

# Package ‘eHDPrep’

September 1, 2022

**Type** Package

**Title** Quality Control and Semantic Enrichment of Datasets

**Version** 1.2

**Maintainer** Ian Overton <I.Overton@qub.ac.uk>

**Description** A tool for the preparation and enrichment of health datasets for analysis. Provides functionality for assessing data quality and for improving the reliability and machine interpretability of a dataset.

eHDPrep also enables semantic enrichment of a dataset where metavariables are discovered from the relationships between input variables determined from user-provided ontologies.

**License** GPL-3

**Encoding** UTF-8

**LazyData** true

**RoxygenNote** 7.2.1

**Imports** ggplot2 (>= 3.3.3),  
dplyr (>= 1.0.3),  
forcats (>= 0.5.0),  
stringr (>= 1.4.0),  
purrr (>= 0.3.4),  
tidyr (>= 1.1.2),  
kableExtra (>= 1.3.1),  
magrittr (>= 2.0.1),  
tibble (>= 3.0.5),  
scales (>= 1.1.1),  
rlang (>= 0.4.10),  
quanteda (>= 2.1.2),  
tm (>= 0.7-8),  
pheatmap (>= 1.0.12),  
igraph (>= 1.2.6),  
tidygraph (>= 1.2.0),  
readr (>= 1.4.0),  
readxl (>= 1.3.1),  
knitr (>= 1.31)

**VignetteBuilder** knitr

**Depends** R (>= 3.6.0)

**Suggests** testthat (>= 2.1.0),  
ggraph (>= 2.0.5)

**Config/testthat/edition** 3

**R topics documented:**

apply_quality_ctrl . . . . .	3
assess_completeness . . . . .	5
assess_quality . . . . .	6
assume_var_classes . . . . .	8
cellspec_lgl . . . . .	9
compare_completeness . . . . .	10
compare_info_content . . . . .	10
compare_info_content_plt . . . . .	11
completeness_heatmap . . . . .	12
count_compare . . . . .	14
discrete.mi . . . . .	15
distant_neg_val . . . . .	16
encode_as_num_mat . . . . .	16
encode_binary_cats . . . . .	17
encode_bin_cat_vec . . . . .	18
encode_cats . . . . .	19
encode_genotypes . . . . .	20
encode_genotype_vec . . . . .	21
encode_ordinals . . . . .	21
entropy . . . . .	22
exact.kde . . . . .	23
example_data . . . . .	24
example_mapping_file . . . . .	25
example_ontology . . . . .	25
export_dataset . . . . .	26
extract_fretext . . . . .	26
identify_inconsistency . . . . .	27
import_dataset . . . . .	29
import_var_classes . . . . .	30
information_content_contin . . . . .	31
information_content_discrete . . . . .	31
join_vars_to_ontol . . . . .	32
max_catchNAs . . . . .	33
mean_catchNAs . . . . .	33
merge_cols . . . . .	34
metavariable_agg . . . . .	35
metavariable_info . . . . .	37
min_catchNAs . . . . .	38
mi_content_discrete . . . . .	39
mod_track . . . . .	39
node_IC_zhou . . . . .	40
normalize . . . . .	41
nums_to_NA . . . . .	42
onehot_vec . . . . .	42
ordinal_label_levels . . . . .	43
plot_completeness . . . . .	44
prod_catchNAs . . . . .	45
report_var_mods . . . . .	45
review_quality_ctrl . . . . .	46
row_completeness . . . . .	47

semantic_enrichment . . . . .	48
skipgram_append . . . . .	50
skipgram_freq . . . . .	51
skipgram_identify . . . . .	52
strings_to_NA . . . . .	53
sum_catchNAs . . . . .	54
validate_consistency_tbl . . . . .	55
validate_mapping_tbl . . . . .	56
validate_ontol_nw . . . . .	57
variable.bw.kde . . . . .	58
variable_completeness . . . . .	59
variable_entropy . . . . .	59
warn_missing_dots . . . . .	60
zero_entropy_variables . . . . .	60

<b>Index</b>	<b>62</b>
--------------	-----------

apply_quality_ctrl	<i>Apply quality control measures to a dataset</i>
--------------------	--

## Description

The primary high level function for quality control. Applies several quality control functions in sequence to input data frame (see Details for individual functions).

## Usage

```
apply_quality_ctrl(
  data,
  id_var,
  class_tbl,
  bin_cats = NULL,
  min_freq = 1,
  to_numeric_matrix = FALSE
)
```

## Arguments

data	A data frame, data frame extension (e.g. a tibble), or a lazy data frame (e.g. from dbplyr or dtplyr).
id_var	An unquoted expression which corresponds to a variable (column) in data which identifies each row (sample).
class_tbl	data frame such as the output tibble from <a href="#">assume_var_classes</a> followed by <a href="#">import_var_classes</a> .
bin_cats	Optional named vector of user-defined values for binary values using binary_label_1 = binary_label_2 syntax (e.g. c("No" = "Yes") would assign level 1 to "No" and 2 to "Yes"). See <a href="#">encode_binary_cats</a> for defaults. Applied to variables (columns) labelled "character" or "factor" in class_tbl.
min_freq	Minimum frequency of occurrence <a href="#">extract_freetext</a> will use to extract groups of proximal words in free-text from variables (columns) labelled "freetext" in class_tbl.

to\_numeric\_matrix

Should QC'ed data be converted to a numeric matrix? Default: FALSE.

## Details

The wrapped functions are applied in the following order:

1. Standardise missing values ([strings\\_to\\_NA](#))
2. Encode binary categorical variables (columns) ([encode\\_binary\\_cats](#))
3. Encode (specific) ordinal variables (columns) ([encode\\_ordinals](#))
4. Encode genotype variables ([encode\\_genotypes](#))
5. Extract information from free text variables (columns) ([extract\\_freetext](#))
6. Encode non-binary categorical variables (columns) ([encode\\_cats](#))
7. Encode output as numeric matrix (optional, [encode\\_as\\_num\\_mat](#))

class\_tbl is used to apply the above functions to the appropriate variables (columns).

## Value

data with several QC measures applied.

## See Also

Other high level functionality: [assess\\_quality\(\)](#), [review\\_quality\\_ctrl\(\)](#), [semantic\\_enrichment\(\)](#)

## Examples

```
data(example_data)
require(tibble)

# create an example class_tbl object
# note that diabetes_type is classes as ordinal and is not modified as its
# levels are not pre-coded
tibble::tribble(~"var", ~"datatype",
  "patient_id", "id",
  "tumoursize", "numeric",
  "t_stage", "ordinal_tstage",
  "n_stage", "ordinal_nstage",
  "diabetes", "factor",
  "diabetes_type", "ordinal",
  "hypertension", "factor",
  "rural_urban", "factor",
  "marital_status", "factor",
  "SNP_a", "genotype",
  "SNP_b", "genotype",
  "free_text", "freetext") -> data_types

data_QC <- apply_quality_ctrl(example_data, patient_id, data_types,
  bin_cats = c("No" = "Yes", "rural" = "urban"), min_freq = 0.6)
```

---

assess\_completeness     *Assess completeness of a dataset*

---

## Description

Assesses and visualises completeness of the input data across both rows (samples) and columns (variables).

## Usage

```
assess_completeness(data, id_var, plot = TRUE)
```

## Arguments

<code>data</code>	A data frame, data frame extension (e.g. a tibble), or a lazy data frame (e.g. from <code>dbplyr</code> or <code>dtplyr</code> ).
<code>id_var</code>	An unquoted expression which corresponds to a variable (column) in data which identifies each row (sample).
<code>plot</code>	Should plots be rendered when function is run? (Default: TRUE)

## Details

Returns a list of completeness assessments:

**variable\_completeness** A tibble detailing completeness of variables (columns) (via [variable\\_completeness](#)).

**row\_completeness** A tibble detailing completeness of rows (via [row\\_completeness](#)).

**completeness\_plot** A plot of row and variable (column) completeness (via [plot\\_completeness](#)).

**completeness\_heatmap** A clustered heatmap of cell completeness (via [completeness\\_heatmap](#)).

**plot\_completeness\_heatmap** A function which creates a clean canvas before plotting the completeness heatmap.

## Value

list of completeness tibbles and plots

## See Also

Other measures of completeness: [compare\\_completeness\(\)](#), [completeness\\_heatmap\(\)](#), [plot\\_completeness\(\)](#), [row\\_completeness\(\)](#), [variable\\_completeness\(\)](#)

## Examples

```
data(example_data)
res <- assess_completeness(example_data, patient_id)

# variable completeness table
res$variable_completeness

# row completeness table
res$row_completeness
```

```
# show completeness of rows and variables as a bar plot
res$completeness_plot

# show dataset completeness in a clustered heatmap
# (this is similar to res$completeness_heatmap but ensures a blank canvas is first created)
res$plot_completeness_heatmap(res)
```

---

assess\_quality

*Assess quality of a dataset*


---

## Description

Provides information on the quality of a dataset. Assesses dataset's completeness, internal consistency, and entropy.

## Usage

```
assess_quality(data, id_var, consis_tbl)
```

## Arguments

data	A data frame, data frame extension (e.g. a tibble), or a lazy data frame (e.g. from dbplyr or dtplyr).
id_var	An unquoted expression which corresponds to a variable (column) in data which identifies each row (sample).
consis_tbl	data frame or tibble containing information on internal consistency rules (see "Consistency Table Requirements" section)

## Details

Wraps several quality assessment functions from eHDPprep and returns a nested list with the following structure:

**completeness** - A list of completeness assessments:

1. Tibble of variable (column) completeness (via [variable\\_completeness](#))
2. Tibble of row (sample) completeness (via [row\\_completeness](#))
3. Plot of row and variable completeness (via [plot\\_completeness](#))
4. Completeness heatmap (via [completeness\\_heatmap](#))
5. A function which creates a clean canvas before plotting the completeness heatmap.

**internal\_inconsistency** - Tibble of internal inconsistencies, if any are present and if a consistency table is supplied (via [identify\\_inconsistency](#)).

**vars\_with\_zero\_entropy** - Names of variables (columns) with zero entropy (via [zero\\_entropy\\_variables](#))

## Value

Nested list of quality measurements

## Consistency Table Requirements

Table must have exactly five character columns. The columns should be ordered according to the list below which describes the values of each column:

1. First column name of data values that will be subject to consistency checking. String. Required.
2. Second column name of data values that will be subject to consistency checking. String. Required.
3. Logical test to compare columns one and two. One of: ">", ">=", "<", "<=", "==", "!=". String. Optional if columns 4 and 5 have non-NA values.
4. Either a single character string or a colon-separated range of numbers which should only appear in column A. Optional if column 3 has a non-NA value.
5. Either a single character string or a colon-separated range of numbers which should only appear in column B given the value/range specified in column 4. Optional if column 3 has a non-NA value.

Each row should detail one test to make. Therefore, either column 3 or columns 4 and 5 must contain non-NA values.

