

Filter Artists

Alessandro Fook

04 de outubro de 2018

From a large frame filter and select a few number of artists.

```
#Manipulate dataframe to remove unnecessary information
dataframe <- read.csv("spotigraph-database.csv")
keeps <- c("artist_name", "playlist_name")
dataframe2 <- dataframe[keeps] %>% unique
dataframe3 <- dataframe2[!grepl(",|&", dataframe2$artist_name),]
dataframe4 <- dataframe3[grepl("^[A-Za-z]+$", dataframe3$artist_name),]

#Create a dataframe to use as filter
frame_filter1 <- as.data.frame(table(unlist(dataframe4$playlist_name)))
frame_filter2 <- frame_filter1[!grepl("Playlist|,|\\|\\|", frame_filter1$Var1),]
frame_filter3 <- frame_filter2[grepl("^[A-Za-z]+$", frame_filter2$Var1),]
frame_filter4 <- subset(frame_filter3, subset = (Freq > 10 & Freq < 20))
frame_filter5 <- frame_filter4[sample(nrow(frame_filter4), 30),]

#Filter dataframe and save to a csv file
subset1 <- subset(dataframe4, playlist_name %in% frame_filter5$Var1)
subset2 <- aggregate(subset1$playlist_name,subset1['artist_name'],paste,collapse=',')
colnames(subset2) <- c("artist", "playlists")

#Specifying the url for desired website to be scrapped
url_base <- 'https://www.vagalume.com.br/'
url_rest <- '/relacionados/'

#Scraping related artists
subset2["related"] <- NA

for (i in 1:dim(subset2)[1]) {
  relacionados <- tryCatch({html_nodes(read_html(paste(url_base,gsub(" ", "-"),tolower(as.character(subset2$related[i]))))
  subset2$related[i] <- paste(unlist(list(relacionados)), collapse = ",")
}

write.csv(subset2, file = "filter-artists.csv", row.names = FALSE)
```