Package 'WordListsAnalytics'

August 30, 2024

Type Package

Title Multiple Data Analysis Tools for Property Listing Tasks

Version 0.2.4

Maintainer Sebastian Moreno <sebastian.moreno.araya@gmail.com>

Description Application to estimate statistical values using properties provided by a group of individuals to describe
concepts using 'shiny'. It estimates the underlying distribution to generate new descriptive words Canessa et al. (2023) <doi:10.3758/s13428-022-01811-w>, applies a new clustering model, and uses simulations to estimate
the probability that two persons describe the same words based on their descriptions Canessa et al. (2022) <doi:10.3758/s13428-022-02030-z>.

License GPL (>= 3)

Encoding UTF-8

LazyData true

RoxygenNote 7.3.1

Imports ggplot2, readr, dplyr, reshape2, grDevices, stats, graphics, lsa

Depends shiny, R (>= 4.2.0)

Collate 'CPN_27.R' 'CPN_120.R' 'tab_upload_data.R' 'tab_estimations.R' 'tab_estimate_participants.R' 'tab_property_simulator.R' 'tab_pa_data.R' 'tab_pa_values.R' 'tab_cluster_image.R' 'tab_cosine_distance.R' 'tab_t_c_a_calculator.R' 'PLT_ui.R' 'fun_generate_norms.R' 'fun_estimate_participant.R' 'fun_property_simulator.R' 'fun_pa_function.R' 'fun_threshold_graph.R' 'fun_cluster_image_function.R' 'fun_tca_calculator.R' 'PLT_server.R' 'PLT_app.R'

NeedsCompilation no

Author Sebastian Moreno [aut, cre],
Cristobal Heredia [aut],
Enrique Canessa [ths],
Sergio Chaigneau [ths]

Repository CRAN

Date/Publication 2024-08-30 11:40:09 UTC

Contents

`clusterImage` *This function receives a property listing task, a given concept, and a threshold. It clusterizes the data according to the order of the listed properties. Given the mentioned properties of all users for a specific concept, the algorithm estimates a similarity among properties, based on the number of words mentioned between properties. For example, if the properties A and B are usually mentioned one after another, their similarity will be higher than the properties A and C which are usually not even mentioned together. The properties with low similarity to all other properties (below the user-defined threshold) are discarded from the plot.*

Description

This function receives a property listing task, a given concept, and a threshold. It clusterizes the data according to the order of the listed properties. Given the mentioned properties of all users for a specific concept, the algorithm estimates a similarity among properties, based on the number of words mentioned between properties. For example, if the properties A and B are usually mentioned one after another, their similarity will be higher than the properties A and C which are usually not even mentioned together. The properties with low similarity to all other properties (below the user-defined threshold) are discarded from the plot.

Usage

```
clusterImage(data, distThreshold, concept = NULL)
```

Arguments

`data`	Data frame with 3 columns: ID, Concept and Property
`distThreshold`	Distance value. It assign properties to specific cluster if their similarity is greater than distThreshold
`concept`	Text value. Clusters will only be generated with properties from this concept.

Value

List with 2 elements: ggplot2 plot and data frame with cluster information

Examples

```
data_cpn = data.frame(CPN_27)
threshold = 0.061
concept = "Ability"
cluster_data = clusterImage(data_cpn, threshold, concept)
```

CPN_120 *CPN Example data*

Description

The CPN120 dataset is a property listing task dataset over 120 concepts (60 concrete and 60 abstract). The dataset was generated from 221 voluntary Chilean university students (71% male, 28.5% female, average age = 23.7 years with s.d. = 6.2 years). Each participant listed up to 10 characteristics for each concept. The dataset had over 32,000 responses, which were categorized into valid and invalid, obtaining 31,864 valid responses.

Usage

```
data(CPN_120)
```

Format

A data frame with 31864 rows and 3 variables:

ID ID for original subject

Concept Concept asked

Property Property given by the subject ...

Source

Fondecyt proyect #1200139, Chilean goverment

CPN_27 *CPN Example data*

Description

The CPN27 dataset is a property listing task dataset over 27 abstract concepts. The dataset was generated from 100 voluntary Chilean university students (51% males, 49% females, mean age = 21.0 years with s.d. = 1.42 years). Each student listed features for 10 of the 27 concepts. The dataset had over 5,000 responses, which were sorted into valid and invalid, obtaining 4697 valid responses.

Usage

```
data(CPN_27)
```

Format

A data frame with 4618 rows and 3 variables:

ID ID for original subject

Concept Concept asked

Property Property given by the subject ...

Source

Fondecyt proyect #1200139, Chilean goverment

`estimate_participant` *Estimate the number of people needed and expected number of unique properties for a determined coverage based on the estimated norms*

Description

Estimate the number of people needed and expected number of unique properties for a determined coverage based on the estimated norms

Usage

```
estimate_participant(est_norms, target_cover)
```

Arguments

`est_norms`	A data frame with the estimated norms (generated by generateNorms function)
`target_cover`	Float between 0 and 1, corresponding to coverage (the fraction of the total incidence probabilities of the reported properties that are in the reference sample)

Value

A vector with the extra number of participant to achieve the specific coverage, and the estimate of the number of unique properties listed by the new amount of suggested people

Examples

```
data_cpn = data.frame(CPN_27)
estimated_norms = generate_norms(data_cpn)
estimated_norms = na.omit(estimated_norms)
estimate_participant(estimated_norms, 0.8)
```

generate_norms *Calculate all the norms from a Conceptual properties*

Description

Calculate all the norms from a Conceptual properties

Usage

```
generate_norms(orig_data)
```

Arguments

orig_data	Data frame with 3 columns: id, concept and properties

Value

Data frame with all the estimations of norms

Examples

```
data_test = data.frame(CPN_27)
generate_norms(data_test)
```

WordListsAnalytics *PLT App function*

Description

The WordListsAnalytics package provides a comprehensive Shiny app designed for analyzing and managing Property Listing Task (PLT) and/or Semantic Fluency Task (SFT) data. The app includes multiple tabs: Upload Data, Estimated Parameters, Sample Size Estimation, Data Simulator, Inputs to Calculate p(a), P(a) Calculation, and Clusters and Shifts.