## See Also

Other high level functionality: [apply\\_quality\\_ctrl\(\)](#), [review\\_quality\\_ctrl\(\)](#), [semantic\\_enrichment\(\)](#)

## Examples

```
# general example
data(example_data)
res <- assess_quality(example_data, patient_id)

# example of internal consistency checks on more simple dataset
# describing bean counts
require(tibble)

# creating `data`:
beans <- tibble::tibble(red_beans = 1:15,
  blue_beans = 1:15,
  total_beans = 1:15*2,
  red_bean_summary = c(rep("few_beans",9), rep("many_beans",6)))

# creating `consis_tbl`
bean_rules <- tibble::tribble(~varA, ~varB, ~lgl_test, ~varA_boundaries, ~varB_boundaries,
  "red_beans", "blue_beans", "==", NA, NA,
  "red_beans", "total_beans", "<=", NA, NA,
  "red_beans", "red_bean_summary", NA, "1:9", "few_beans",
  "red_beans", "red_bean_summary", NA, "10:15", "many_beans")

# add some inconsistencies
beans[1, "red_bean_summary"] <- "many_beans"
beans[1, "red_beans"] <- 10

res <- assess_quality(beans, consis_tbl = bean_rules)

# variable completeness table
res$completeness$variable_completeness
```

```
# row completeness table
res$completeness$row_completeness

# show completeness of rows and variables as a bar plot
res$completeness$completeness_plot

# show dataset completeness in a clustered heatmap
res$completeness$plot_completeness_heatmap(res$completeness)

# show any internal inconsistencies
res$internal_inconsistency

# show any variables with zero entropy
res$vars_with_zero_entropy
```

---

assume_var_classes	<i>Assume variable classes in data</i>
--------------------	--

---

## Description

Classes/data types of data variables are assumed with this function and exported to a .csv file for amendment. Any incorrect classes can then be corrected and imported using [import\\_var\\_classes](#).

## Usage

```
assume_var_classes(data, out_file = "./datatypes.csv")
```

## Arguments

data	data frame
out_file	file where variables and their assumed classes are stored for user verification. Default: "./datatypes.csv"

## Value

Writes a .csv file in the working directory (default 'datatypes.csv') containing the variables and their assumed data types / classes.

## See Also

[import\\_var\\_classes](#)

## Examples

```
# example below assumes incorrectly for several variables
## Not run:
## Not run:
  data(example_data)
  assume_var_classes(example_data)

## End(Not run)
## End(Not run)
```



---

cellspec_lgl	<i>Kable logical data highlighting</i>
--------------	--

---

**Description**

Adds colour highlighting to cell values if they are encoded as logical values. Output should then be passed to knitr's kable function.

**Usage**

```
cellspec_lgl(.data, rg = FALSE)
```

**Arguments**

.data	Table to be highlighted.
rg	Should red and green be used for TRUE and FALSE values, respectively? If FALSE (default), colour-blind-friendly colours are applied.

**Details**

This is useful for identifying the encoding used in a value (e.g. the difference between the string "TRUE" and truth value of logic TRUE). This highlighting can also be useful when visually assessing cell values in a table. The colour naming format (HTML or LaTeX) is automatically detected. There are four cell types considered:

1. non-logical cells are coloured black)
2. TRUE cells are coloured red (default) or green if rg is TRUE
3. FALSE cells are coloured cyan (default) or red if rg is TRUE
4. NA cells are coloured gray

Note: When passed to kable(), the escape parameter should be FALSE for colours to be rendered correctly.

**Value**

Table with cell colours specified.

**See Also**

[kable\\_cell\\_spec](#)

**Examples**

```
require(knitr)
require(kableExtra)
require(magrittr)
# create dataframe with a logical variable and non-logical variable
data.frame(lgls = c(TRUE,TRUE,FALSE,NA), non_lgls = c("true","true","false","NA")) %>%
  eHDPrep::cellspec_lgl() %>%
  kable(escape = FALSE)
```

---

compare_completeness	<i>Compare Completeness between Datasets</i>
----------------------	--

---

**Description**

Produces a density plot comparing the completeness of two datasets (tbl\_a and tbl\_b) for variables (if dim == 2, default) or row (if dim == 1). The label used to identify the dataset's density curve can be specified using tbl\_a\_lab and tbl\_b\_lab.

**Usage**

```
compare_completeness(tbl_a, tbl_b, dim = 2, tbl_a_lab = NULL, tbl_b_lab = NULL)
```

**Arguments**

tbl_a	Data frame of the first data frame to compare.
tbl_b	Data frame of the second data frame to compare.
dim	Integer. Dimension to measure completeness on. 2 (Default) measures completeness by variable. 1 measures completeness by row.
tbl_a_lab	String to be used to label tbl_a on the output plot.
tbl_b_lab	String to be used to label tbl_b on the output plot.

**Value**

Plot showing densities of completeness across both datasets.

**See Also**

Other measures of completeness: [assess\\_completeness\(\)](#), [completeness\\_heatmap\(\)](#), [plot\\_completeness\(\)](#), [row\\_completeness\(\)](#), [variable\\_completeness\(\)](#)

**Examples**

```
data(example_data)
compare_completeness(example_data, strings_to_NA(example_data), dim = 2,
                      "raw", "cleaned")
```

---

compare_info_content	<i>Information Content Comparison Table</i>
----------------------	---

---

**Description**

Used to quantify the amount of information loss, if any, which has occurred in a merging procedure between two discrete variables.

**Usage**

```
compare_info_content(input1, input2, composite)
```

**Arguments**

input1	Character vector. First variable to compare
input2	Character vector. Second variable to compare
composite	Character vector. Composite variable, resultant of merging input1 and input2.

**Details**

The function requires the two discrete variables which have been merged (input1 and input2) and the composite variable (output). For each input, information content is calculated using [information\\_content\\_discrete](#) along with each input's mutual information content with the composite variable using [mi\\_content\\_discrete](#). The function returns a table describing these measures.

If the mutual information content between an input variable and the composite variable is equal to the information content of the input variable, it is confirmed that all information in the input variable has been incorporated into the composite variable. However, if one or both input variables' information content is not equal to their mutual information with the composite variables, information loss has occurred.

**Value**

Table containing information content for input1 and input2 and their mutual information content with composite.

**See Also**

[compare\\_info\\_content\\_plt](#)

**Examples**

```
data(example_data)
require(dplyr)
require(magrittr)
example_data %>%
  mutate(diabetes_merged = coalesce(diabetes_type, diabetes)) %>%
  select(starts_with("diabetes")) ->
  merged_data

compare_info_content(merged_data$diabetes,
                     merged_data$diabetes_type,
                     merged_data$diabetes_merged)
```

---

compare\_info\_content\_plt

*Information Content Comparison Plot*

---

**Description**

This function requires the output from [compare\\_info\\_content](#). It is used to visualise the amount of information loss, if any, which has occurred in a merging procedure between two discrete variables.

**Usage**

```
compare_info_content_plt(compare_info_content_res)
```

**Arguments**

```
compare_info_content_res
```

Output from [compare\\_info\\_content](#).

**Details**

If the mutual information content between an input variable and the composite variable is equal to the information content of the input variable, it is confirmed that all information in the input variable has been incorporated into the composite variable.

**Value**

Plot of measures calculated in [compare\\_info\\_content](#).

**See Also**

[compare\\_info\\_content](#)

**Examples**

```
data(example_data)
require(dplyr)
require(magrittr)
example_data %>%
  mutate(diabetes_merged = coalesce(diabetes_type, diabetes)) %>%
  select(starts_with("diabetes")) ->
  merged_data

compare_info_content(merged_data$diabetes,
                     merged_data$diabetes_type,
                     merged_data$diabetes_merged) %>%
  compare_info_content_plt()
```

---

completeness\_heatmap    *Completeness Heatmap*

---

**Description**

Produces a heatmap visualising completeness across a dataset.

**Usage**

```
completeness_heatmap(
  data,
  id_var,
  annotation_tbl = NULL,
  method = 1,
  show_rownames = FALSE,
  ...
)
```

**Arguments**

<code>data</code>	Data frame to be analysed.
<code>id_var</code>	Character constant of row identifier variable name.
<code>annotation_tbl</code>	Data frame containing variable annotation data. Column 1 should contain variable names, column 2 should contain an annotation label.
<code>method</code>	Integer between 1 and 3. Default: 1. See Details for more information.
<code>show_rownames</code>	Boolean. Should rownames be shown. Default: False.
<code>...</code>	Parameters to be passed to <a href="#">pheatmap</a> .

**Details**

- Method 1: Missing values are numerically encoded with a highly negative number, numerically distant from all values in data, using [distant\\_neg\\_val](#). Values in categorical variables are replaced with the number of unique values in the variable. Clustering uses these values. Cells are coloured by presence (yellow = missing; blue = present).
- Method 2: Same as Method 1 but cells are coloured by values used to cluster.
- Method 3: Values in data are encoded as Boolean values for clustering (present values = 1; missing values = 0). Cells are coloured by presence (yellow = missing; blue = present).

**Value**

completeness heatmap

**Note**

If the heatmap overlaps with other plots on the current device, it is recommended that users run the [grid.newpage](#) function to ensure a clean page is used for this plot.

**References**

Kolde R (2019). `_pheatmap: Pretty Heatmaps_`. R package version 1.0.12, <<https://CRAN.R-project.org/package=pheatmap>>.

**See Also**

[pheatmap](#)

Other measures of completeness: [assess\\_completeness\(\)](#), [compare\\_completeness\(\)](#), [plot\\_completeness\(\)](#), [row\\_completeness\(\)](#), [variable\\_completeness\(\)](#)

**Examples**

```
data(example_data)
completeness_heatmap(example_data,patient_id)

# with variable-level annotations
## create a dataframe containing variable annotations
tibble::tribble(~"var", ~"datatype",
  "patient_id", "id",
  "tumoursize", "numeric",
  "t_stage", "ordinal_tstage",
  "n_stage", "ordinal_nstage",
  "diabetes", "factor",
```

```

"diabetes_type", "ordinal",
"hypertension", "factor",
"rural_urban", "factor",
"marital_status", "factor",
"SNP_a", "genotype",
"SNP_b", "genotype",
"free_text", "freetext") -> data_types

completeness_heatmap(example_data, patient_id, annotation_tbl = data_types)

```

count\_compare

*Compare unique values before and after data modification***Description**

Performs comparison of variables before and after a change has been applied in order to allow manual inspection and review of modifications made during the dataset preparation process.

**Usage**

```

count_compare(
  cols2compare,
  before_tbl = NULL,
  after_tbl = NULL,
  only_diff = FALSE,
  kableout = TRUE,
  caption = NULL,
  latex_wrap = FALSE
)

```

**Arguments**

cols2compare	Variables to compare between tables.
before_tbl	Data frame from before modification was made.
after_tbl	Data frame from after modification was made.
only_diff	Keep only rows which differ between the tables (good for variables with many unique values, such as numeric variables).
kableout	Should output be a kable from knitr? If not, returns a tibble. (Default: TRUE)
caption	Caption for kable's caption parameter.
latex_wrap	Should tables be aligned vertically rather than horizontally? Useful for wide table which would otherwise run off a page in LaTeX format.

**Details**

The purpose of this function is to summarise individual alterations in a dataset and works best with categorical variables. The output contains two tables derived from the parameters `before_tbl` and `after_tbl`. Each table shows the unique combinations of values in variables specified in the parameter `cols2compare` if the variable is present. The tables are presented as two sub-tables and therefore share a single table caption. This caption is automatically generated describing the content

of the two sub-tables when the parameter `caption` is not specified. The default output is a kable containing two sub-kables however if the parameter `kableout` is `FALSE`, a list containing the two tibbles are returned. This may preferable for further analysis on the tables' contents.

### Value

Returns list of two tibbles or a kable (see `kableout` argument), each tallying unique values in specified columns in each input table.

