

# Galileo: Cluster Analysis Algorithms

Justin Grimmer      Gary King

Version  
September 4, 2008

# affinity: Computes Similarity (affinity) between documents

## Description

As currently implemented, takes an object from undergrad and computes the affinity between the documents. Currently, only cosine is implemented.

## Usage

```
affinity(undergrad, tfidf = T, type = "cosine", stop = T, extra = F, extra.dat = NULL)
```

## Arguments

<code>undergrad</code>	An object from the undergrad function
<code>tfidf</code>	If <code>tfidf=T</code> , calculate tf-idf weights to aid in topic detection
<code>type</code>	Type of affinity to compute, currently only cosine is available
<code>stop</code>	If <code>stop=T</code> , stop words are removed
<code>extra</code>	If <code>extra=T</code> affinity checks for an additional set of words to remove from the analysis
<code>extra.dat</code>	Used only if <code>extra=T</code> , the additional set of words to remove from the undergrad object

## Details

Computes the cosine between documents. Its output is used by Galileo to find topics in the data.

## Value

If there are K- documents, returns a K-K matrix of affinities between documents

## Warning

....

## Note

further notes

Make other sections like Warning with

# 1 Warning

....

## Author(s)

Justin Grimmer and Gary King

## See Also

objects to See Also as `help`,

# galileo: Function for Topic Estimation

## Description

This function performs a variety of clustering algorithms for topic-discovery in text.

## Usage

```
galileo(affinity, model = 'hclust', ...)
```

## Arguments

**affinity** Output from affinity, matrix describing the affinity between observations

**model** The clustering method desired for analysis. Currently, the following methods are available: agglomerative hierarchical clustering **hclust**, divisive hierarchical clustering **divisive**, hybrid hierarchical clustering **hybridHclust**, k-means **kmeans**, k-medoids **kmedoids**, affinity Propagation **affprop**, trimmed k-means **trimkmeans**, fuzzy k-means **fuzzy**, k-means for large data sets **lkmeans**, hard competitive learning **hardcl**, neural gas clustering **neuralgas**, qt clustering **qtclust**, and spectral clustering (normalized) **spectral**. See the help-files for each method for a full description about the strengths and weaknesses of each method of clustering.

...

parameters passed to the various methods, see the help files for each method.

## Details

**galileo** is a function that provides several different clustering algorithms, all using the output from **affinity**.

## Value

Returns a vector describing the cluster assignments for each document.

## Author(s)

Justin Grimmer and Gary King implemented **galileo** in R

## References

put references to the literature/web site here

## See Also

The help files for each method

## Examples

```
##---- Should be DIRECTLY executable !! ----  
##-- ==> Define data, use random,  
##--    or do help(data=index) for the standard data sets.
```

# mutinf: Cluster Labels by Mutual Information

## Description

Computes a set of labels for each document based upon the additional information each word provides for predicting the members of each cluster.

## Usage

```
mutinf(cluster, undergrad, stop = T, extra = F, extra.dat = NULL)
```

## Arguments

<code>cluster</code>	Cluster is an object from galileo
<code>undergrad</code>	An undergrad object used to compute cluster
<code>stop</code>	Remove stop words? Should agree with the call to galileo
<code>extra</code>	Remove additional words? Should agree with call to galileo
<code>extra.dat</code>	Additional words to be removed—Should agree with call to galileo

## Details

Takes an undergrad object and a clustering solution as arguments and returns the mutual information for each and cluster. The mutual information identifies words that are common for topics within cluster, but rare outside of the cluster.

## Value

If there are  $K$  topics and  $W$  stems, returns a  $K$  by  $W$  matrix of each word's mutual information for each topic.

## Warning

....

## Note

further notes

Make other sections like Warning with

## 2 Warning

....

**Author(s)**

Justin Grimmer and Gary King

**References**

Stanford Book

# clustTable: Latex Table of Cluster Output

## Description

This function provides a formatted table to display the output from the clustering routine. `clustTable` returns the proportion of total documents in each topic, along with a user specified number of stems to label each topic.

## Usage

```
clustTable(cluster, mut.inf, num = 5)
```

## Arguments

<code>cluster</code>	An object from <code>galileo</code>
<code>mut.inf</code>	An object from the <code>mutinf</code> function
<code>num</code>	A user specified number of stems to label each category

## Value

A LaTeX ready table to display the results of text classification

## Warning

....

## Note

further notes

Make other sections like Warning with

## 3 Warning

....

## Author(s)

Justin Grimmer and Gary King

## Contributors

Justin Grimmer and Gary King implemented `clustTable`.

## References