Package ‘clustringr’

October 12, 2022

Type Package

Title Cluster Strings by Edit-Distance

Version 1.0

Author Dan S. Reznik

Maintainer Dan S. Reznik <dreznik@gmail.com>

Description Returns an edit-distance based clusterization of an input vector of strings. Each cluster will contain a set of strings w/ small mutual edit-distance (e.g., Levenshtein, optimum-sequence-alignment, Damerau-Levenshtein), as computed by stringdist::stringdist(). The set of all mutual edit-distances is then used by graph algorithms (from package 'igraph') to single out subsets of high connectivity.

License MIT + file LICENSE

Encoding UTF-8

LazyData true

Imports magrittr, dplyr, stringi, stringr, stringdist, igraph, assertthat, forcats, rlang, tidygraph, ggraph, ggplot2

Depends R (>= 3.1)

RoxygenNote 6.1.1

NeedsCompilation no

Repository CRAN

Date/Publication 2019-03-30 16:10:03 UTC

R topics documented:

cluster_plot *Plot string clusters as graph.*

Description

Plot string clusters as graph.

Usage

```
cluster_plot(cluster, min_cluster_size = 2, label_size = 2.5,
  repel = T)
```

Arguments

`cluster`	string clusters returned from ‘cluster_strings()‘
`min_cluster_size`	minimum size for clusters to be plotted.
`label_size`	how big should the cluster name fonts be.
`repel`	whether to "repel" (so cluster names won’t overlap)

Value

a graph plot (using ‘ggraph‘) of the string clusters.

Examples

```
s_vec <- c("alcool","alcohol","alcoholic","brandy","brandie","cachaça")
s_clust <- cluster_strings(s_vec,method="lv",max_dist=3,algo="cc")
cluster_plot(s_clust,min_cluster_size=1)
```

cluster_strings *Cluster Strings by Edit-Distance*

Description

Cluster Strings by Edit-Distance

Usage

```
cluster_strings(s_vec, clean = T, method = "osa", max_dist = 3,
  algo = "cc")
```

Arguments

`s_vec`	a vector of character strings
`clean`	whether to space-squish and de-duplicate s_vec
`method`	one of "osa","lv","dl" (as in ‘stringdist‘)
`max_dist`	max distance (typically damerau-levenshtein) between related strings.
`algo`	one of "cc" (connected components) or "eb" (edge betweeness)

Value

a data frame containing cluster membership for each input string

Examples

```
s_vec <- c("alcool","alcohol","alcoholic","brandy","brandie","cachaça")
s_clust <- cluster_strings(s_vec,method="lv",max_dist=3,algo="cc")
s_clust$df_clusters
```

`quijote_words` *Distinct words in Cervantes' "Don Quijote".*

Description

Dataframe listing all distinct words (length>3), their length, and frequency of appearance in text.

Usage

```
quijote_words
```

Format

A data frame w/ ~22k rows and 3 cols:

word the unique word, in Spanish

len the word's length

freq number of appearances in text

Source

http://www.gutenberg.org/cache/epub/2000/pg2000.txt

Index