### Examples

```
# merge data as the example modification
example_data_merged <- merge_cols(example_data, diabetes_type, diabetes,
  "diabetes_merged", rm_in_vars = TRUE)

# review the differences between the input and output of the variable merging step above:
count_compare(before_tbl = example_data,
  after_tbl = example_data_merged,
  cols2compare = c("diabetes", "diabetes_type", "diabetes_merged"),
  kableout = FALSE)
```

---

discrete.mi

---

*Calculate mutual information of a matrix of discrete values*


---

### Description

Compute mutual information between all rows of a matrix containing discrete outcomes.

### Usage

```
discrete.mi(mat, progress.bar = FALSE)
```

### Arguments

<code>mat</code>	A matrix of discrete values
<code>progress.bar</code>	Outputs status to terminal when set to 'text', or no status updates are output when set to <code>FALSE</code> .

### Details

Note that only the lower triangle of the matrix is populated for speed, as the result is symmetric. Takes a matrix as input.

### Value

A lower triangular matrix where element `[i,j]` contains the mutual information in bits between row `i` and row `j` of the input matrix

### Author(s)

Alexander Lyulph Robert Lubbock, Ian Overton

---

distant_neg_val	<i>Find highly distant value for data frame</i>
-----------------	---

---

### Description

Returns a numeric value which is distant from the values in data using the following equation:

$$output = -2 * (max(data) - min(data))$$

### Usage

```
distant_neg_val(data)
```

### Arguments

data	data frame.
------	-------------

### Value

Numeric vector of length 1

### Examples

```
data(mtcars)
eHDPrep::distant_neg_val(mtcars)
```

---

encode_as_num_mat	<i>Convert data frame to numeric matrix</i>
-------------------	---

---

### Description

Converts all columns to numeric and uses the row identifier column (id\_var) as row names.

### Usage

```
encode_as_num_mat(data, id_var)
```

### Arguments

data	A data frame, data frame extension (e.g. a tibble), or a lazy data frame (e.g. from dbplyr or dtplyr).
id_var	An unquoted expression which corresponds to a variable in data which identifies each row.

### Value

Numeric matrix with id\_var values as row names



**Examples**

```
require(dplyr)
require(magrittr)
mtcars %>%
  dplyr::as_tibble(rownames = "id") %>%
  encode_as_num_mat(id)
```

---

encode_binary_cats	<i>Encode categorical variables as binary factors</i>
--------------------	---

---

**Description**

In a data frame, converts binary categories to factors. Ordering of levels is standardised to: negative\_finding, positive\_finding. This embeds a standardised numeric relationship between the binary categories while preserving value labels.

**Usage**

```
encode_binary_cats(data, ..., values = NULL)
```

**Arguments**

data	A data frame, data frame extension (e.g. a tibble), or a lazy data frame (e.g. from dbplyr or dtplyr).
...	<tidy-select> One or more unquoted expressions separated by commas. Variable names can be used as if they were positions in the data frame, so expressions like x:y can be used to select a range of variables.
values	Optional named vector of user-defined values for binary values using binary_label_1 = binary_label_2 syntax (e.g. c("No" = "Yes") would assign level 1 to "No" and 2 to "Yes").

**Details**

Binary categories to convert can be specified with a named character vector, specified in values. The syntax of the named vector is: negative\_finding = positive\_finding. If values is not provided, the default list will be used: "No"="Yes", "No/unknown" = "Yes", "no/unknown" = "Yes", "Non-user" = "User", "Never" = "Ever", "WT" = "MT".

**Value**

dataset with specified binary categories converted to factors.

**Examples**

```
# use built-in values. Note: rural_urban is not modified
# Note: diabetes is not modified because "missing" is interpreted as a third category.
# strings_to_NA() should be applied first
encode_binary_cats(example_data, hypertension, rural_urban)

# use custom values. Note: rural_urban is now modified as well.
encoded_data <- encode_binary_cats(example_data, hypertension, rural_urban,
  values = c("No" = "Yes", "rural" = "urban"))
```



---

encode_cats	<i>Encode categorical variables using one-hot encoding.</i>
-------------	---

---

## Description

Variables specified in `...` are replaced with new variables describing the presence of each unique category. Generated variable names have space characters replaced with `"_"` and commas are removed.

## Usage

```
encode_cats(data, ...)
```

## Arguments

<code>data</code>	A data frame, data frame extension (e.g. a tibble), or a lazy data frame (e.g. from <code>dbplyr</code> or <code>dtplyr</code> ).
<code>...</code>	<code>&lt;tidy-select&gt;</code> One or more unquoted expressions separated by commas. Variable names can be used as if they were positions in the data frame, so expressions like <code>x:y</code> can be used to select a range of variables.

## Value

Tibble with converted variables.

## Examples

```
require(magrittr)
require(dplyr)

data(example_data)

# encode one variable
encode_cats(example_data, marital_status) %>%
  select(starts_with("marital_status"))

# encode multiple variables
encoded <- encode_cats(example_data, diabetes, marital_status)

select(encoded, starts_with("marital_status"))
# diabetes_type included below but was not modified:
select(encoded, starts_with("diabetes"))
```

---

encode_genotypes	<i>Encode genotype/SNP variables in data frame</i>
------------------	--

---

## Description

Standardises homozygous SNPs (e.g. recorded as "A") to two character form (e.g. "A/A") and orders heterozygous SNPs alphabetically (e.g. "GA" becomes "A/G"). The SNP values are then converted from a character vector to an ordered factor, ordered by observed allele frequency (in the supplied cohort). The most frequent allele is assigned level 1, the second most frequent value is assigned level 2, and the least frequent values is assigned level 3). This method embeds the numeric relationship between the allele frequencies while preserving value labels.

## Usage

```
encode_genotypes(data, ...)
```

## Arguments

data	A data frame, data frame extension (e.g. a tibble), or a lazy data frame (e.g. from dbplyr or dtplyr).
...	<tidy-select> One or more unquoted expressions separated by commas. Variable names can be used as if they were positions in the data frame, so expressions like x:y can be used to select a range of variables.

## Value

‘data’ with variables (...) encoded as standardised genotypes

## Examples

```
data(example_data)
require(dplyr)
require(magrittr)

# one variable
encode_genotypes(example_data, SNP_a) %>%
  select(SNP_a)

# multiple variables
encode_genotypes(example_data, SNP_a, SNP_b) %>%
  select(SNP_a, SNP_b)

# using tidyselect helpers
encode_genotypes(example_data, dplyr::starts_with("SNP")) %>%
  select(starts_with("SNP"))
```

---

encode_genotype_vec	<i>Encode a genotype/SNP vector</i>
---------------------	-------------------------------------

---

### Description

Standardises homozygous SNP alleles (e.g. recorded as 'A') to two character form (e.g. 'A/A') and orders heterozygous SNP alleles alphabetically (e.g. "GA" becomes "A/G"). The SNP values are then converted from a character vector to an ordered factor, ordered by SNP allele frequency (e.g. most frequent SNP allele is 1, second most frequent value is 2, and least frequent values is 3). This method embeds the numeric relationship between the SNP allele frequencies while preserving value labels.

### Usage

```
encode_genotype_vec(x)
```

### Arguments

x	input vector containing genotype data
---	---------------------------------------

### Value

Ordered factor, ordered by allele frequency in variable

### Examples

```
data(example_data)
vec <- example_data[, "SNP_a", drop = TRUE]
eHDPRep:::encode_genotype_vec(vec)
```

---

encode_ordinals	<i>Encode ordinal variables</i>
-----------------	---------------------------------

---

### Description

Converts character or factor variables in the input data frame to ordered factors embedding numeric relationship between values while preserving value labels.

### Usage

```
encode_ordinals(data, ord_levels, ..., strict_levels = TRUE)
```

### Arguments

data	A data frame, data frame extension (e.g. a tibble), or a lazy data frame (e.g. from dbplyr or dtplyr).
ord_levels	character vector containing values in desired order (lowest to highest).
...	<tidy-select> One or more unquoted expressions separated by commas. Variable names can be used as if they were positions in the data frame, so expressions like x:y can be used to select a range of variables.

`strict_levels` logical constant. If TRUE, variables in ... which contain values other than `ord_levels` (including NA) are not modified and a warning is given. If FALSE, values not in `ord_levels` are converted to NA.

**Value**

dataframe with specified variables encoded as ordered factors.

**Examples**

```
data(example_data)
require(dplyr)
require(magrittr)
encode_ordinals(example_data, ord_levels = c("N0", "N1", "N2"), n_stage)

# Note: "unequivocal" is present in t_stage but not in `ord_levels`.
# with `strict_levels` TRUE, t_stage is unmodified and a warning message is given:

encode_ordinals(example_data,
  ord_levels = c("T1", "T2", "T3a", "T3b", "T4"), strict_levels = TRUE, t_stage) %>%
  select(t_stage)

# with `strict_levels` FALSE, it is replaced with NA:

encode_ordinals(example_data,
  ord_levels = c("T1", "T2", "T3a", "T3b", "T4"), strict_levels = FALSE, t_stage) %>%
  select(t_stage)
```

---

entropy	<i>Calculate Entropy of a Vector</i>
---------	--------------------------------------

---

**Description**

Calculates Shannon Entropy of a vector in bits (default) or natural units. Missing values are omitted from the calculation.

**Usage**

```
entropy(x, unit = c("bits"))
```

**Arguments**

- `x` Input vector
- `unit` Unit to measure entropy. Either "bits" (default) or "nats".

**Value**

Entropy of input variable

**References**

Shannon, C. E. A mathematical theory of communication. The Bell System Technical Journal 27, 379–423 (1948).

**Examples**

```
# no entropy:
vec <- c(1,1,1,1,1,1)
entropy(vec)

# entropy
vec <- c(1,2,3,4,5,6)
entropy(vec)
```

---

exact.kde	<i>Exact kernel density estimation</i>
-----------	--

---

**Description**

Calculates KDE for a set of points exactly, rather than an approximation as per the `density()` core function.

**Usage**

```
exact.kde(x, bw, output.domain = x, na.rm = FALSE)
```

**Arguments**

<code>x</code>	A numeric vector of values
<code>bw</code>	The bandwidth to use - either a single value, or a vector of values the same length as <code>x</code> if using adaptive bandwidth estimation (with each value giving the bandwidth at the corresponding data point).
<code>output.domain</code>	The domain of values over which to estimate the density. Defaults to <code>x</code> . To use the same domain of <code>x</code> values as <code>R</code> 's <code>density</code> , set to <code>NULL</code> .
<code>na.rm</code>	Remove missing values if <code>TRUE</code>

**Details**

Only tractable for around 10,000 data points or less - otherwise consider using the `density()` core function for a close approximation.

The `density()` core function approximation is normally a very good approximation, but some small values close to zero may become zero rather than just very small. This makes it less suitable for mutual information estimation.

**Value**

The exact kernel density estimate as a `density` object, compatible with `R`'s `density` function.

