Package ‘pickmax’

July 15, 2025

Type Package

Title Split and Coalesce Duplicated Records

Version 0.1.0

Maintainer Sbonelo Chamane <SChamane@hsrc.ac.za>

Description Deduplicates datasets by retaining the most complete and informative records. Identifies duplicated entries based on a specified key column, calculates completeness scores for each row, and compares values within groups. When differences between duplicates exceed a user-defined threshold, records are split into unique IDs; otherwise, they are coalesced into a single, most complete entry. Returns a list containing the original duplicates, the split entries, and the final coalesced dataset. Useful for cleaning survey or administrative data where duplicated IDs may reflect minor data entry inconsistencies.

License GPL-3

Encoding UTF-8

Imports dplyr, rlang, magrittr

RoxygenNote 7.3.2

NeedsCompilation no

Author Sbonelo Chamane [aut, cre] (ORCID: 0000-0001-5350-5203),
Musawenkosi Mabaso [aut],
Ronel Sewpaul [aut],
Sean Jooste [aut],
Kutloano Skhosana [aut],
Khangelani Zuma [aut]

Repository CRAN

Date/Publication 2025-07-15 11:40:05 UTC

Contents

pickmax *Split and Coalesce Duplicated Records*

Description

Deduplicates datasets by retaining the most complete and informative records. Identifies duplicated entries based on a specified key column, calculates completeness scores for each row, and compares values within groups. When differences between duplicates exceed a user-defined threshold, records are split into unique IDs; otherwise, they are coalesced into a single, most complete entry. Returns a list containing the original duplicates, the split entries, and the final coalesced dataset. Useful for cleaning survey or administrative data where duplicated IDs may reflect minor data entry inconsistencies.

Usage

```
pickmax(df, key_col = "id", diff_cutoff = 0.5)
```

Arguments

`df`	A data frame or tibble.
`key_col`	Character. Name of the column to identify duplicates.
`diff_cutoff`	Numeric between 0 and 1. Proportion of comparable fields that must differ for a duplicated record to be split into its own ID. Differences below this cutoff are treated as acceptable and those rows will stay merged under the original key. Defaults to 0.5 (50 percent).

Details

This function: 1. Computes a completeness percentage for each record. 2. Flags duplicates and checks if the proportion of differing fields relative to the most complete record exceeds ‘diff_cutoff‘. - Records exceeding the threshold are split with new IDs. - Others are merged using the most complete non-NA values.

Value

A named list with three data frames:

`duplicates_df` All rows flagged as duplicates, ordered by completeness.

`split_df` Rows split off because they exceeded `diff_cutoff`.

`coalesced_df` The final deduplicated data frame, coalesced prioritising completeness.

Examples

```
# Create a small sample with real duplicates
df <- data.frame(
  id    = c(1, 1, 2, 2, 3, 4, 4),
  value = c(10, 10, NA, 20, 5, 3, 3),
  tag   = c("A", "A", NA, "B", "C", "X", NA),
  stringsAsFactors = FALSE
)

# Run pickmax with default diff_cutoff (50%)
res <- pickmax(df, key_col = "id", diff_cutoff = 0.5)

# Show the duplicates flagged
print(res$duplicates_df)

# Show records that got split per diff_cutoff
print(res$split_df)

# Show final cleaned dataset
print(res$coalesced_df)
```

Index