**Version:** 1.0

**Date:** 2016-12-17

**Title:** merge2()

**Author:**  Jeremy Guinta

**Depends:** dplyr

January 19, 2017

**Description:** Prepares a full outer join of two data frame objects and provides summary statistics and flagging of the merge on the output data.

R topics documented:

Contents

[Description 2](#_Toc470099319)

[Usage 2](#_Toc470099320)

[Arguments 2](#_Toc470099321)

[Variables Created 2](#_Toc470099322)

[Example 3](#_Toc470099323)

merge2() Better merging of data frames

# Description

This process standardizes the merge function and adds key summary statistics and flagging on output the dataframe after merging on two individual data frames. This function assumes the use of a two data frame objects in memory and can take any number of joining fields. The joining fields do not need to be named the same, but do need to be in the proper paired order to facilitate the merge. The function uses the base R merge() function and adds summary output and flagging to the output dataframe that indicates if the recorded matched, did not match and came from first dataframe specified or did not match and came from the second dataframe specified. Relies on dplyr functions to prepare summary output.

# Usage

merge2(dataframe1, datafame2, by.x=c(), by.y=c(), explosion=FALSE)

# Arguments

dataframe1 an object of class data.frame in which the selected

by.x fields can be found.

dataframe2 an object of class data.frame in which the selected

by.y fields can be found.

by.x a character vector of the variable names from data, dataframe1. The order of these should be the same order as by.y

by.y a character vector of the variable names from data, dataframe2. The order of these should be the same order as by.x

explosion Defaults to FALSE. When TRUE, this option will error if any records from either dataframe1 or dataframe2 explode.

# Variables Created

.m This variable can take on three values: 1, 2 or 3. 1 indicates a record that was contained in dataframe1, but not dataframe2. 2 indicates a record that was contained in dataframe2, but not dataframe1. 3 indicates a record that matched between dataframe1 and dataframe2.

.expl This variable can take on two values: 0 or 1. 0 indicates that the data did not explode when matched. 1 indicates that the data did explode when matched. This variable will be turned on when a many-to-one match occurs or when a many-to-many match occurs.

# Example

#Create Data

authors <- data.frame(

surname = I(c("Tukey", "Venables", "Tierney", "Ripley", "McNeil")),

nationality = c("US", "Australia", "US", "UK", "Australia"),

deceased = c("yes", rep("no", 4)))

books <- data.frame(

name = I(c("Tukey", "Venables", "Tierney",

"Ripley", "Ripley", "McNeil", "R Core")),

title = c("Exploratory Data Analysis",

"Modern Applied Statistics ...",

"LISP-STAT",

"Spatial Statistics", "Stochastic Simulation",

"Interactive Data Analysis",

"An Introduction to R"),

other.author = c(NA, "Ripley", NA, NA, NA, NA,

"Venables & Smith"))

#Merge

authors\_books<-merge2(authors, books, by.x=c("surname"), by.y=c("name"))

# [1] "Summary of Match"

# .m cnt descr

# 1 2 1 Righthand Side

# 2 3 6 Match

# [1] "Summary of Explosion"

# orig\_cnt matched\_cnt descr

# 1 5 6 Lefthand Side

# 2 7 7 Righthand Side

# 3 7 7 Total

table(authors\_books$.m, ifany="ifany")

#2 3

#1 6

table(authors\_books$.expl, useNA="ifany")

#0 1

#5 2

authors\_books

surname nationality deceased title other.author .m .expl

1 McNeil Australia no Interactive Data Analysis <NA> 3 0

2 R Core <NA> <NA> An Introduction to R Venables & Smith 2 0

3 Ripley UK no Spatial Statistics <NA> 3 1

4 Ripley UK no Stochastic Simulation <NA> 3 1

5 Tierney US no LISP-STAT <NA> 3 0

6 Tukey US yes Exploratory Data Analysis <NA> 3 0

7 Venables Australia no Modern Applied Statistics ... Ripley 3 0