**Author(s)**

Alexander Lyulph Robert Lubbock, Ian Overton

## Examples

```
## Not run:
x <- c(rnorm(100,sd=2),runif(50),4*rnorm(75,mean=2,sd=4))
plot(exact.kde(x,bw.nrd0(x),output.domain=NULL),xlab="x",
      col="black",lty=2,main="Exact vs approximate KDE")
lines(density(x),col="red",lty=1)
legend("topright",c("Exact KDE","Approx KDE"),lty=2:1,col=c("black","red"))

## End(Not run)
```

---

example\_data

*Example data for eHDPprep*

---

## Description

A dataset containing synthetic example values to demonstrate functionality of eHDPprep

## Usage

example\_data

## Format

A data frame with 1,000 rows and 10 variables:

**patient\_id** 1 to 1000, effectively row numbers

**tumoursize** double. random values with a mean of 50 and SD of 20

**t\_stage** character. T stage random values

**n\_stage** character. N stage random values

**diabetes** character. Patient diabetes category

**diabetes\_type** character. Patient diabetes type category

**hypertension** character. Patient hypertension category

**rural\_urban** character. Patient domestic address category

**marital\_status** character. Patient marital status category

**SNP\_a** character. Single Nucleotide Polymorphism (SNP) of the patient

**SNP\_b** character. Another SNP of the patient

**free\_text** character. sentences from the stringr package as an example of short free text variables

## Source

synthetic



---

example_mapping_file	<i>Example mapping file for semantic enrichment</i>
----------------------	---

---

**Description**

A data frame containing mappings between variables in 'example\_data' and 'example\_onto'. Used to demonstrate semantic enrichment.

**Usage**

```
example_mapping_file
```

**Format**

A data frame:

**variable** character. names of variables in post-QC 'example\_data'.

**onto\_entity** character. names of mapped entities in 'example\_onto'.

**Details**

Maps variables in eHDPRep's 'example\_data' to eHDPRep's 'example\_ontology'.

**Source**

synthetic

---

example_ontology	<i>Example ontology as a network graph for semantic enrichment</i>
------------------	--

---

**Description**

A small custom network graph to demonstrate semantic enrichment.

**Usage**

```
example_ontology
```

**Format**

```
tidygraph graph
```

**Details**

Contains semantic links of variables in eHDPRep's 'example\_data' following quality control.

**Source**

synthetic

---

export_dataset	<i>Export data to delimited file</i>
----------------	--------------------------------------

---

### Description

Save dataset in .csv or .tsv format. A wrapper function for readr's [write\\_csv](#) and [write\\_tsv](#).

### Usage

```
export_dataset(x, file, format = "csv", ...)
```

### Arguments

x	A data frame or tibble to write to disk.
file	File or connection to write to.
format	Character constant. "csv" (default) or "tsv"
...	parameters to pass to <a href="#">write_csv</a> or <a href="#">write_tsv</a> .

### See Also

[write\\_csv](#) and [write\\_tsv](#)

Other import to/export from R functions: [import\\_dataset\(\)](#)

---

extract_freetext	<i>Extract information from free text</i>
------------------	---

---

### Description

Extracts information from specified free text variables (...) which occur in a minimum amount of rows (min\_freq) and appends new variables to data.

### Usage

```
extract_freetext(data, id_var, min_freq = 1, ...)
```

### Arguments

data	Data frame to append skipgram variables to.
id_var	An unquoted expression which corresponds to a variable in data which identifies each row.
min_freq	Minimum percentage frequency of skipgram occurrence to return. Default = 1.
...	Unquoted expressions of free text variable names from which to extract information.

### Details

New variables report the presence of skipgrams (proximal words in the text) with a minimum frequency (min\_freq, default = 1%).

**Value**

data with additional Boolean variables describing skipgrams in . . .

**References**

Guthrie, D., Allison, B., Liu, W., Guthrie, L. & Wilks, Y. A Closer Look at Skip-gram Modelling. in Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06) (European Language Resources Association (ELRA), 2006).

Benoit K, Watanabe K, Wang H, Nulty P, Obeng A, Müller S, Matsuo A (2018). “quanteda: An R package for the quantitative analysis of textual data.” *Journal of Open Source Software*, \*3\*(30), 774. doi:10.21105/joss.00774 <<https://doi.org/10.21105/joss.00774>>, <<https://quanteda.io>>.

Feinerer I, Hornik K (2020). *\_tm: Text Mining Package\_*. R package version 0.7-8, <<https://CRAN.R-project.org/package=tm>>.

Ingo Feinerer, Kurt Hornik, and David Meyer (2008). Text Mining Infrastructure in R. *Journal of Statistical Software* 25(5): 1-54. URL: <https://www.jstatsoft.org/v25/i05/>.

**See Also**

Principle underlying function: [tokens\\_ngrams](#)

Other free text functions: [skipgram\\_append\(\)](#), [skipgram\\_freq\(\)](#), [skipgram\\_identify\(\)](#)

**Examples**

```
data(example_data)
extract_freertext(example_data, patient_id, min_freq = 0.6, free_text)
```

---

identify\_inconsistency

*Identify inconsistencies in a dataset*

---

**Description**

Tests pairs of variables for consistency between their values according to a table of rules or 'consistency table'.

**Usage**

```
identify_inconsistency(data = NULL, consis_tbl = NULL, id_var = NULL)
```

**Arguments**

data	data frame which will be checked for internal consistency
consis_tbl	data frame or tibble containing information on internal consistency rules (see "Consistency Table Requirements" section)
id_var	An unquoted expression which corresponds to a variable in data which identifies each row.

## Details

Multiple types of checks for inconsistency are supported:

1. Comparing by logical operators (<, <=, ==, !=, >=, >)
2. Comparing permitted categories (e.g. cat1 in varA only if cat2 in varB)
3. Comparing permitted numeric ranges (e.g. 20-25 in varC only if 10-20 in varD)
4. Mixtures of 2 and 3 (e.g. cat1 in varA only if 20-25 in varC)

The consistency tests rely on such rules being specified in a separate data frame (consis\_tbl; see section "Consistency Table Requirements").

Variable A is given higher priority than Variable B when A is a category. If A (as char) is not equal to the value in col 4, the check is not made. This is to account for one way dependencies (i.e. VarA is fruit, VarB is apple)

## Value

tibble detailing any identified internal inconsistencies in data, if any are found. If no inconsistencies are found, data is returned invisibly.

## Consistency Table Requirements

Table must have exactly five character columns. The columns should be ordered according to the list below which describes the values of each column:

1. First column name of data values that will be subject to consistency checking. String. Required.
2. Second column name of data values that will be subject to consistency checking. String. Required.
3. Logical test to compare columns one and two. One of: ">", ">=", "<", "<=", "==", "!=". String. Optional if columns 4 and 5 have non-NA values.
4. Either a single character string or a colon-separated range of numbers which should only appear in column A. Optional if column 3 has a non-NA value.
5. Either a single character string or a colon-separated range of numbers which should only appear in column B given the value/range specified in column 4. Optional if column 3 has a non-NA value.

Each row should detail one test to make. Therefore, either column 3 or columns 4 and 5 must contain non-NA values.

## See Also

Other internal consistency functions: [validate\\_consistency\\_tbl\(\)](#)

## Examples

```
require(tibble)
# example with synthetic dataset on number of bean counts
# there is a lot going on in the function so a simple dataset aids this example
#
# creating `data`:
beans <- tibble::tibble(red_beans = 1:15,
blue_beans = 1:15,
```

```

total_beans = 1:15*2,
red_bean_summary = c(rep("few_beans",9), rep("many_beans",6)))
#
# creating `consis_tbl`
bean_rules <- tibble::tribble(~varA, ~varB, ~lgl_test, ~varA_boundaries, ~varB_boundaries,
"red_beans", "blue_beans", "==", NA, NA,
"red_beans", "total_beans", "<=", NA,NA,
"red_beans", "red_bean_summary", NA, "1:9", "few_beans",
"red_beans", "red_bean_summary", NA, "10:15", "many_beans")

identify_inconsistency(beans, bean_rules)

# creating some inconsistencies as examples
beans[1, "red_bean_summary"] <- "many_beans"
beans[1, "red_beans"] <- 10

identify_inconsistency(beans, bean_rules)

```

import\_dataset

*Import data into R*

## Description

Imports a rectangular single table into R from a .xls, .xlsx, .csv, or .tsv file.

## Usage

```
import_dataset(file, format = "excel", ...)
```

## Arguments

file	Character constant. Path to file.
format	Character constant. "excel" (default, for .xls or .xlsx files), "csv", or "tsv".
...	Parameters to pass to <a href="#">read_excel</a> , <a href="#">read_csv</a> or <a href="#">read_tsv</a>

## Details

First row is interpreted as column headers by default. For more details see [read\\_excel](#) (.xlsx/.xls), [read\\_csv](#) (.csv), or [read\\_tsv](#) (.tsv).

## Value

data as a tibble

## See Also

[read\\_excel](#) for additional parameters for importing .xls or .xlsx files, [read\\_csv](#) for .csv files, [read\\_tsv](#) for .tsv files

Other import to/export from R functions: [export\\_dataset\(\)](#)

## Examples

```
## Not run:
# This code will not run
# ./dataset.xlsx should be replaced with path to user's dataset

# excel
import_dataset(file = "./dataset.xlsx", format = "excel")
#csv
import_dataset(file = "./dataset.csv", format = "csv")
#tsv
import_dataset(file = "./dataset.tsv", format = "tsv")

## End(Not run)
```

---

import_var_classes	<i>Import corrected variable classes</i>
--------------------	--

---

## Description

Reads in output of [assume\\_var\\_classes](#), ensures all specified datatypes are one of ("id", "numeric", "double", "integer", "character", "factor", "ordinal", "genotype", "freetext", "logical") as required for high level eHDPprep functions.

## Usage

```
import_var_classes(file = "./datatypes.csv")
```

## Arguments

file	character string. Path to output of <a href="#">assume_var_classes</a> which should be manually verified outside of R and corrected where any data type is incorrect.
------	---

## Value

data frame containing the data type values of variables, as described in file

## See Also

[assume\\_var\\_classes](#)

## Examples

```
## Not run:
## Not run:
data(example_data)
assume_var_classes(example_data)
import_var_classes()

## End(Not run)
## End(Not run)
```

---

`information_content_contin`*Calculate Information Content (Continuous Variable)*

---

**Description**

Calculates information content of a continuous (numeric) vector in bits (default) or natural units. Missing values are omitted from the calculation.

**Usage**

```
information_content_contin(x, unit = c("bits"))
```

**Arguments**

<code>x</code>	Input vector
<code>unit</code>	Unit to measure entropy. Either "bits" (default) or "nats".

**Value**

Information content of input variable

**Examples**

```
data(example_data)
information_content_contin(example_data$tumoursize)
```

---

`information_content_discrete`*Calculate Information Content (Discrete Variable)*

---

**Description**

Calculates information content of a discrete (categorical or ordinal) vector in bits (default) or natural units. Missing values are omitted from the calculation.

**Usage**

```
information_content_discrete(x, unit = c("bits"))
```

**Arguments**

<code>x</code>	Input vector
<code>unit</code>	Unit to measure entropy. Either "bits" (default) or "nats".

**Value**

Information content of input variable

**Examples**

```
data(example_data)
information_content_discrete(example_data$marital_status)
```

---

join_vars_to_ontol	<i>Join Mapping Table to Ontology Network Graph</i>
--------------------	---

---

**Description**

This function creates new nodes representing dataset variables and joins them to an input ontology network using a mapping file. Prior to joining, the information content of all nodes is calculated using [node\\_IC\\_zhou](#).

**Usage**

```
join_vars_to_ontol(ontol_graph, var2entity_tbl, mode = "in", root, k = 0.5)
```

**Arguments**

ontol_graph	Graph containing the chosen ontology. Must be in tidygraph format or coercible to this format.
var2entity_tbl	Edge table containing dataset variable names in first column and entities in ontologies to which they are mapped in the second column.
mode	Character constant specifying the directionality of the edges. One of "in" or "out".
root	name of root node identifier in column 1 to calculate node depth from.
k	numeric value to adjust the weight of the two items of information content equation (relative number of hyponyms/descendants and relative node depth). Default = 0.5

**Details**

- The user-defined mappings between variables in a dataset and entities/terms in an ontology are provided in an edge table (`var2entity_tbl`).
- A node attribute column, `node_category` is generated to describe if a node is one of "Dataset Variable", "Annotation", or "Annotation Ancestor".

