Aprendizegem Automática I

Luís Freitas PG38347, Luís Maia

21/12/2020

Em baixo, consegue-se observar o código da escolha da diretoria do projeto, inserindo o caminho na variavél D

#Chose dir

D = "C:/workspaceschool/Project\_answerCP/"  
#D = "/mnt/c/workspaceschool/Project\_answerCP/"  
setwd(D)  
print(paste0("A diretoria do projeto é: ",D[1]))

## [1] "A diretoria do projeto é: C:/workspaceschool/Project\_answerCP/"

é apresentado no código em baico as extensões necessárias para o desenvolvimento do projeto

##Install Extensions

#This extension is used for the "countLines" function in Insert Data   
#install.packages("R.utils")  
library(R.utils)

## Warning: package 'R.utils' was built under R version 3.6.3

## Loading required package: R.oo

## Warning: package 'R.oo' was built under R version 3.6.3

## Loading required package: R.methodsS3

## Warning: package 'R.methodsS3' was built under R version 3.6.3

## R.methodsS3 v1.8.1 (2020-08-26 16:20:06 UTC) successfully loaded. See ?R.methodsS3 for help.

## R.oo v1.24.0 (2020-08-26 16:11:58 UTC) successfully loaded. See ?R.oo for help.

##   
## Attaching package: 'R.oo'

## The following object is masked from 'package:R.methodsS3':  
##   
## throw

## The following objects are masked from 'package:methods':  
##   
## getClasses, getMethods

## The following objects are masked from 'package:base':  
##   
## attach, detach, load, save

## R.utils v2.10.1 (2020-08-26 22:50:31 UTC) successfully loaded. See ?R.utils for help.

##   
## Attaching package: 'R.utils'

## The following object is masked from 'package:utils':  
##   
## timestamp

## The following objects are masked from 'package:base':  
##   
## cat, commandArgs, getOption, inherits, isOpen, nullfile, parse,  
## warnings

#This extension is used for the "laf\_open" function  
#install.packages("LaF")  
library(LaF)

## Warning: package 'LaF' was built under R version 3.6.3

Em baixo é apresentado o código de inserção dos dados, o dataset lectures, o dataset questions, e uma amostra do dataset train

##Insert Data

#insert dataset questions  
Filename1 ="database\_aa1/questions.csv"   
questions = read.csv2((paste0(D,Filename1)), sep = ",")  
  
#insert dataset lectures  
Filename2 ="database\_aa1/lectures.csv"  
lectures = read.csv2((paste0(D,Filename2)), sep = ",")  
  
#insert dataset train\_sample  
Filename3 ="database\_aa1/sample/train\_sample.csv"  
train\_sample = read.csv2(paste0(D,Filename3), sep = ",", header = FALSE)  
  
#insert headers in data frame train\_sample  
Filename4 ="database\_aa1/train.csv"  
columns\_name = read.csv2(paste0(D,Filename4), sep = ",", header = FALSE, nrows=1)  
for (i in 1:dim(train\_sample)[2]) {  
 names(train\_sample)[i] <- as.character(columns\_name[1,i])  
}  
  
#remove temporary variables  
rm(Filename1, Filename2, Filename3, Filename4, columns\_name, i)

Em baixo o dataset “questions” (explicar o dataset)

#head of dataset questions  
head(questions)

## question\_id bundle\_id correct\_answer part tags  
## 1 0 0 0 1 51 131 162 38  
## 2 1 1 1 1 131 36 81  
## 3 2 2 0 1 131 101 162 92  
## 4 3 3 0 1 131 149 162 29  
## 5 4 4 3 1 131 5 162 38  
## 6 5 5 2 1 131 149 162 81

Em baio o dataset “lectures”

#head of dataset lectures  
head(lectures)

## lecture\_id tag part type\_of  
## 1 89 159 5 concept  
## 2 100 70 1 concept  
## 3 185 45 6 concept  
## 4 192 79 5 solving question  
## 5 317 156 5 solving question  
## 6 335 114 2 concept

Em baixo o dataset “train”

#head of dataset train  
head(train\_sample)

