Cleaning Excerpts - Set 2

Jorge Olmos

15/10/2020

1. Set up: libraries, directory and dataset

1.2 Renaming variable “Type of Response: Missing” so all datasets have the same name for this var. Then, rbind the datasets into one by matching column names.

AO <- AO %>% setNames( gsub("[ [:punct:]]", "\_", colnames(AO)))  
AO <- AO %>% setNames( gsub("\_\_", "\_", colnames(AO), fixed = T))  
AP <- AP %>% setNames( gsub("[ [:punct:]]", "\_", colnames(AP)))  
AP <- AP %>% setNames( gsub("\_\_", "\_", colnames(AP), fixed = T))  
DS <- DS %>% setNames( gsub("[ [:punct:]]", "\_", colnames(DS)))  
DS <- DS %>% setNames( gsub("\_\_", "\_", colnames(DS), fixed = T))  
DZ <- DZ %>% setNames( gsub("[ [:punct:]]", "\_", colnames(DZ)))  
DZ <- DZ %>% setNames( gsub("\_\_", "\_", colnames(DZ), fixed = T))  
NI <- NI %>% setNames( gsub("[ [:punct:]]", "\_", colnames(NI)))  
NI <- NI %>% setNames( gsub("\_\_", "\_", colnames(NI), fixed = T))  
  
  
# Renaming var "type of response: missing" to match AO  
AP <- AP %>% dplyr::rename(Code\_Res\_Missing\_Applied = Code\_Type\_of\_Response\_Res\_Missing\_Applied)  
DS <- DS %>% dplyr::rename(Code\_Res\_Missing\_Applied = Code\_Type\_of\_Response\_Res\_Missing\_Applied)  
DZ <- DZ %>% dplyr::rename(Code\_Res\_Missing\_Applied = Code\_Type\_of\_Response\_Res\_Missing\_Applied)  
NI <- NI %>% dplyr::rename(Code\_Res\_Missing\_Applied = Code\_Type\_of\_Response\_Res\_Missing\_Applied)  
  
  
# Function to combine dataframes, keeping only the columns that have the same names. This identifies which columns in the smaller dataframe match the columns in the larger one. https://amywhiteheadresearch.wordpress.com/2013/05/13/combining-dataframes-when-the-columns-dont-match/  
  
rbind.match.columns <- function(input1, input2) {  
 n.input1 <- ncol(input1)  
 n.input2 <- ncol(input2)  
   
 if (n.input2 < n.input1) {  
 TF.names <- which(names(input2) %in% names(input1))  
 column.names <- names(input2[, TF.names])  
 } else {  
 TF.names <- which(names(input1) %in% names(input2))  
 column.names <- names(input1[, TF.names])  
 }  
   
 return(rbind(input1[, column.names], input2[, column.names]))  
}  
  
APDS <- rbind.match.columns(AP, DS) # 1,082 OBS OF 128 VAR  
AP\_DS\_DZ <- rbind.match.columns(APDS, DZ) # 1,754 of 128 var  
AP\_DS\_DZ\_NI <- rbind.match.columns(AP\_DS\_DZ, NI) # 2,266 obs of 128 var  
exc0 <- rbind.match.columns(AP\_DS\_DZ\_NI, AO) # 2,896 obs of 126 var

1. Removing columns (vars) with no relevant information and rename remaining ones

exc1 <- select(exc0, Media\_Title:Excerpt\_Copy,  
 grep("\_Applied$", colnames(exc0)),  
 Code\_Crime\_CrimeID\_Weight,  
 Codes\_Applied\_Combined,  
 Code\_Crime\_Speakers\_relationship\_to\_crime\_Applied,  
 Code\_Crime\_Speakers\_relationship\_to\_crime\_Rel\_Personal\_Applied)  
  
