Package 'stmCorrViz'

July 24, 2016

Type Package

Title A Tool for Structural Topic Model Visualizations

Version 1.3

Date 2016-07-03

Imports jsonlite, stm, tm, SnowballC, stats, methods, utils

Depends R(>= 2.10.0)

Author Antonio Coppola, Margaret Roberts, Brandon Stewart, Dustin Tingley

Maintainer Antonio Coppola <acoppola@alumni.harvard.edu>

Description Generates an interactive visualization of topic correlations/ hierarchy in a Structural Topic Model (STM) of Roberts, Stewart, and Tingley. The package performs a hierarchical clustering of topics which are then exported to a JSON object and visualized using D3.

License GPL (>= 2)

RoxygenNote 5.0.1

NeedsCompilation no

Repository CRAN

Date/Publication 2016-07-24 09:28:14

R topics documented:

stmCorrViz-package *Hierarchical Correlation View of STM Models*

Description

Generates an interactive visualization of topic correlations/hierarchy in a Structural Topic Model (STM) of Roberts, Stewart, and Tingley. The package performs a hierarchical clustering of topics which are then exported to a JSON object and visualized using D3.

Details

Package:	stmCorrViz
Type:	Package
Version:	1.3
Date:	2016-07-03
License:	GPL (>= 2)

Author(s)

Antonio Coppola, Margaret E. Roberts, Brandon M. Stewart, and Dustin Tingley

Maintainer: Antonio Coppola <acoppola@alumni.harvard.edu>

References

Margaret E. Roberts, Brandon M. Stewart and Dustin Tingley (2014). stm: R Package for Structural Topic Models.

findThreshold *Find appropriate threshold range*

Description

This function performs a grid search over potential clustering thresholds to identify a valid range, and inspect the varying levels of aggregation within it.

Usage

```
findThreshold(mod, documents_raw=NULL, documents_matrix=NULL,
              range_min=.05, range_max=5, step=.05)
```

Arguments

mod	A fitted STM object from **stm**.
documents_raw	The raw documents used to generate the STM model. A character vector where each entry is the full text of a document.
documents_matrix	Document-term matrix representation of the raw documents, as generated by the prepDocuments function.
range_min	Lower bound of the range to be searched.
range_max	Upper bound of the range to be searched.
step	Step size for the grid search.

Value

A data frame containing the following columns:

1. *threshold*: Threshold value.
2. *valid*: Binary value; 1 if clustering is successful using given threshold; 0 if not.
3. *juncture_points*: Number of juncture points in the resulting clustering tree; -1 if run is unsuccessful. Lower threshold values yield a higher number of juncture points, corresponding to more binary splits and deeper trees. Higher threshold values produce fewer juncture points, corresponding to trees that have significant breadth rather than depth.

See Also

stmCorrViz

---

immigration_perceptions

Sample STM Model

Description

This is an example of a fitted STM object, alongside the raw document data used to fit the model. The model has 20 topics, 341 documents and a 455 word dictionary. For background about the underlying dataset, see gadarian.

Usage

```
data(immigration_perceptions)
```

Format

A list with the following elements:

model The STM model object.

raw_documents A character vector containing the raw documents.

documents_matrix Processed documents in bag-of-words matrix representation.

See Also

gadarian

---

stmCorrViz *Generate STM Correlation Tree*

Description

This function generates an interactive, full-model HTML visualization of topic hierachies for a fitted STM model. The visualization highlights the correlations among topics, and can be used to view the model at differing levels of complexity. The function makes use of the D3.js visualization library. The visualization needs to be viewed in a compatible web browser.

Usage

```
stmCorrViz(mod, file_out, documents_raw=NULL, documents_matrix=NULL,
           title="STM Model", clustering_threshold=FALSE,
           search_options = list(range_min=.05, range_max=5, step=.05),
           labels_number=7, display=TRUE, verbose=FALSE)
```

Arguments

mod	A fitted STM object from **stm**.
file_out	Name of the output file that will be generated by the function. This should end with an HTML extension.
documents_raw	The raw documents used to generate the STM model. A character vector where each entry is the full text of a document.
documents_matrix	Document-term matrix representation of the raw documents, as generated by the prepDocuments function.
title	Root node label. This defaults to "STM Model".
clustering_threshold	A parameter specifying the level of aggregation in the hierarchical clustering routine for topics. Lower threshold values produce more binary splits and deeper trees, while higher threshold values produce more aggregation and trees that have significant breadth rather than depth. See below for more details. If FALSE, a grid search is performed to find valid thresholds is performed using findThreshold. The valid clustering threshold resulting in a median level of tree complexity is chosen.
search_options	List specifying the grid search parameters to be used by findThreshold. Only necessary if clustering_threshold is FALSE.
labels_number	The number of top words used to label each node (topic or topical cluster) in the visualization.
display	Boolean. If set to TRUE, the visualization is launched in the system's default web browser upon function execution.
verbose	Boolean. If set to TRUE, displays function progress in the console during execution.