**Value**

A [tidygraph](#) resulting from the joining of `var2entity_tbl` and `ontol_graph`.

**See Also**

[node\\_IC\\_zhou](#)

Other semantic enrichment functions: [metavariable\\_agg\(\)](#), [metavariable\\_info\(\)](#)

**Examples**

```
data(example_ontology)
join_vars_to_ontol(example_ontology, example_mapping_file, root = "root", mode = "in")
```



---

max_catchNAs	<i>Find maximum of vector safely</i>
--------------	--------------------------------------

---

**Description**

This low-level function is deployed as part of the semantic enrichment process. Calculates maximum of values in numeric vector (ignoring NAs). If all values in input vector are NA, returns NA (rather than -Inf),

**Usage**

```
max_catchNAs(x)
```

**Arguments**

x	numeric vector
---	----------------

**Value**

maximum value of x

**Examples**

```
x <- c(1,2,3)
eHDPrep:::max_catchNAs(x)
```

---

mean_catchNAs	<i>Find mean of vector safely</i>
---------------	-----------------------------------

---

**Description**

This low-level function is deployed as part of the semantic enrichment process. Averages values in numeric vector (ignoring NAs). If all values in numeric vector are NA, returns NA (rather than NaN),

**Usage**

```
mean_catchNAs(x)
```

**Arguments**

x	numeric vector
---	----------------

**Value**

mean of x

## Examples

```
x <- c(1,2,3)
eHDPrep::mean_catchNAs(x)

# with NA present:
x <- c(1,2,3, NA)
eHDPrep::mean_catchNAs(x)
```

---

merge_cols	<i>Merge columns in data frame</i>
------------	------------------------------------

---

## Description

Merges two columns in a single data frame. The merging draws on the functionality of dplyr's [coalesce](#) where missing values from one vector are replaced by corresponding values in a second variable. The name of the merged variable is specified in `merge_var_name`. `primary_var` and `secondary_var` can be removed with `rm_in_vars`. Variables must be combinable (i.e. not a combination of numeric and character).

## Usage

```
merge_cols(
  data,
  primary_var,
  secondary_var,
  merge_var_name = NULL,
  rm_in_vars = FALSE
)
```

## Arguments

<code>data</code>	data frame containing <code>primary_var</code> and <code>secondary_var</code> .
<code>primary_var</code>	Data variable which contains the best quality / most detailed information. Missing values will be supplied by values in corresponding rows from <code>secondary_var</code> .
<code>secondary_var</code>	Data variable which will be used to fill missing values in <code>primary_var</code> .
<code>merge_var_name</code>	character constant. Name for merged variable. Default: <code>[primary_var]_[secondary_var]_merged</code>
<code>rm_in_vars</code>	logical constant. Should <code>primary_var</code> and <code>secondary_var</code> be removed? Default = <code>FALSE</code> .

## Value

data frame with coalesced `primary_var` and `secondary_var`

## See Also

[coalesce](#)

## Examples

```
data(example_data)

# preserve input variables (default)
res <- merge_cols(example_data, diabetes_type, diabetes)
dplyr::select(res, dplyr::starts_with("diabetes"))

# remove input variables
res <- merge_cols(example_data, diabetes_type, diabetes, rm_in_vars = TRUE)
dplyr::select(res, dplyr::starts_with("diabetes"))
```

---

metavariable_agg	<i>Aggregate Data by Metavariable</i>
------------------	---------------------------------------

---

## Description

Variables in a numeric data frame are aggregated into metavariables via their most informative common ancestors identified in an ontological graph object (see [metavariable\\_info](#)). Metavariables are appended to the data frame.

## Usage

```
metavariable_agg(graph, data, label_attr = "name", normalize_vals = TRUE)
```

## Arguments

graph	Graph containing ontological and dataset nodes. Must be in <a href="#">tidygraph</a> format or coercible to this format. Must have been processed using <a href="#">metavariable_info</a> .
data	Numeric data frame or matrix containing variables which are also in graph.
label_attr	Node attribute containing labels used for column names when creating metavariable aggregations. Default: "name"
normalize_vals	Should values be normalized before aggregation? Default: TRUE

## Details

Metavariables are created from the aggregation of data variables via their most informative common ancestor (expected to have been calculated in [metavariable\\_info](#)). Metavariables are labelled using the syntax: MV\_[label\_attr]\_[Aggregation function]. The data variables are aggregated row-wise by their maximum, minimum, mean, sum, and product. Metavariables with zero entropy (no information) are not appended to the data. See examples for where this function should be applied in the semantic enrichment workflow.

## Value

data with semantic aggregations derived from common ontological ancestry (metavariables) appended as new columns, each prefixed with "MV\_" and suffixed by their aggregation function (e.g. "\_SUM").

**Note**

A warning may be shown regarding the '.add' argument being deprecated, this is believed to be an issue with tidygraph which may be resolved in a future release: <<https://github.com/thomasp85/tidygraph/issues/131>>. Another warning may be shown regarding the 'neimode' argument being deprecated, this is believed to be an issue with tidygraph which may be resolved in a future release: <<https://github.com/thomasp85/tidygraph/issues/>>. These warning messages are not believed to have an effect on the functionality of eHDPRep.

**See Also**

Other semantic enrichment functions: `join_vars_to_ontol()`, `metavariable_info()`

**Examples**

```
require(magrittr)
require(dplyr)
data(example_ontology)
data(example_mapping_file)
data(example_data)

#' # define datatypes
tibble::tribble(~"var", ~"datatype",
  "patient_id", "id",
  "tumoursize", "numeric",
  "t_stage", "ordinal_tstage",
  "n_stage", "ordinal_nstage",
  "diabetes_merged", "character",
  "hypertension", "factor",
  "rural_urban", "factor",
  "marital_status", "factor",
  "SNP_a", "genotype",
  "SNP_b", "genotype",
  "free_text", "freetext") -> data_types

# create post-QC data
example_data %>%
  merge_cols(diabetes_type, diabetes, "diabetes_merged", rm_in_vars = TRUE) %>%
  apply_quality_ctrl(patient_id, data_types,
    bin_cats = c("No" = "Yes", "rural" = "urban"),
    to_numeric_matrix = TRUE) %>%
  suppressMessages() ->
  post_qc_data

# minimal example on first four columns of example data:
dplyr::slice(example_ontology, 1:7, 24) %>%
  join_vars_to_ontol(example_mapping_file[1:3,], root = "root") %>%
  metavariable_info() %>%
  metavariable_agg(post_qc_data[1:10, 1:4]) -> res
# see Note section of documentation for information on possible warnings.

# summary of result:
tibble::glimpse(res)

## Not run:
# full example:
example_ontology %>%
  join_vars_to_ontol(example_mapping_file, root = "root") %>%
```

```

    metavariable_info() %>%
    metavariable_agg(post_qc_data) -> res
# see Note section of documentation for information on possible warnings.

# summary of result:
tibble::glimpse(res)

## End(Not run)

```

---

metavariable_info	<i>Compute Metavariable Information</i>
-------------------	---

---

## Description

Calculates attributes for each node in a graph object pertaining to their suitability and rank as metavariables; primarily if they are the most informative common ancestor (see [node\\_IC\\_zhou](#)) of a set of nodes representing a dataset variable.

## Usage

```
metavariable_info(graph, mode = "in")
```

## Arguments

graph	Graph containing ontological and dataset nodes. Must be in <a href="#">tidygraph</a> format or coercible to this format.
mode	Character constant specifying the directionality of the edges. One of: "in" or "out".

## Details

The added attributes are:

**min\_dist\_to\_var** Integer. The minimum distance of an ontology node in the graph to a node representing a dataset variable.

**is\_metavariable** Logical. If the node has at least two descendants in the graph which represent dataset variables.

**variable\_descendants** List. The names of variables of which a node is an ancestor.

**variable\_set** Integer. An identifier for the unique set of descendants in the graph which represent dataset variables. The assigned number corresponds to the order in which a unique set was identified when scanning through the node table.

**highest\_IC** Logical. If the node possesses the highest information content of all other nodes which are common ancestors of the same variable set. Information content is expected to have been calculated in [join\\_vars\\_to\\_ontol](#).

## Value

A modified graph object with additional node attributes pertaining to their status as a metavariable.

**See Also**[node\\_IC\\_zhou](#)Other semantic enrichment functions: [join\\_vars\\_to\\_ontol\(\)](#), [metavariable\\_agg\(\)](#)**Examples**

```
data(example_ontology)
require(magrittr)
example_ontology %>%
  join_vars_to_ontol(example_mapping_file, root = "root") -> joined_ontol

metavariable_info(joined_ontol)
```

---

min_catchNAs	<i>Find minimum of vector safely</i>
--------------	--------------------------------------

---

**Description**

This low-level function is deployed as part of the semantic enrichment process. Calculates minimum of values in numeric vector (ignoring NAs). If all values in numeric vector are NA, returns NA (rather than Inf),

**Usage**

```
min_catchNAs(x)
```

**Arguments**

x	numeric vector
---	----------------

**Value**

minimum value of x

**Examples**

```
x <- c(1,2,3)
eHDPrep::min_catchNAs(x)
```

---

mi_content_discrete	<i>Calculate Mutual Information Content</i>
---------------------	---

---

**Description**

Calculates mutual information content between two variables in bits. Missing values are omitted from the calculation.

**Usage**

```
mi_content_discrete(x, y)
```

**Arguments**

x	First variable
y	Second variable

**Value**

Mutual information content of x and y

**Examples**

```
data(example_data)
mi_content_discrete(example_data$diabetes, example_data$diabetes_type)
```

---

mod_track	<i>Data modification tracking</i>
-----------	-----------------------------------

---

**Description**

This function produces a table where each row represents a value in a variable which is present in the cleaned dataset and which has been modified. The identifier, original and modified value, modification type, and variable names in the original and modified datasets are recorded.

**Usage**

```
mod_track(before_tbl, after_tbl, id_var, plot = FALSE, vars2compare)
```

**Arguments**

before_tbl	Data frame from before modifications were made.
after_tbl	Data frame from after modifications were made.
id_var	An unquoted expression which corresponds to a variable in both before_tbl and after_tbl which identifies each row. Required.
plot	Should a plot be returned instead of a table of results? Default: FALSE.
vars2compare	Character vectors of variable names to compare.

**Value**

Table containing row-level modification records or plot summarising modifications.

**Examples**

```
# merge data as the example modification

require(magrittr)

# example with one modification type (removal)
# return table
mod_track(example_data, strings_to_NA(example_data), patient_id)

# return plot
mod_track(example_data, strings_to_NA(example_data), patient_id, plot = TRUE)

# example with multiple modification types (removal, substitution and addition)
example_data %>%
  strings_to_NA() %>%
  merge_cols(diabetes_type, diabetes) ->
  modded_data

# return table
mod_track(example_data, modded_data, patient_id, vars2compare = c("t_stage",
"diabetes_type_diabetes_merged" = "diabetes", "diabetes_type_diabetes_merged"
= "diabetes_type"), plot = FALSE)

# return plot
mod_track(example_data, modded_data, patient_id, vars2compare = c("t_stage",
"diabetes_type_diabetes_merged" = "diabetes", "diabetes_type_diabetes_merged"
= "diabetes_type"), plot = TRUE)
```

---

node\_IC\_zhou

---

*Calculate Node Information Content (Zhou et al 2008 method)*


---

**Description**

Computes the information content for each node in a directed graph according to the equation developed by Zhou *et al.* (2008).

