LAB-EXERCISE#5_ECHAVERIA

Luigi Echaveria

2024-03-18

// Lab Exercise 4 cleaning

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

# Load Arxiv Scraped Dataset
arxiv <- read_csv("Arxiv_papers_on_Data_Scrape.csv")

# Extracting the date from the meta column
arxiv_date_only <- str_extract(arxiv$meta, "\\d+\\s[A-Za-z]+\\s\\d+")

# Changing to date type
arxiv_date_type <- as.Date(arxiv_date_only, format = "%d %b %Y")
head(arxiv_date_type)

# Removing meta and number column and appending the new date column
# Mutating all while converting other columns to lowercase, removing parenthesis text in the subject col
cleaned_arxiv <- arxiv %>%
  mutate(date = arxiv_date_type,
         subject = gsub("\\s\\(.*\\)", "", subject),
         across(where(is.character), tolower)) %>%
  select(-meta, -...1)

# Writing to CSV
write.csv(cleaned_arxiv, "cleaned_arxiv.csv")
```

// Lab Exercise 5 Cleaning

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

# Load Arxiv Scraped Dataset
movie_reviews <- read_csv("All Reviews.csv")

# Extracting the date from the meta column and changing to date type
reviews_date_type <- as.Date(str_extract(movie_reviews$date, "\\d+\\s[A-Za-z]+\\s\\d+"), format = "%d %b

# Extracting the rating from the rating column and changing to integer
reviews_ratings_integer <- as.integer(str_extract(movie_reviews$ratings, "\\d+\\.\\d+"))
```

```
# Removing all emoticons from the columns
movie_reviews$title <- gsub("\\p{So}", "", movie_reviews$title, perl = TRUE)
movie_reviews$reviewer <- gsub("\\p{So}", "", movie_reviews$reviewer, perl = TRUE)
movie_reviews$review <- gsub("\\p{So}", "", movie_reviews$review, perl = TRUE)

# Removing non-alphabetical languages from the columns
movie_reviews$title <- gsub("[^a-zA-Z ]", "", movie_reviews$title)
movie_reviews$reviewer <- gsub("[^a-zA-Z ]", "", movie_reviews$reviewer)
movie_reviews$review <- gsub("[^a-zA-Z ]", "", movie_reviews$review)

# Replace all blank string with NA
movie_reviews$title <- na_if(movie_reviews$title, "")
movie_reviews$reviewer <- na_if(movie_reviews$reviewer, "")
movie_reviews$review <- na_if(movie_reviews$review, "")

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

# Combine all together
cleaned_reviews <- movie_reviews %>%
  mutate(date = reviews_date_type, ratings = reviews_ratings_integer)

# Writing to CSV
write.csv(cleaned_reviews, "cleaned_reviews.csv")
```