  
exc1 <- exc1 %>% select(participant = Media\_Title,   
 exc\_range = Excerpt\_Range,  
 exc\_creator = Excerpt\_Creator,  
 exc\_date = Excerpt\_Date,  
 exc\_raw = Excerpt\_Copy,  
 codes = Codes\_Applied\_Combined,  
 justicia = Code\_Attitudes\_toward\_justicia\_por\_propia\_mano\_Applied,  
 state = Code\_Attitudes\_toward\_state\_Applied,  
 crimeID = Code\_Crime\_CrimeID\_Applied,  
 crimeID\_num = Code\_Crime\_CrimeID\_Weight,  
 speaker = Code\_Crime\_Speakers\_relationship\_to\_crime\_Applied,  
 personal = Code\_Crime\_Speakers\_relationship\_to\_crime\_Rel\_Personal\_Applied,  
 heard = Code\_Crime\_Speakers\_relationship\_to\_crime\_Rel\_Heard\_Witnessed\_Applied,  
 hypothetical = Code\_Crime\_Speakers\_relationship\_to\_crime\_Rel\_Hypothetical\_General\_Applied,  
 crime = Code\_Crime\_Type\_of\_crime\_Applied,  
 assault = Code\_Crime\_Type\_of\_crime\_Cri\_Assault\_con\_violencia\_Applied,  
 dom\_vio = Code\_Crime\_Type\_of\_crime\_Cri\_Domestic\_violence\_Applied,  
 drug = Code\_Crime\_Type\_of\_crime\_Cri\_Drug\_trafficking\_dealing\_Applied,  
 extort = Code\_Crime\_Type\_of\_crime\_Cri\_Extortion\_Applied,  
 kidnap = Code\_Crime\_Type\_of\_crime\_Cri\_Kidnapping\_Applied,  
 murder = Code\_Crime\_Type\_of\_crime\_Cri\_Murder\_Applied,  
 rob = Code\_Crime\_Type\_of\_crime\_Cri\_Robbery\_Property\_Loss\_Applied,  
 sexual = Code\_Crime\_Type\_of\_crime\_Cri\_Sexual\_Applied,   
 verbal = Code\_Crime\_Type\_of\_crime\_Cri\_Verbal\_threat\_and\_intimidation\_Applied,   
 measure = Code\_Legal\_Extrajudicial\_measures\_Applied,  
 commu\_illegal = Code\_Legal\_Extrajudicial\_measures\_LE\_Community\_extrajudicial\_Applied,   
 commu\_legal = Code\_Legal\_Extrajudicial\_measures\_LE\_Community\_legal\_Applied,   
 state\_illegal = Code\_Legal\_Extrajudicial\_measures\_LE\_State\_extrajudicial\_Applied,   
 state\_legal = Code\_Legal\_Extrajudicial\_measures\_LE\_State\_legal\_Applied,   
 moral = Code\_Moral\_reasoning\_Applied,   
 conseq = Code\_Moral\_reasoning\_MR\_Consequentialist\_Applied,   
 dehuman = Code\_Moral\_reasoning\_MR\_Dehumanization\_Applied,   
 deonto = Code\_Moral\_reasoning\_MR\_Deontological\_Applied,   
 human = Code\_Moral\_reasoning\_MR\_Empathy\_humanization\_Applied,   
 response = Code\_Type\_of\_Response\_Applied,   
 arrest = Code\_Type\_of\_Response\_Res\_Arrest\_Applied,   
 death = Code\_Type\_of\_Response\_Res\_Capital\_punishment\_death\_Applied,   
 commu\_serv = Code\_Type\_of\_Response\_Res\_Community\_service\_Applied,   
 life = Code\_Type\_of\_Response\_Res\_Life\_sentence\_Applied,   
 missing = Code\_Res\_Missing\_Applied,   
 no\_resp = Code\_Type\_of\_Response\_Res\_No\_response\_Applied,   
 physical = Code\_Type\_of\_Response\_Res\_Physical\_punishment\_torture\_Applied,   
 prevention = Code\_Type\_of\_Response\_Res\_Prevention\_Applied,   
 reparations = Code\_Type\_of\_Response\_Res\_Reparations\_compensations\_Applied,   
 reintegration = Code\_Type\_of\_Response\_Res\_Social\_reintegration\_Applied,   
 jail = Code\_Type\_of\_Response\_Res\_Some\_jail\_time\_Applied)

1. Replacing string variables (False - True) with dummy variables (0 - 1)

exc2 <- exc1  
  
exc2[exc2=="False"] <- '0'  
exc2[exc2=="True"] <- '1'  
exc2[7:46] <- sapply(exc2[7:46], as.numeric)

1. Rectifying dummy variables ‘speaker’, ‘crime’, ‘measures’, ‘moral’, and ‘response’

exc3 <- exc2  
  
crimes <- c('assault', 'dom\_vio', 'drug', 'extort', 'kidnap', 'murder', 'rob', 'sexual', 'verbal')  
measures <- c('commu\_illegal', 'commu\_legal', 'state\_illegal', 'state\_legal')  
morals <- c('conseq', 'dehuman', 'deonto', 'human')  
responses <- c("arrest", "death", "commu\_serv", "life", "missing", "no\_resp", "physical", "prevention", "reparations", "reintegration", "jail")  
  
exc3$speaker <- ifelse(exc3$personal > 0 | exc3$heard > 0 | exc3$hypothetical > 0, 1, 0)  
exc3$crime <- ifelse(rowSums(exc3[crimes]) > 0, 1, 0)  
exc3$measure <- ifelse(rowSums(exc3[measures]) > 0, 1, 0)  
exc3$moral <- ifelse(rowSums(exc3[morals]) > 0, 1, 0)  
exc3$response <- ifelse(rowSums(exc3[responses]) > 0, 1, 0)

1. Create new variables:
2. ‘exc\_start’, character position where excerpt starts
3. ‘exc\_end’, character position where excerpt ends

exc4 <- exc3 %>% mutate(exc\_start = as.numeric(sub("\\-.\*", "", exc\_range)),  
 exc\_end = as.numeric(sub(".\*\\-", "", exc\_range)), .after = exc\_range)

1. Creating new variables that counts the different Types of Crime (crime\_count), Measures (measure\_count), Moral Reasonings (moral\_count), and Types of Responses (response\_count) by excerpt or row.

exc5 <- exc4 %>% mutate(crime\_count = rowSums(exc4[crimes]), .before = measure)  
exc5 <- exc5 %>% mutate(measure\_count = rowSums(exc5[measures]), .before = moral)  
exc5 <- exc5 %>% mutate(moral\_count = rowSums(exc5[morals]), .before = response)  
exc5 <- exc5 %>% mutate(response\_count = rowSums(exc5[responses])) # 2,896 obs of 54 var

1. Creating new three-level variable ‘rel\_crime’ derived from variables ‘personal’, ‘heard’, and ‘hypothetical’. This variable describes the Speaker’s Relationship to the Crime.

exc6 <- exc5 %>% add\_column(rel\_crime = NA, .after = 'speaker')  
exc6$rel\_crime <- ifelse(exc6$personal == 1, 'Personal',  
 ifelse(exc6$heard == 1, 'Heard',   
 ifelse(exc6$hypothetical == 1, 'Hypothetical', exc6$rel\_crime)))#2,896 obs of 55 var

1. Writing clean dataset

write\_xlsx(exc6, '../02\_Data/02\_Clean/excerpts\_clean\_set2.xlsx')