## row\_id timestamp user\_id content\_id content\_type\_id task\_container\_id  
## 1 6685040 948334 146132347 10685 0 10  
## 2 91098206 103010638 1933608435 973 0 67  
## 3 42132926 10881193788 893185638 1002 0 1384  
## 4 24157444 30027741885 516901151 5068 0 494  
## 5 54513578 86079486 1155436014 8648 0 28  
## 6 12571626 3832206391 272339193 3382 0 1639  
## user\_answer answered\_correctly prior\_question\_elapsed\_time  
## 1 0 1 20000.0  
## 2 1 1 18000.0  
## 3 3 1 25000.0  
## 4 1 0 32000.0  
## 5 1 1 15000.0  
## 6 3 1 14333.0  
## prior\_question\_had\_explanation  
## 1 True  
## 2 True  
## 3 True  
## 4 True  
## 5 True  
## 6 True

#create function to take a sample from a dataset #sample1 <- function(file, n) { # lf <- laf\_open(detect\_dm\_csv(file, sep = “,”, header = TRUE, factor\_fraction = -1)) # return(read\_lines(lf, sample(1:nrow(lf), n))) #}

#count lines of a data set #n\_rows = countLines(“C:/workspaceschool/Project\_answerCP/database/train.csv”) #n\_rows = n\_rows[1] #cat(“This is a number of rows of dataset train:”,n\_rows) #train = sample1(“C:/workspaceschool/Project\_answerCP/database/train.csv”, 100) #train = sampleCSV(“train.csv”, nrLines = 100, header=TRUE) #train = read.csv.ffdf(file=“C:/workspaceschool/Project\_answerCP/database/train.csv”, header=TRUE, VERBOSE=TRUE, first.rows=10000, next.rows=50000)

##Data Preparation  
  
```r  
#Divide dataset train in questions and lectures   
train\_question = train\_sample[train\_sample$content\_type\_id == 0,]  
train\_lecture = train\_sample[train\_sample$content\_type\_id == 1,]  
#rm(train\_sample)  
  
  
  
names(questions)[1] <- "content\_id"  
names(lectures)[1] <- "content\_id"  
train\_question = merge(x = train\_question, y = questions, by = "content\_id", all.x = TRUE)  
train\_lecture = merge(x = train\_lecture, y = lectures, by = "content\_id", all.x = TRUE)  
  
head(train\_question)

## content\_id row\_id timestamp user\_id content\_type\_id task\_container\_id  
## 1 0 99476171 1189417585 2110788479 0 59  
## 2 0 83326546 32899659650 1775472333 0 2215  
## 3 0 59984442 9217027681 1273477311 0 797  
## 4 0 39540247 184154486 840831219 0 36  
## 5 0 86108085 3866013540 1833263456 0 466  
## 6 0 8729112 521308099 189923712 0 132  
## user\_answer answered\_correctly prior\_question\_elapsed\_time  
## 1 2 0 20000.0  
## 2 1 0 22000.0  
## 3 0 1 22000.0  
## 4 0 1 12000.0  
## 5 0 1 21000.0  
## 6 0 1 22000.0  
## prior\_question\_had\_explanation bundle\_id correct\_answer part tags  
## 1 True 0 0 1 51 131 162 38  
## 2 True 0 0 1 51 131 162 38  
## 3 True 0 0 1 51 131 162 38  
## 4 True 0 0 1 51 131 162 38  
## 5 True 0 0 1 51 131 162 38  
## 6 True 0 0 1 51 131 162 38

#head(train\_lecture)  
train\_question = train\_question[-c(1,2,5,7)]  
head(train\_question)

## timestamp user\_id task\_container\_id answered\_correctly  
## 1 1189417585 2110788479 59 0  
## 2 32899659650 1775472333 2215 0  
## 3 9217027681 1273477311 797 1  
## 4 184154486 840831219 36 1  
## 5 3866013540 1833263456 466 1  
## 6 521308099 189923712 132 1  
## prior\_question\_elapsed\_time prior\_question\_had\_explanation bundle\_id  
## 1 20000.0 True 0  
## 2 22000.0 True 0  
## 3 22000.0 True 0  
## 4 12000.0 True 0  
## 5 21000.0 True 0  
## 6 22000.0 True 0  
## correct\_answer part tags  
## 1 0 1 51 131 162 38  
## 2 0 1 51 131 162 38  
## 3 0 1 51 131 162 38  
## 4 0 1 51 131 162 38  
## 5 0 1 51 131 162 38  
## 6 0 1 51 131 162 38

#head(lectures)  
#train\_lecture = train\_lecture[-c(1,5,7,8,9,10)]  
#head(lectures)