**Usage**

```
node_IC_zhou(graph, mode = "in", root, k = 0.5)
```

**Arguments**

graph	tidygraph directed graph.
mode	Character constant specifying the directionality of the edges. One of "in" or "out".
root	name of root node identifier in column 1 to calculate node depth from.
k	numeric value to adjust the weight of the two items of information content equation (relative number of hyponyms/descendants and relative node depth). Default = 0.5



**Value**

tidygraph with additional node attribute "information\_content"

**Note**

For use in semantic enrichment, this should be applied before joining an ontology with nodes representing data variables (i.e. before applying [join\\_vars\\_to\\_ontol](#).

**References**

Zhou, Z., Wang, Y. & Gu, J. A New Model of Information Content for Semantic Similarity in WordNet. in 2008 Second International Conference on Future Generation Communication and Networking Symposia vol. 3 85–89 (2008).

**Examples**

```
data(example_ontology)
node_IC_zhou(example_ontology, mode = "in", root = "root")
```

---

normalize	<i>Min max normalization</i>
-----------	------------------------------

---

**Description**

Normalizes values in x to be between 0 and 1 using min-max normalization.

**Usage**

```
normalize(x, na.rm = TRUE)
```

**Arguments**

- x                    numeric vector
- na.rm                a logical indicating whether missing values should be removed. Default = TRUE.

**Value**

normalised x

---

nums_to_NA	<i>Replace numeric values in numeric columns with NA</i>
------------	--

---

### Description

Replaces specified numbers in numeric columns with NA.

### Usage

```
nums_to_NA(data, ..., nums_to_replace = NULL)
```

### Arguments

data	A data frame, data frame extension (e.g. a tibble), or a lazy data frame (e.g. from dbplyr or dtplyr).
...	<tidy-select> One or more unquoted expressions separated by commas. Variable names can be used as if they were positions in the data frame, so expressions like x:y can be used to select a range of variables.
nums_to_replace	numeric vector of values to be replaced with NA. Case is ignored.

### Details

Columns to process can be specified in ... or the function will be applied to all numeric columns.

### Value

data with specified values replaced with NA

### Examples

```
data(example_data)

# replace all 1,2, and 3 from tumoursize and patient_id with NA.
nums_to_NA(data = example_data, tumoursize, patient_id, nums_to_replace = c(1,2,3))
```

---

onehot_vec	<i>One hot encode a vector</i>
------------	--------------------------------

---

### Description

Uses one-hot encoding to convert nominal vectors to a tibble containing variables for each of the unique values in input vector.

### Usage

```
onehot_vec(x, prefix)
```

**Arguments**

x	non-numeric vector
prefix	prefix to append to output variable names

**Value**

tibble

**Examples**

```
vec <- c("blue", "red", "green", "red", "green")
eHDPrep::onehot_vec(vec, "example")
```

---

ordinal\_label\_levels    *Extract labels and levels of ordinal variables in a dataset*

---

**Description**

This function enables preservation of the text labels for ordinal variables in a dataset in preparation for conversion to a numeric matrix. A table is produced which retains the mappings between the text labels and the numerical labels for future reference.

**Usage**

```
ordinal_label_levels(data, out_path = NULL)
```

**Arguments**

data	data frame with ordinal variables with labels and levels to be extracted.
out_path	Optional string. Path to write output to. If not supplied, R object will be returned.

**Value**

Tibble of text label and (numerical) level mappings

**Examples**

```
require(magrittr) # for %>%

# create an example class_tbl object
# note that diabetes_type is classed as ordinal yet is not modified as its
# levels are not pre-coded. It should instead be encoded with encode_ordinals().
tibble::tribble(~"var", ~"datatype",
  "patient_id", "id",
  "tumoursize", "numeric",
  "t_stage", "ordinal_tstage",
  "n_stage", "ordinal_nstage",
  "diabetes", "factor",
  "diabetes_type", "ordinal",
  "hypertension", "factor",
  "rural_urban", "factor",
```

```

"marital_status", "factor",
"SNP_a", "genotype",
"SNP_b", "genotype",
"free_text", "freetext") -> data_types

# show unique values for t_stage in pre-QC example_data
unique(example_data$t_stage)

# apply quality control to example_data
apply_quality_ctrl(example_data, patient_id, data_types,
bin_cats =c("No" = "Yes", "rural" = "urban"), min_freq = 0.6) %>%
ordinal_label_levels -> res

# examine the labels and levels of t_stage in post-QC example_data
dplyr::filter(res, variable == "t_stage")

```

---

plot_completeness	<i>Plot Completeness of a Dataset</i>
-------------------	---------------------------------------

---

## Description

Generates a bar plot of percentage completeness for one or both data frame dimensions (rows/columns).

## Usage

```
plot_completeness(data, id_var, plot = c("variables", "rows"))
```

## Arguments

data	Data frame in tidy format (see <a href="https://tidyr.tidyverse.org/">https://tidyr.tidyverse.org/</a> ).
id_var	Row identifier variable name.
plot	Character vector containing one or both of variables and rows.

## Value

Completeness bar plot.

## See Also

Other measures of completeness: [assess\\_completeness\(\)](#), [compare\\_completeness\(\)](#), [completeness\\_heatmap\(\)](#), [row\\_completeness\(\)](#), [variable\\_completeness\(\)](#)

## Examples

```

data(example_data)
plot_completeness(example_data, patient_id, "variables")

```

---

prod_catchNAs	<i>Find product of vector safely</i>
---------------	--------------------------------------

---

**Description**

This low-level function is deployed as part of the semantic enrichment process. Calculates product of values in numeric vector (ignoring NAs). If all values in numeric vector are NA, returns NA (rather than Inf),

**Usage**

```
prod_catchNAs(x)
```

**Arguments**

x	numeric vector
---	----------------

**Value**

product of x

**Examples**

```
x <- c(1,2,3)
eHDPrep:::prod_catchNAs(x)
```

---

report_var_mods	<i>Track changes to dataset variables</i>
-----------------	---

---

**Description**

Reports if variables have been added, removed, or are preserved between two data frames. Intended to be used to review quality control / data preparation.

**Usage**

```
report_var_mods(before_tbl = NULL, after_tbl = NULL)
```

**Arguments**

before_tbl	Data frame from before modifications were made.
after_tbl	Data frame from after modifications were made.

**Value**

Tibble containing two columns. 'variable' contains name of each variable. 'presence' contains the presence of the variable in after\_tbl.

**Examples**

```
example_data_merged <- merge_cols(example_data, diabetes_type,
  diabetes, "diabetes_merged", rm_in_vars = TRUE)

report_var_mods(example_data, example_data_merged)
```

---

review_quality_ctrl	<i>Review Quality Control</i>
---------------------	-------------------------------

---

**Description**

Provides information on modifications made to a dataset at both variable (column) and value (sample) levels, designed for review of quality control measures.

**Usage**

```
review_quality_ctrl(before_tbl, after_tbl, id_var)
```

**Arguments**

before_tbl	Data frame from before modifications were made.
after_tbl	Data frame from after modifications were made.
id_var	An unquoted expression which corresponds to a variable in both before_tbl and after_tbl which identifies each row. Required.

**Details**

Modifications are identified by comparing the original and modified dataset.

QC review functions are applied in the following order:

1. Variable-level modifications ([report\\_var\\_mods](#))
2. Value-level modifications ([mod\\_track](#))
3. Value-level modifications (plot) ([mod\\_track](#))

A list containing each of these functions' outputs is returned.

**Value**

List containing data for review of quality control

**See Also**

Other high level functionality: [apply\\_quality\\_ctrl\(\)](#), [assess\\_quality\(\)](#), [semantic\\_enrichment\(\)](#)

**Examples**

```

data(example_data)
require(tibble)

tibble::tribble(~"var", ~"datatype",
  "patient_id", "id",
  "tumoursize", "numeric",
  "t_stage", "ordinal_tstage",
  "n_stage", "ordinal_nstage",
  "diabetes", "factor",
  "diabetes_type", "ordinal",
  "hypertension", "factor",
  "rural_urban", "factor",
  "marital_status", "factor",
  "SNP_a", "genotype",
  "SNP_b", "genotype",
  "free_text", "freetext") -> data_types

# create QC'ed dataset
post_QC_example_data <- apply_quality_ctrl(example_data,
                                           patient_id,
                                           data_types,
                                           bin_cats = c("No" = "Yes",
                                                         "rural" = "urban"),
                                           min_freq = 0.6)

# review QC
QC_review <- review_quality_ctrl(before_tbl = example_data,
                                 after_tbl = post_QC_example_data,
                                 id_var = patient_id)

# view variable level changes
QC_review$variable_level_changes

# view value level changes
QC_review$value_level_changes

# view value level changes as a plot
QC_review$value_level_changes_plt

```

---

row\_completeness

---

*Calculate Row Completeness in a Data Frame*


---

**Description**

Calculates the completeness of each row/observation in a data frame.

**Usage**

```
row_completeness(data, id_var)
```

**Arguments**

data	Data frame.
id_var	Row identifier variable.

**Details**

Row completeness is measured by comparing the number of NA to non-NA values. Returns the count of NA as well as the percentage of NA values and the percentage completeness.

**Value**

Tibble detailing completeness statistics for each row in input data.

**See Also**

Other measures of completeness: [assess\\_completeness\(\)](#), [compare\\_completeness\(\)](#), [completeness\\_heatmap\(\)](#), [plot\\_completeness\(\)](#), [variable\\_completeness\(\)](#)

**Examples**

```
data(example_data)
row_completeness(example_data, patient_id)
```

---

semantic_enrichment	<i>Semantic enrichment</i>
---------------------	----------------------------

---

**Description**

Enriches a dataset with additional (meta-)variables derived from the semantic commonalities between variables (columns).

**Usage**

```
semantic_enrichment(data, ontology, mapping_file, mode = "in", root, ...)
```

**Arguments**

data	Numeric data frame or matrix containing variables present in the mapping file.
ontology	Graph containing the chosen ontology. Must be in <a href="#">tidygraph</a> format or coercible to this format.
mapping_file	Path to csv file or data frame containing mapping information. Should contain two columns only. The first column should contain column names, present in the data frame. The second column should contain the name of entities present in the ontology object.
mode	Character constant specifying the directionality of the edges. One of: "in" or "out".
root	name of root node identifier in column 1 to calculate node depth from.
...	additional arguments to pass to <a href="#">read_csv</a>



## Details

Semantic enrichment generates meta-variables from the aggregation of data variables (columns) via their most informative common ancestor. Meta-variables are labelled using the syntax: `MV_[label_attr]_[Aggregation function]`. The data variables are aggregated row-wise by their maximum, minimum, mean, sum, and product. Meta-variables with zero entropy (no information) are not appended to the data. See the "Semantic Enrichment" section in eHDPRep's vignette for more information: `vignette("Introduction_to_eHDPRep", package = "eHDPRep")`

This function requires three inputs:

**data** a numeric data frame or matrix

**ontology** an ontology in the form of a directed network graph

**mapping\_file** A mapping file/data frame linking variables in the data frame with nodes in the ontology object

## Value

Semantically enriched dataset

## Note

A warning may be shown regarding the 'add' argument being deprecated, this is believed to be an issue with tidygraph which may be resolved in a future release: <<https://github.com/thomasp85/tidygraph/issues/131>>. Another warning may be shown regarding the 'neimode' argument being deprecated, this is believed to be an issue with tidygraph which may be resolved in a future release: <<https://github.com/thomasp85/tidygraph/issues/>>. These warning messages are not believed to have an effect on the functionality of eHDPRep.

