PAC4\_neteja

Edgar Pardo

5/16/2020

# Introducció

Aquest algoritme correspon al procés de neteja del conjunt de dades que s’emprarà en el projecte de la matèria *M2.959 - Visualització de dades* del Màster en Data Science de la UOC.

## Procés de neteja

Primer de tot, importarem el fitxer provinent de *peaceagreeements.org*[1] de tots els acords de pau on es tracten temes de reconstrucció socioeconòmica:

dataset <- read.csv('D:\\UOC\\VD\\PAC4\\PAX\_dataset.csv',stringsAsFactors = FALSE)  
dim(dataset)[1]

## [1] 933

dim(dataset)[2]

## [1] 265

Comptem amb 933 acords de pau i 265 atributs per cadascun d’ells. El nostre objectiu serà reduir el nombre d’atributs a tots aquells que estan categoritzats com a *Basic information* al *Codebook*[2] i els que seran objecte d’estudi: *Human rights and Equality*.

Primer ens quedarem amb els atributs que representen aquesta informació bàsica:

column\_names <- c('ï..Con','Contp','PP','PPName','Reg','AgtId','Agt','Dat','Status','Lgt','N\_characters','Agtp','Stage','StageSub','Part','ThrdPart','OthAgr','Loc1ISO','Loc2ISO','Loc1GWNO','Loc2GWNO','UcdpCon','UcdpAgr','PamAgr','CowWar')  
final\_dataset <- dataset[,column\_names]  
cat('Tenim ',dim(final\_dataset)[2], ' atributs')

## Tenim 25 atributs

Ara tractarem a part els atributs corresponents a drets humans. Recodificarem les variables binàries i categòriques numèriques en cadenes de text, de manera que siguin fàcils d’interpretar i l’usuari no necessiti haver llegit el *Codebook*.

length(unique(dataset$AgtId)) #933 DISTINCT VALUES

## [1] 933

column\_names\_hr <- c('HrGen','EqGen','HrDem','HrFra','HrfSp','HrfBor','HrfTinc','AgtId')  
hr\_dataset <- dataset[,column\_names\_hr]

En aquest subset del dataframe original no només ens hem quedat amb les variables de drets humans a estudiar, sinó que també comptem amb la variable *AgtId* perquè ens servirà posteriorment per fer una join amb el dataset final. Hem comprovat que tots els valor d’aquest últim atribut siguin únics per garantir que en ser unequívocs cada registre del subset anirà a un registre del dataset final.

La codificació serà la següent:

* Variable categòrica binària
  + Valor 0 passarà a etiquetar-se com a “Not mentioned”
  + Valor 1 passarà a etiquetar-se com a “Mentioned”
* Variable categòrica numèrica
  + Valor 0 passarà a etiquetar-se com a “Not provisioned”
  + Valor 1 passarà a etiquetar-se com a “Rhetorical provisions”
  + Valor 2 passarà a etiquetar-se com a “Substantive provisions”
  + Valor 3 passarà a etiquetar-se com a “Detailed substantive provisions”

hr\_dataset <- hr\_dataset %>% mutate(HrGen = ifelse(HrGen == "1",'Mentioned','Not mentioned'))  
hr\_dataset <- hr\_dataset %>% mutate(HrfSp = ifelse(HrfSp == "1",'Mentioned','Not mentioned'))  
hr\_dataset <- hr\_dataset %>% mutate(HrfBor = ifelse(HrGen == "1",'Mentioned','Not mentioned'))  
hr\_dataset <- hr\_dataset %>% mutate(HrfTinc = ifelse(HrGen == "1",'Mentioned','Not mentioned'))  
hr\_dataset$EqGen <- factor(hr\_dataset$EqGen)  
levels(hr\_dataset$EqGen) <- list(  
 'Not provisioned' = c("0"),  
 'Rhetorical provisions' = c("1"),  
 'Substantive provisions' = c("2"),  
 'Detailed substantive provisions' = c("3"))  
  
hr\_dataset$HrDem <- factor(hr\_dataset$HrDem)  
levels(hr\_dataset$HrDem) <- list(  
 'Not provisioned' = c("0"),  
 'Rhetorical provisions' = c("1"),  
 'Substantive provisions' = c("2"),  
 'Detailed substantive provisions' = c("3"))  
  
hr\_dataset$HrFra <- factor(hr\_dataset$HrFra)  
levels(hr\_dataset$HrFra) <- list(  
 'Not provisioned' = c("0"),  
 'Rhetorical provisions' = c("1"),  
 'Substantive provisions' = c("2"),  
 'Detailed substantive provisions' = c("3"))

unique(hr\_dataset$HrDem) ## Una variable amb 4 categories

## [1] Substantive provisions Rhetorical provisions   
## [3] Not provisioned Detailed substantive provisions  
## 4 Levels: Not provisioned ... Detailed substantive provisions

unique(hr\_dataset$HrGen) ## Una variable binària

## [1] "Mentioned" "Not mentioned"

Un cop transformat és moment de fer la join:

final\_dataset <- merge(final\_dataset,hr\_dataset )  
dim(final\_dataset)[1]

## [1] 933

dim(final\_dataset)[2]

## [1] 32

El resultat final ha estat 933 acords de pau amb 32 atributs. Ara és moment de exportar-lo a csv per tractar-lo amb la nostra eina de visualització de dades.

write.csv(final\_dataset, file = "PAX\_clean.csv")

### Bibliografia

[1] Peaceagreements.org. 2020. PA-X: Peace Agreements Database - Site. [online] Disponible a: <https://www.peaceagreements.org/> [Accedit 16 maig 2020]. [2] Peaceagreements.org. 2020. [online] Disponible a: <https://www.peaceagreements.org/files/PAX%> 20codebook%20Version3.pdf [Accedit 16 maig 2020].