Usage

```
WordListsAnalytics()
```

Details

To launch the Shiny app, call the WordListsAnalytics() function without using parameters.

Value

None (it executes a shiny application).

Tabs Details

Upload Data:: The "Upload Data" tab is the initial interface for users to upload their Property Listing Task (PLT) data. This data must consist of three columns: subject, concept, and property. Users also have the option to load example data (CPN-27, Canessa & Chaigneau, 2020) to familiarize themselves with the app's functionalities. In this tab, users can apply several data cleaning options:

- **Convert to Lower Case:** Change all data entries to lower case.
- **Delete Repeated Rows:** Remove duplicate rows to ensure unique data entries.
- **Delete Punctuation Marks:** Eliminate punctuation marks from the data.
- **Delete Spaces from Words:** Remove spaces within words for uniformity.

Users can preview the data to see the applied changes in real-time before proceeding with further analysis.

Estimated Parameters:: The "Estimated Parameters" tab allows researchers to view metrics for each listed concept. The metrics available in the table are:

- **Q1:** Number of properties reported by exactly one participant (singletons).
- **Q2:** Number of properties reported by exactly two participants (doubletons).
- **T:** Total number of participants who listed properties for a concept.
- **S_obs:** Observed semantic richness (unique properties listed for a concept).
- **U:** Total number of properties listed by all participants for a concept.
- **S_hat:** Estimated semantic richness (total unique properties if sampled infinitely).
- **sd_S_hat:** Standard deviation of the estimated semantic richness.
- **CI_L and CI_U:** Lower and upper bounds of the 95% confidence interval for the estimated semantic richness.
- **C_T:** Estimated coverage (proportion of total properties captured in the sample).

If there is insufficient data to calculate the metrics, the concept is added to the list of "Omitted NA's".

Sample Size Estimation:: This tab allows researchers to calculate coverage for a list of concepts in the data. Coverage is defined as the fraction of the total number of properties in the population captured in the sample for each concept (Canessa et al., 2023). By adjusting the expected coverage, researchers can determine if their data meets the required level of comprehensiveness. The tab displays a table with the following columns: Concept, T_star, S_hat_star, and Warning.

- **T_star:** Number of additional subjects needed to achieve the desired coverage.
- **S_hat_star:** Estimate of the semantic richness after including additional subjects.
- **Warning:** Indicates if Q2=0, meaning T_star cannot be calculated and further actions are required.

Data Simulator:: The property_simulator function generates synthetic data by modeling a probability distribution from which properties are sampled in a Property Listing Task (PLT). It is used to illustrate the incremental sampling procedure and does not need to accurately model any real probability distribution (Canessa et al., 2023). The function takes three parameters: "concept," "additional unique properties," and "number of subjects to generate." The "concept" parameter specifies the concept for which synthetic data will be generated. "Additional unique properties" is the number of new properties with a frequency of 1 to be added to the empirical distribution. "Number of subjects to generate" specifies the number of artificial subjects. The function returns a table with synthetic properties listed by each artificial subject.

Inputs to Calculate p(a):: This tab provides the necessary information to calculate the agreement probability. Researchers must select a concept. The tab displays a table listing each property mentioned for that specific concept and its frequency. An additional value, 's', is calculated for each concept, representing the average number of properties listed by subjects for a given concept in a PLT. The 's' value is repeated for each property row that belongs to the same concept to improve readability.

P(a) Calculation:: This tab calculates the agreement probability (P(a)) between pairs of concepts. Users can choose to calculate agreement probability for all concepts against themselves or for specific pairs. Users can adjust several parameters to improve the calculation:

- **Number of Repetitions:** Sets how many times the entire simulation process is repeated.
- **Number of Iterations:** Specifies the number of iterations within each repetition.
- **Moving Average Window Size:** Defines how many of the last iterations are averaged together to calculate the agreement probability (P(a)).

Clusters and Shifts:: This tab displays graphs for the "average number of clusters per subject," "average number of shifts per subject," and the "similarity matrix and clusters for a concept." Researchers must select a concept and set the "threshold for clustering" to generate these graphs. The threshold defines the minimum similarity required for two words to be included in the same cluster. The graphs obtained are:

- **Similarity Matrix and Clusters:** Shows how closely related pairs of words are based on their positions in lists generated by subjects.
- **Average Number of Clusters per Subject:** Indicates how many distinct groups of related words each subject creates on average.
- **Average Number of Shifts per Subject:** Reflects the fluidity of a subject's thought process and how often they switch contexts while listing words.

Users can adjust the resolution of the graphs and download them.

References

Canessa, E., & Chaigneau, S. E. (2020). Mathematical regularities of data from the property listing task. Journal of Mathematical Psychology, 97 doi:10.1016/j.jmp.2020.102376

Examples

```
if (interactive()) {
  WordListsAnalytics()
}
```

Index