## See Also

Other high level functionality: `apply_quality_ctrl()`, `assess_quality()`, `review_quality_ctrl()`

## Examples

```
require(magrittr)
require(dplyr)
data(example_ontology)
data(example_mapping_file)
data(example_data)

#' # define datatypes
tibble::tribble(~"var", ~"datatype",
  "patient_id", "id",
  "tumoursize", "numeric",
  "t_stage", "ordinal_tstage",
  "n_stage", "ordinal_nstage",
  "diabetes_merged", "character",
  "hypertension", "factor",
  "rural_urban", "factor",
  "marital_status", "factor",
  "SNP_a", "genotype",
  "SNP_b", "genotype",
  "free_text", "freetext") -> data_types

# create post-QC data
example_data %>%
```

```

merge_cols(diabetes_type, diabetes, "diabetes_merged", rm_in_vars = TRUE) %>%
apply_quality_ctrl(patient_id, data_types,
  bin_cats = c("No" = "Yes", "rural" = "urban"),
  to_numeric_matrix = TRUE) %>%
suppressMessages() ->
post_qc_data

# minimal example on first four columns of example data:
semantic_enrichment(post_qc_data[1:10,1:4],
  dplyr::slice(example_ontology, 1:7,24),
  example_mapping_file[1:3,], root = "root") -> res
# see Note section of documentation for information on possible warnings.

# summary of result:
tibble::glimpse(res)

## Not run:
# full example:
res <- semantic_enrichment(post_qc_data, example_ontology,
  example_mapping_file, root = "root")
# see Note section of documentation for information on possible warnings.

## End(Not run)

```

---

skipgram\_append

*Append Skipgram Presence Variables to Dataset*


---

## Description

Adds new variables to data which report the presence of skipgrams (either those specified in skipgrams2append or, if not specified, skipgrams with a minimum frequency (min\_freq, default = 1)).

## Usage

```
skipgram_append(skipgram_tokens, skipgrams2append, data, id_var, min_freq = 1)
```

## Arguments

skipgram_tokens	Output of <a href="#">skipgram_identify</a> .
skipgrams2append	Which skipgrams in skipgram_tokens to append to dataset.
data	Data frame to append skipgram variables to.
id_var	An unquoted expression which corresponds to a variable in data which identifies each row.
min_freq	Minimum percentage frequency of skipgram occurrence to return. Default = 1.

## Value

data with additional variables describing presence of skipgrams

## References

- Guthrie, D., Allison, B., Liu, W., Guthrie, L. & Wilks, Y. A Closer Look at Skip-gram Modelling. in Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06) (European Language Resources Association (ELRA), 2006).
- Benoit K, Watanabe K, Wang H, Nulty P, Obeng A, Müller S, Matsuo A (2018). “quanteda: An R package for the quantitative analysis of textual data.” *Journal of Open Source Software*, \*3\*(30), 774. doi:10.21105/joss.00774 <<https://doi.org/10.21105/joss.00774>>, <<https://quanteda.io>>.
- Feinerer I, Hornik K (2020). *\_tm: Text Mining Package\_*. R package version 0.7-8, <<https://CRAN.R-project.org/package=tm>>.
- Ingo Feinerer, Kurt Hornik, and David Meyer (2008). Text Mining Infrastructure in R. *Journal of Statistical Software* 25(5): 1-54. URL: <https://www.jstatsoft.org/v25/i05/>.

## See Also

Principle underlying function: [tokens\\_ngrams](#)

Other free text functions: [extract\\_freetext\(\)](#), [skipgram\\_freq\(\)](#), [skipgram\\_identify\(\)](#)

## Examples

```
data(example_data)
# identify skipgrams
toks_m <- skipgram_identify(x = example_data$free_text,
                           ids = example_data$patient_id,
                           max_interrupt_words = 5)
# add skipgrams by minimum frequency
skipgram_append(toks_m,
                id_var = patient_id,
                min_freq = 0.6,
                data = example_data)
# add specific skipgrams
skipgram_append(toks_m,
                id_var = patient_id,
                skipgrams2append = c("sixteen_week", "bad_strain"),
                data = example_data)
```

---

skipgram\_freq

*Report Skipgram Frequency*

---

## Description

Measures the frequency of skipgrams (non-contiguous words in free text), reported in a tibble. Frequency is reported as both counts and percentages.

## Usage

```
skipgram_freq(skipgram_tokens, min_freq = 1)
```

## Arguments

skipgram\_tokens

Output of [skipgram\\_identify](#).

min\_freq

Minimum skipgram percentage frequency of occurrence to retain. Default = 1.

**Value**

Data frame containing frequency of skipgrams in absolute count and relative to the length of input variable.

**References**

Guthrie, D., Allison, B., Liu, W., Guthrie, L. & Wilks, Y. A Closer Look at Skip-gram Modelling. in Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06) (European Language Resources Association (ELRA), 2006).

Benoit K, Watanabe K, Wang H, Nulty P, Obeng A, Müller S, Matsuo A (2018). “quanteda: An R package for the quantitative analysis of textual data.” *Journal of Open Source Software*, \*3\*(30), 774. doi:10.21105/joss.00774 <<https://doi.org/10.21105/joss.00774>>, <<https://quanteda.io>>.

Feinerer I, Hornik K (2020). *\_tm: Text Mining Package\_*. R package version 0.7-8, <<https://CRAN.R-project.org/package=tm>>.

Ingo Feinerer, Kurt Hornik, and David Meyer (2008). Text Mining Infrastructure in R. *Journal of Statistical Software* 25(5): 1-54. URL: <https://www.jstatsoft.org/v25/i05/>.

**See Also**

Principle underlying function: [tokens\\_ngrams](#)

Other free text functions: [extract\\_freertext\(\)](#), [skipgram\\_append\(\)](#), [skipgram\\_identify\(\)](#)

**Examples**

```
data(example_data)
toks_m <- skipgram_identify(x = example_data$free_text,
                           ids = example_data$patient_id,
                           max_interrupt_words = 5)
skipgram_freq(toks_m, min_freq = 0.5)
```

---

skipgram_identify	<i>Identify Neighbouring Words (Skipgrams) in a free-text vector</i>
-------------------	--

---

**Description**

Identifies words which appear near each other in the free-text variable (var), referred to as "Skipgrams". Supported languages for stop words and stemming are danish, dutch, english, finnish, french, german, hungarian, italian, norwegian, portuguese, russian, spanish, and swedish.

**Usage**

```
skipgram_identify(
  x,
  ids,
  num_of_words = 2,
  max_interrupt_words = 2,
  words_to_rm = NULL,
  lan = "english"
)
```

**Arguments**

<code>x</code>	Free-text character vector to query.
<code>ids</code>	Character vector containing IDs for each element of <code>var</code> .
<code>num_of_words</code>	Number of words to consider for each returned skipgram. Default = 2.
<code>max_interrupt_words</code>	Maximum number of words which can interrupt proximal words. Default = 2.
<code>words_to_rm</code>	Character vector of words which should not be considered.
<code>lan</code>	Language of <code>var</code> . Default: english.

**Value**

Tibble containing skipgrams as variables and patient values as rows.

**References**

Guthrie, D., Allison, B., Liu, W., Guthrie, L. & Wilks, Y. A Closer Look at Skip-gram Modelling. in Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06) (European Language Resources Association (ELRA), 2006).

Benoit K, Watanabe K, Wang H, Nulty P, Obeng A, Müller S, Matsuo A (2018). “quanteda: An R package for the quantitative analysis of textual data.” *Journal of Open Source Software*, \*3\*(30), 774. doi:10.21105/joss.00774 <<https://doi.org/10.21105/joss.00774>>, <<https://quanteda.io>>.

Feinerer I, Hornik K (2020). *tm: Text Mining Package*. R package version 0.7-8, <<https://CRAN.R-project.org/package=tm>>.

Ingo Feinerer, Kurt Hornik, and David Meyer (2008). Text Mining Infrastructure in R. *Journal of Statistical Software* 25(5): 1-54. URL: <https://www.jstatsoft.org/v25/i05/>.

**See Also**

Principle underlying function: [tokens\\_ngrams](#)

Other free text functions: [extract\\_freetext\(\)](#), [skipgram\\_append\(\)](#), [skipgram\\_freq\(\)](#)

**Examples**

```
data(example_data)
skipgram_identify(x = example_data$free_text,
                  ids = example_data$patient_id,
                  max_interrupt_words = 5)
```

---

strings\_to\_NA

---

*Replace values in non-numeric columns with NA*


---

**Description**

Replaces specified or pre-defined strings in non-numeric columns with NA.

**Usage**

```
strings_to_NA(data, ..., strings_to_replace = NULL)
```

**Arguments**

<code>data</code>	A data frame, data frame extension (e.g. a tibble), or a lazy data frame (e.g. from <code>dbplyr</code> or <code>dtplyr</code> ).
<code>...</code>	<tidy-select> One or more unquoted expressions separated by commas. Variable names can be used as if they were positions in the data frame, so expressions like <code>x:y</code> can be used to select a range of variables.
<code>strings_to_replace</code>	character vector of values to be replaced with NA.

**Details**

Columns to process can be specified in custom arguments (`...`) or will be applied to all non-numeric columns. Default strings which will be replaced with NA are as follows: "Undetermined", "unknown", "missing", "fail", "fail / unknown", "equivocal", "equivocal / unknown", "\*". String search is made using `grepl` and supports `regex` so metacharacters (`.` `\` `|` `()` `[]` `{}` `^` `$` `*` `+` `?` `$`) should be escaped with a `"\"` prefix. Matches are case sensitive by default but can ignore case with the parameter: `ignore.case = TRUE` in `...`.

**Value**

data with specified values replaced with NA.

**Examples**

```
data(example_data)

# original unique values in diabetes column:
unique(example_data$diabetes)
# Using default values
res <- strings_to_NA(example_data)
unique(res$diabetes)

# original unique values in diabetes_type column:
unique(example_data$diabetes_type)
# Using custom values
res <- strings_to_NA(example_data, strings_to_replace = "Type I")
unique(res$diabetes_type)
```

---

sum\_catchNAs

*Sum vector safely for semantic enrichment*


---

**Description**

sums values in `x` (ignoring NAs). If all values in `x` are NA, returns NA (rather than 0),

**Usage**

```
sum_catchNAs(x)
```

**Arguments**

x                      numeric vector

**Value**

sum of x

**Examples**

```
x <- c(1,2,3)
eHDPrep:::sum_catchNAs(x)
```

---

validate\_consistency\_tbl

*Validate internal consistency table*

---

**Description**

Runs a series of checks on a table of internal consistency rules (see Consistency Table Requirements) in preparation for [identify\\_inconsistency](#).

**Usage**

```
validate_consistency_tbl(data, consis_tbl)
```

**Arguments**

data                      data frame which will be checked for internal consistency

consis\_tbl                data frame or tibble containing information on internal consistency rules (see "Consistency Table Requirements" section)

**Value**

Error message or successful validation message is printed. The dataset is returned invisibly.

