distance measurements and average quality scores for a sample of 505 matched pairs of documents evaluated by human coders.

References

Mozer et al. (2019) \"Matching with Text Data: An Experimental Evaluation of Methods for Matching Documents and of Measuring Match Quality\". *Political Analysis*, Forthcoming.

get_CEM *Create a data frame of pairs of documents obtained through coarsened exact matching (CEM) within a specified number of bins and return indices for matched sets*

Description

Create a data frame of pairs of documents obtained through coarsened exact matching (CEM) within a specified number of bins and return indices for matched sets

Usage

```
get_CEM(x, Z, rep.name, cuts, caliper_fun, verbose = FALSE, SR = NULL)
```

Arguments

`x`	a text representation
`Z`	a vector of treatment indicators
`rep.name`	a string or character with the name of the matching method
`cuts`	a function for how the variables will be binned. Defaults to "median"

Value

A data.frame of indices for matched pairs of documents

get_matches *Similarity and distance computation between documents or features*

Description

These functions compute distance matrices from a text representation where each row is a document and each column is a feature to measure distance over based on treatment indicator Z

Usage

```
get_matches(dist, Z, dist.name, caliper_fun, verbose = FALSE)
```

Arguments

`Z`	a vector of treatment indicators
`dist.name`	a string or character with the name of the matching method
`caliper_fun`	an optional function specifying the caliper to enforce when matching
`x`	a matrix of pairwise distances for all potential matches of treatment and control units. See pair_distances.

get_similarity_scores *This function calculates an input character vector's similarity matrix according to the measures contained in the predictive model.*

Description

This function calculates an input character vector's similarity matrix according to the measures contained in the predictive model.

Usage

```
get_similarity_scores(x)
```

Arguments

`x`	A character vector where each element is a document

Value

A data frame of rows (n * n-1) and columns 16; each column is one of the constituent similarity measures

`get_word2vec_glove` *This function calculates the Word2Vec embeddings*

Description

This function calculates the Word2Vec embeddings

Usage

```
get_word2vec_glove(dir.source, corpus)
```

Arguments

`dat`	FoxCNN corpus to calculate Word2Vec scores for

Value

A list of data frames containing the Word2Vec projections of the corpus

`makeMatches` *Create a data frame of matched pairs of documents and return indices for matched sets*

Description

Create a data frame of matched pairs of documents and return indices for matched sets

Usage

```
makeMatches(match.obj, Z)
```

Arguments

`match.obj`	a matched data set
`Z`	a vector of treatment indicators

Value

A data.frame of indices for matched pairs of documents `pair_distances` *Similarity and distance computation between documents or features*

Description

These functions compute distance matrices from a text representation where each row is a document and each column is a feature to measure distance over based on treatment indicator Z

Usage

```
pair_distances(dat, Z, propensity.method, all.counts = NULL,
  include = c("cosine", "euclidean", "mahalanobis", "propensity"),
  exclude = "jaccard", form = "data.frame", verbose = FALSE)
```

Arguments

`Z`	A logical or binary vector indicating treatment and control for each unit in the study. TRUE or 1 represents a treatment unit, FALSE of 0 represents a control unit.
`propensity.method`	Either GLM or MNIR for propensity score estimation
`form`	Should the distances be returned as a list of matrices or condensed into a single data frame?
`x`	a matrix text representation with rows corresponding to each document in a corpus and columns that represent summary measures of the text (e.g., word counts, topic proportions, etc.). Acceptable forms include a valid **quanteda** `dfm` object, a **tm** Document-Term Matrix, or a matrix of estimated topic proportions.

Value

A matrix showing pairwise distances for all potential matches of treatment and control units under various distance metrics

`quality_model` *Fitted model for pairwise match quality as a function of 117 distance metrics calculated in Mozer et al. (2019). Trained on "FoxCNNsurvey" dataset.*

Description

Fitted model for pairwise match quality as a function of 117 distance metrics calculated in Mozer et al. (2019). Trained on "FoxCNNsurvey" dataset.

Usage

```
quality_model
```

Format

A `glmnet` model object.

Details

Fitted model for predicting the match quality score for a given pair of text documents as a function of 117 distance measurements.

References

Mozer et al. (2019) \"Matching with Text Data: An Experimental Evaluation of Methods for Matching Documents and of Measuring Match Quality\". *Political Analysis*, Forthcoming.

`textmatch` *This function runs the main ML model as specified in Mozer et al. (2018)*

Description

This function runs the main ML model as specified in Mozer et al. (2018)

Usage

```
textmatch(x, outcome = "matrix", verbose = TRUE)
```

Arguments

`x`	A character vector or subset of the FoxCNN corpus where each element is a document

Value

A vector of predicted match quality scores on a scale of 0-10.

`transform_dfm` *Applies bounds, weights, and/or coarsening schemes to a dfm or document frequency matrix to reduce the dimension of the data, reduce noise, or apply other design rules (e.g. - to exclude words that occur in too few or too many documents).*

Description

Applies bounds, weights, and/or coarsening schemes to a dfm or document frequency matrix to reduce the dimension of the data, reduce noise, or apply other design rules (e.g. - to exclude words that occur in too few or too many documents).

Usage

```
transform_dfm(x, bounds, tfidf = FALSE, verbose = TRUE)
```

Arguments

`x`	a matrix text representation with rows corresponding to each document in a corpus and columns that represent summary measures of the text (e.g., word counts, topic proportions, etc.). Acceptable forms include a valid **quanteda** `dfm` object, a **tm** Document-Term Matrix, or a matrix of estimated topic proportions.
`bounds`	a vector of lower and upper bounds to enforce. Defaults to excluding any terms that appear in only one document and any terms that appear in every document
`tfidf`	optional scheme to use for weighting the DTM. Defaults to `FALSE`.
`verbose`	indicator for verbosity

Value

A bounded DFM

`transform_stm` *Refits a STM with a content-based covariate so that all document-level topic-proportions are estimated "as-treated". Also allows for calculation of the SR sufficient reduction and optional coarsening to reduce the dimension of the data, reduce noise, or apply other design rules (e.g. - to exclude words that occur in too few or too many documents).*

Description

Refits a STM with a content-based covariate so that all document-level topic-proportions are estimated "as-treated". Also allows for calculation of the SR sufficient reduction and optional coarsening to reduce the dimension of the data, reduce noise, or apply other design rules (e.g. - to exclude words that occur in too few or too many documents).

Usage

```
transform_stm(mod, out, Z, calc.SR = TRUE, coarsen = FALSE)
```

Arguments

`mod`	a fitted `stm` object
`out`	the original call to the STM
`Z`	an indicator for treatment assignment
`calc.SR`	an indicator for returning the sufficient reduction. Default is TRUE.
`coarsen`	an indicator for returning the coarsened STM

Value

A bounded DFM

Index