Details

This function generates a full-model, interactive, general-purpose hierarchical representation of an STM model. First a hierarchy of topics is created using hierarchical clustering as implemented in `hclust`. Then the hierarchy is written out to a JSON object using `stmJSON`. Finally D3.js is used to create an interactive visualization.

The visualization is built as a HTML page, and as such requires a web browser for inspection. The function does not return and object, but writes HTML output to disk.

The visualization takes the form of an indented tree. The leaves of the tree correspond to topics. The leaf nodes are grouped in topic clusters. This allows the model to be visualized at differing levels of aggregation. The function uses the D3.js library for visualization purpose. The visualization is largely built on top of Mike Bostock's Collapsible Indented Tree block. A nested JSON structure representing the hierarchical model is produced using the `stmJSON` function.

References

Bostock M, Vadim O, Jeffrey H. D3: Data-Driven Documents. Visualization and Computer Graphics, IEEE Transactions on 17.12 (2011): 2301-2309.

Margaret E. Roberts, Brandon M. Stewart and Dustin Tingley (2014). stm: R Package for Structural Topic Models.

See Also

`stmJSON`

Examples

```
data(immigration_perceptions)

stmCorrViz(immigration_perceptions$model, "corrviz.html",
  documents_raw=immigration_perceptions$raw_documents,
  documents_matrix=immigration_perceptions$documents_matrix)
```

`stmJSON` *Generate JSON Representation of STM Model*

Description

This function generates a nested JSON structure representing a fitted Structual Topic Model (STM). Used internally by `stmCorrViz`. Most users will not need to call this directly.

Usage

```
stmJSON(mod, documents_raw=NULL, documents_matrix=NULL,
         title="STM Model", clustering_threshold=1.5,
         labels_number=7, verbose)
```

Arguments

`mod`	An `STM` fitted model from the **stm** package.
`documents_raw`	The raw documents used to generate the STM model. A character vector where each entry is the full text of a document.
`documents_matrix`	Document-term matrix representation of the raw documents, as generated by the `prepDocuments` function.
`title`	Root node label. This defaults to "STM Model".
`clustering_threshold`	A parameter specifying the level of aggregation in the hierarchical clustering routine for topics. Lower threshold values produce more binary splits and deeper trees, while higher threshold values produce more aggregation and trees that have significant breadth rather than depth. See below for more details.
`labels_number`	The number of top words used to label each node (topic or topical cluster) in the visualization.
`verbose`	Logical. If set to TRUE, displays function progress in the console during execution.

Details

A nested JSON structure representing the hierarchical model is produced as follows. The function first retrieves the theta matrix from the STM object; accordingly computes correlations among topics; and then uses the correlation metrics to compute distances. The function finally performs hierarchical clustering on the topics by calling the `hclust` function.

The function finds all binary splits in the middle of the clustering tree whose clustering height measure is below the threshold specified in the `clustering_threshold` argument. All these splits are marked as aggregation points. The routine retrieves the merge matrix from the output of `hclust`, and produces a new merge list by deleting all the splits performed at aggregation points along the tree. While the `hclust` merge matrix only contains binary splits, the new merge list can contain non-binary cluster splits.

The merge list is transformed into a structure of nested lists with a recursive function call. Each level of this nested structure corresponds to a node in the hierarchical representation of the STM model. The data structure is eventually transformed into a JSON object by using the **jsonlite** package.

New beta matrices and top words are computed for each of the topic clusters according to their membership, by marginalizing over content covariates. The clusters are labeled accordingly.

Value

A JSON string representing the full STM model.

References

Margaret E. Roberts, Brandon M. Stewart and Dustin Tingley (2014). stm: R Package for Structural Topic Models.

See Also

stmCorrViz

Index