Galileo: Cluster Analysis Algorithms

Justin Grimmer Gary King

Version
March 31, 2008

`affinity`: Computes Similarity (affinity) between documents

Description

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

Usage

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

Arguments

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

Details

Computes the cosine between documents. Its output it 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

References

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(S, method = c('Affinity','Spectral'), p=NULL, lambda=0.6, maxits
= 500, convits = 50, num=NULL)
```

Arguments

`S`	Output from affinity, matrix describing the affinity between documents
`method`	Either 'Affinity' for Dueck and Frey (2007) Affinity propagation, or 'Spectral' for Ng, Jordan, and Weiss(2001) Spectral Clustering
`p`	
`lambda`	Dampening term for Affinity propagation
`maxits`	Maximum number of iterations, Affinity propagation
`convits`	Number of iterations to establish convergence, Affinity Propagation
`num`	Number of clusters, Spectral Clustering

Details

If method = 'Affinity' then approximate mixture of von Mises-Fisher clustering is performed using Affinity propogation (Dueck and Frey (2007)). The number of components is estimated from the data conditional on the prior (1 + log(prior, no. components)).

If method = 'Spectral' the spectral clustering method outline in Ng, Jordan, Weiss (2001) is performed.

Value

Returns a vector describing the cluster assignments for each document.

Note

Author(s)

Justin Grimmer and Gary King implemented both Algorithms in R.

References

put references to the literature/web site here

See Also

objects to See Also as `help`,

`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

`cluster`	Cluster is an object from galileo
`undergrad`	An undergrad object used to compute cluster
`stop`	Remove stop words? Should agree with the call to galileo
`extra`	Remove additional words? Should agree with call to galileo
`extra.dat`	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 ouput 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

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

Details

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

References

Contributors

Justin Grimmer and Gary King implemented `clustTable`.

References