Split reference list helper for pilot and collaborative screening rounds

Coralie Williams

2022-12-02

When screening for a systematic review or meta-analysis, we conduct several pilot screening rounds. Pilot screenings help us refine our search string, decision tree, and increase the overall accuracy of our screening1.

During a pilot screening, we want to select a random subset of references that would be a representative sample of the full set to determine whether our hit rate is sufficient. If possible, it is encouraged to work in collaboration with one or two other reviewers. An independent reviewer can perform duplicate screening or screen a subset of the full reference list.

There are two reasons we’d want to automate the selection and splitting of a reference list:

1. It is time consuming to randomly select papers (>100 papers is tedious to select by hand!)
2. We are not really good at selecting at random (actually computers aren’t really good at being truly random either\*)

Below is the R ([www.r-project.org](https://www.r-project.org/)) code to run two functions that may come in useful when conducting your pilot and collaborative screenings with Rayyan (<https://rayyan.ai/>), or any other software where you can upload your pilot reference list.

### 1. Select random pilot set:

First, load the getpilotref function below.

# -----------------------------------  
# getpilotref function   
# -----------------------------------  
## Description:   
# Function to obtain a random subset of references for pilot screening.  
#  
# Arguments  
# - x: data frame with reference list  
# - n: number of papers for pilot subset (default is 10)  
# - write: logical argument whether to save the pilot list as a csv file in current working directory (default is FALSE).  
# - fileName: name of file (default is "pilot")  
  
getpilotref <- function(x, n=10, write=FALSE, fileName="pilot"){  
   
 if (length(n) == 1L && n%%1==0 && n>0 && n<=nrow(x)) {   
   
 # sample randomly the vector n of row indexes and remove id column in the final dataset  
 x$ids <- 1:nrow(x)  
 pdat <- x[which(x$ids %in% sample(x$ids, n)),]  
 pilot <<- pdat[,-which(colnames(pdat)=="ids")]  
   
 } else {  
 # error message n value provided is not valid   
 stop("Incompatible value n supplied, please check. n must be a positive integer no higher than the total number of references provided.")   
 }  
   
 if (write==T){  
   
 # save generated pilot list in working directory using the name provided  
 write\_csv(pilot, paste(fileName, ".csv", sep=""), na="")  
   
 # print out summary of saved file name  
 cat(paste("Pilot random sample set of ", n, " articles is saved as: ", fileName, ".csv", sep=""))  
   
 }  
}

As an example, load this reference list of

# Load example butterfly reference list (exported as csv from Rayyan)  
articles<-read.csv("https://raw.githubusercontent.com/coraliewilliams/2022/main/data/articles\_butterfly.csv")

Now,

### 2. Split reference list with another collaborator

Load the splitref\_prop function in your environment:

# -----------------------------------  
# splitref\_prop function   
# -----------------------------------  
## Description:   
# Function to split in two a reference list based on input proportions.  
#  
## Arguments:   
# - x: data frame with reference list  
# - p: vector of two numerical proportions for each split, it must have two positive numerical values that sum to 1.  
# - write: logical argument whether to save the pilot list as csv in current working directory  
# - fileName: name to give to the suffix of the two split files.   
  
splitref\_prop <- function(x, p=c(0.5, 0.5), fileName = "split", write = F) {  
   
 if (length(p) == 2L && is.numeric(p) && sum(p) == 1 && all(p > 0)) {  
 # get random list of indexes for each reference  
 rids <- sample(1:nrow(x))  
   
 # get index of row to split on using the first value of proportion of vector  
 spl <- floor(p[-length(p)] \* nrow(x))  
   
 # get separate split data frame based on split ids indexes  
 indx1 <- rids[1:spl]  
 indx2 <- rids[(spl + 1):nrow(x)]  
   
 # get separate datasets  
 split1 <<- x[indx1,]  
 split2 <<- x[indx2,]  
   
 # print out message  
 cat(paste(c("Reference list was randomly split into",length(p), "proportions of", p[1]\*100, "% and", p[2]\*100, "%")))  
   
 if (write == T) {  
 # save files  
 write\_csv(split1, paste(fileName, "\_set1", ".csv", sep = ""), na ="")  
 write\_csv(split2, paste(fileName, "\_set2", ".csv", sep = ""), na ="")  
   
 # print out summary of splitting and data files saved  
 p <- ifelse(length(prop) == 1, "equal", "unequal")  
 cat(paste(c("\n",k,"files were saved:\n",paste(getwd(), "/", fileName, "\_set1", ".csv", sep = ""))))  
   
 }  
   
 } else {  
 # error message if provided n value is not valid  
 stop("Incompatible values for p (proportions) supplied, please check.  
 Proportion values must be positive integers less than 1, and the total sum of all proportions should equal to 1.")  
 }  
}

Work on more functionalities (e.g. split between multiple collaborators) is upcoming… Any comments, questions or feedback, please contact me at [coralie.williams@unsw.edu.au](mailto:coralie.williams@unsw.edu.au)

\*most often random number generators produced by computers are “pseudo-random”, meaning they are produced from a deterministic mathematical model. These pseudo-random numbers are usually good (and better than what any human replicate). In statistics, a selection process is random if each item of a set has an equal probability of being chosen.

There exist true random number generators based . If you’re curious check out these websites: <https://www.random.org/>; <https://qrng.anu.edu.au/random-colours/>

### notrandom

### References

1. Foo, Y. Z., O’Dea, R. E., Koricheva, J., Nakagawa, S. & Lagisz, M. A practical guide to question formation, systematic searching and study screening for literature reviews in ecology and evolution. Methods in Ecology and Evolution 12, 1705–1720 (2021).

(Foo et al., 2021)

Foo, Y. Z., O’Dea, R. E., Koricheva, J., Nakagawa, S., & Lagisz, M. (2021). A practical guide to question formation, systematic searching and study screening for literature reviews in ecology and evolution. *Methods in Ecology and Evolution*, *12*(9), 1705–1720. https://doi.org/10.1111/2041-210X.13654