**Consistency Table Requirements**

Table must have exactly five character columns. The columns should be ordered according to the list below which describes the values of each column:

1. First column name of data values that will be subject to consistency checking. String. Required.
2. Second column name of data values that will be subject to consistency checking. String. Required.
3. Logical test to compare columns one and two. One of: ">", ">=", "<", "<=", "==", "!=". String. Optional if columns 4 and 5 have non-NA values.
4. Either a single character string or a colon-separated range of numbers which should only appear in column A. Optional if column 3 has a non-NA value.
5. Either a single character string or a colon-separated range of numbers which should only appear in column B given the value/range specified in column 4. Optional if column 3 has a non-NA value.

Each row should detail one test to make. Therefore, either column 3 or columns 4 and 5 must contain non-NA values.

### See Also

Other internal consistency functions: `identify_inconsistency()`

### Examples

```
require(tibble)
# example with synthetic dataset on number of bean counters
# there is a lot going on in the function so a simple dataset aids this example
#
# creating `data`:
beans <- tibble::tibble(red_beans = 1:15,
  blue_beans = 1:15,
  total_beans = 1:15*2,
  red_bean_summary = c(rep("few_beans",9), rep("many_beans",6)))
#
# creating `consis_tbl`
bean_rules <- tibble::tribble(~varA, ~varB, ~lgl_test, ~varA_boundaries, ~varB_boundaries,
  "red_beans", "blue_beans", "==", NA, NA,
  "red_beans", "total_beans", "<=", NA, NA,
  "red_beans", "red_bean_summary", NA, "1:9", "few_beans",
  "red_beans", "red_bean_summary", NA, "10:15", "many_beans")

validate_consistency_tbl(beans, bean_rules)
```

---

validate_mapping_tbl	<i>Validate mapping table for semantic enrichment</i>
----------------------	---

---

### Description

Applies tests to a mapping table to ensure it is valid for use with the data frame and ontological graph, in preparation for semantic enrichment.

### Usage

```
validate_mapping_tbl(mapping_tbl, data, ontol_graph)
```

### Arguments

mapping_tbl	data frame. Contains two columns. First column contains variable names of a primary dataset. Second column contains entities in an ontological graph to which the primary dataset's variable names are mapped.
data	data frame. Primary dataset which contains variable names referred to in first column of the mapping table
ontol_graph	ontological graph which contains entity names/IDs referred to in second column of the mapping table

### Value

Any warnings and the mapping table returned invisibly



## Examples

```
data(example_mapping_file)
data(example_data)
data(example_ontology)

# warns that some variables don't correspond between the data and the mapping file:
eHDPrep::validate_mapping_tbl(example_mapping_file, example_data, example_ontology)
```

---

validate_ontol_nw	<i>Validate ontology network for semantic enrichment</i>
-------------------	--

---

## Description

Performs tests on a graph object in preparation for semantic enrichment.

## Usage

```
validate_ontol_nw(graph)
```

## Arguments

graph	graph object to validate.
-------	---------------------------

## Details

The tests are:

1. Is graph coercible to [tidygraph](#) format?
2. Is graph directed?
3. Does graph contains one component (is one ontology)?

## Value

input graph or validation errors

## Examples

```
data(example_ontology)
eHDPrep::validate_ontol_nw(example_ontology)
```

---

variable.bw.kde	<i>Variable bandwidth Kernel Density Estimation</i>
-----------------	---

---

## Description

Calculates variable bandwidth KDE using Abramson's two stage estimator.

## Usage

```
variable.bw.kde(x, output.domain = x, na.rm = FALSE, adjust.factor = 0.5)
```

## Arguments

x	A numeric vector of values for estimating density
output.domain	The domain of values over which to estimate the density. Defaults to x. To use the same domain of x values as R's density, set to NULL.
na.rm	Remove missing values if TRUE
adjust.factor	A scaling factor (exponent) applied to the variable bandwidth calculation. Larger factors result in greater deviation from the fixed bandwidth (a value of 0 gives the fixed bandwidth case).

## Details

Bandwidth is first calculated using Silverman's estimator, then refined in a second stage to allow local bandwidth variations in the data based on the initial estimate.

## Value

The kernel density estimate as a density object, compatible with R's density function.

## Author(s)

Alexander Lyulph Robert Lubbock, Ian Overton

## References

Abramson, I. S. On Bandwidth Variation in Kernel Estimates-A Square Root Law. Ann. Statist. 10, 1217-1223 (1982).

---

variable_completeness	<i>Calculate Variable Completeness in a Data Frame</i>
-----------------------	--

---

**Description**

Calculates the completeness of each variable in a data frame.

**Usage**

```
variable_completeness(data)
```

**Arguments**

data	Data frame.
------	-------------

**Details**

This is achieved by comparing the number of NA to non-NA values. Returns the count of NA as well as the percentage of NA values and the percentage completeness.

**Value**

Tibble detailing completeness statistics for each variable.

**See Also**

Other measures of completeness: [assess\\_completeness\(\)](#), [compare\\_completeness\(\)](#), [completeness\\_heatmap\(\)](#), [plot\\_completeness\(\)](#), [row\\_completeness\(\)](#)

**Examples**

```
data(example_data)
variable_completeness(example_data)
```

---

variable_entropy	<i>Calculate Entropy of Each Variable in Data Frame</i>
------------------	---

---

**Description**

Calculates Shannon entropy of all variables in a data frame in bits (default) or natural units. Missing values are omitted from the calculation.

**Usage**

```
variable_entropy(data, unit = "bits")
```

**Arguments**

data	Data Frame to compute on
unit	Unit to measure entropy. Either "bits" (default) or "nats".

**Value**

Named numeric vector containing entropy values

**References**

Shannon, C. E. A mathematical theory of communication. The Bell System Technical Journal 27, 379–423 (1948).

**Examples**

```
a <- matrix(c(c(1,1,1,1,1,1, 1,2,3,4,5,6)),ncol = 2, dimnames =
list(seq(1,6), c("no_entropy","entropy")))
variable_entropy(as.data.frame(a))
```

---

warn_missing_dots	<i>Missing dots warning</i>
-------------------	-----------------------------

---

**Description**

Internal function. Warns if dots (...) argument have not been supplied

**Usage**

```
warn_missing_dots(test)
```

**Arguments**

test	expression to test.
------	---------------------

**Value**

warning to user that no values were modified

---

zero_entropy_variables	<i>Identify variables with zero entropy</i>
------------------------	---

---

**Description**

Calculates Shannon entropy of variables in a data frame in bits (default) or natural units. Missing values are omitted from the calculation. Names of variables with zero entropy are returned.

**Usage**

```
zero_entropy_variables(data, unit = "bits")
```

**Arguments**

data	Data Frame to compute on
unit	Unit to measure entropy. Either "bits" (default) or "nats".

### **Value**

Character vector of variable names with zero entropy

### **References**

Shannon, C. E. A mathematical theory of communication. The Bell System Technical Journal 27, 379–423 (1948).

### **Examples**

```
data(example_data)
zero_entropy_variables(example_data)
```

# Index

- \* **datasets**
    - example\_data, 24
    - example\_mapping\_file, 25
    - example\_ontology, 25
  - \* **free text functions**
    - extract\_freetext, 26
    - skipgram\_append, 50
    - skipgram\_freq, 51
    - skipgram\_identify, 52
  - \* **high level functionality**
    - apply\_quality\_ctrl, 3
    - assess\_quality, 6
    - review\_quality\_ctrl, 46
    - semantic\_enrichment, 48
  - \* **import to/export from R functions**
    - export\_dataset, 26
    - import\_dataset, 29
  - \* **internal consistency functions**
    - identify\_inconsistency, 27
    - validate\_consistency\_tbl, 55
  - \* **measures of completeness**
    - assess\_completeness, 5
    - compare\_completeness, 10
    - completeness\_heatmap, 12
    - plot\_completeness, 44
    - row\_completeness, 47
    - variable\_completeness, 59
  - \* **networks**
    - discrete.mi, 15
  - \* **semantic enrichment functions**
    - join\_vars\_to\_ontol, 32
    - metavariable\_agg, 35
    - metavariable\_info, 37
- apply\_quality\_ctrl, 3, 7, 46, 49
- assess\_completeness, 5, 10, 13, 44, 48, 59
- assess\_quality, 4, 6, 46, 49
- assume\_var\_classes, 3, 8, 30
- cell\_spec, 9
- cellspec\_lgl, 9
- coalesce, 34
- compare\_completeness, 5, 10, 13, 44, 48, 59
- compare\_info\_content, 10, 11, 12
- compare\_info\_content\_plt, 11, 11
- completeness\_heatmap, 5, 6, 10, 12, 44, 48, 59
- count\_compare, 14
- discrete.mi, 15
- distant\_neg\_val, 13, 16
- encode\_as\_num\_mat, 4, 16
- encode\_bin\_cat\_vec, 18
- encode\_binary\_cats, 3, 4, 17
- encode\_cats, 4, 19
- encode\_genotype\_vec, 21
- encode\_genotypes, 4, 20
- encode\_ordinals, 4, 21
- entropy, 22
- exact.kde, 23
- example\_data, 24
- example\_mapping\_file, 25
- example\_ontology, 25
- export\_dataset, 26, 29
- extract\_freetext, 3, 4, 26, 51–53
- grepl, 54
- grid.newpage, 13
- identify\_inconsistency, 6, 27, 55, 56
- import\_dataset, 26, 29
- import\_var\_classes, 3, 8, 30
- information\_content\_contin, 31
- information\_content\_discrete, 11, 31
- join\_vars\_to\_ontol, 32, 36–38, 41
- kable, 9
- max\_catchNAs, 33
- mean\_catchNAs, 33
- merge\_cols, 34
- metavariable\_agg, 32, 35, 38
- metavariable\_info, 32, 35, 36, 37
- mi\_content\_discrete, 11, 39
- min\_catchNAs, 38
- mod\_track, 39, 46

node\_IC\_zhou, [32](#), [37](#), [38](#), [40](#)  
normalize, [41](#)  
nums\_to\_NA, [42](#)  
  
onehot\_vec, [42](#)  
ordinal\_label\_levels, [43](#)  
  
pheatmap, [13](#)  
plot\_completeness, [5](#), [6](#), [10](#), [13](#), [44](#), [48](#), [59](#)  
prod\_catchNAs, [45](#)  
  
read\_csv, [29](#), [48](#)  
read\_excel, [29](#)  
read\_tsv, [29](#)  
regex, [54](#)  
report\_var\_mods, [45](#), [46](#)  
review\_quality\_ctrl, [4](#), [7](#), [46](#), [49](#)  
row\_completeness, [5](#), [6](#), [10](#), [13](#), [44](#), [47](#), [59](#)  
  
semantic\_enrichment, [4](#), [7](#), [46](#), [48](#)  
skipgram\_append, [27](#), [50](#), [52](#), [53](#)  
skipgram\_freq, [27](#), [51](#), [51](#), [53](#)  
skipgram\_identify, [27](#), [50–52](#), [52](#)  
strings\_to\_NA, [4](#), [53](#)  
sum\_catchNAs, [54](#)  
  
tidygraph, [32](#), [35](#), [37](#), [48](#), [57](#)  
tokens\_ngrams, [27](#), [51–53](#)  
  
validate\_consistency\_tbl, [28](#), [55](#)  
validate\_mapping\_tbl, [56](#)  
validate\_ontol\_nw, [57](#)  
variable.bw.kde, [58](#)  
variable\_completeness, [5](#), [6](#), [10](#), [13](#), [44](#), [48](#),  
[59](#)  
variable\_entropy, [59](#)  
  
warn\_missing\_dots, [60](#)  
write\_csv, [26](#)  
write\_tsv, [26](#)  
  
zero\_entropy\_variables, [6](#), [60](#)