CS102_IT2C_JUNTANILLA_LAB5

Jose Roland Juntanilla

2024-03-25

#LAB 4 DATASET CLEANING(ARXIV)

```r
library(readr)
library(stringr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
# Reading dataset CSV LAB 4
cleandata <- read_csv("extractarxiv.csv")
```

```
## New names:
## * `` -> `...1`

## Rows: 110 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (5): Title, Author, Subject, Abstract, Meta
## dbl (1): ...1
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
#Extracting the date from the meta column
arxiv_clean_date <- str_extract(cleandata$Meta, "\\d+\\s[A-Za-z]+\\s\\d+")

#Datatype Changing
arxiv_date_datatype <- as.Date(arxiv_clean_date, format = "%d %b %Y")
head(arxiv_date_datatype)
```

```
## [1] "2024-03-10" "2024-03-09" "2024-03-08" "2024-03-07" "2024-02-23"
## [6] "2024-02-29"
```

```r
#Combining Data and Cleaning Dataset
cleanarxiv <- cleandata %>%
  mutate(date = arxiv_date_datatype,
         Subject = gsub("\\s\\(.*\\)", "", Subject),
         across(where(is.character), tolower)) %>%
  select(-Meta, -...1)

#Writing Final Arxiv CSV
write.csv(cleanarxiv,file = "cleanextractarxiv.csv")
```

#LAB 5 DATASET CLEANING(PRODUCT REVIEWS)

```
library(readr)
library(stringr)
library(dplyr)

#Reading Dataset CSV
scrapedreviews <- read_csv("finalreviews.csv")
```

```
## New names:
## Rows: 2500 Columns: 7
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (6): Name, Rate, Type_of_Review, Reviewers_name, Reviews, Data_of_Reviews dbl
## (1): ...1
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```
#Getting the date from the meta column and changing to date type
reviews_clean_date <- as.Date(str_extract(scrapedreviews$Data_of_Reviews, "\\d+\\s[A-Za-z]+\\s\\d+"), fo

#Changing and Cleaning Ratings
reviews_clean_rating <- as.integer(str_extract(scrapedreviews$Rate, "\\d+\\.\\d+"))

# Cleaning and Removing Emoticons in the Dataset
scrapedreviews$Name <- gsub("\\p{So}", "", scrapedreviews$Name, perl = TRUE)
scrapedreviews$Reviewers_name <- gsub("\\p{So}", "", scrapedreviews$Reviewers_name, perl = TRUE)
scrapedreviews$Reviews <- gsub("\\p{So}", "", scrapedreviews$Reviews, perl = TRUE)

#Cleaning diiferent languages letter and non alphabetical letters
scrapedreviews$Name <- gsub("[^a-zA-Z ]", "", scrapedreviews$Name)
scrapedreviews$Reviewers_name <- gsub("[^a-zA-Z ]", "", scrapedreviews$Reviewers_name)
scrapedreviews$Reviews  <- gsub("[^a-zA-Z ]", "",scrapedreviews$Reviews )

#Replacing White Spaces with NA
scrapedreviews$Name <- na_if(scrapedreviews$Name, "")
scrapedreviews$Reviewers_name <- na_if(scrapedreviews$Reviewers_name, "")
scrapedreviews$Reviews <- na_if(scrapedreviews$Reviews, "")

#Changing all columns to lowercase
scrapedreviews <- scrapedreviews %>%
  mutate(across(where(is.character), tolower)) %>%
  select(-...1)

#Combine all columns
final_cleaned_reviews <- scrapedreviews %>%
  mutate(Data_of_Reviews = reviews_clean_date,
         Rate=reviews_clean_rating )
```

```r
#Writing Final Product Reviews CSV
write.csv(final_cleaned_reviews, "cleanefinalreviews.csv")
```