lab_exercise#5

Mel Adry Olivo

2024-03-14

Lab Exercise 4 Cleaning

```
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
```

```
# Load Arxiv Scraped Dataset
arxiv <- read_csv("dataset/arxiv_ai.csv")
```

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

```
## Rows: 150 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.
```

```
# 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)
```

```
## [1] "2024-03-05" "2024-03-05" "2024-03-04" "2024-03-01" "2024-02-19"
## [6] "2024-03-01"
```

```
# 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 co
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_datasets/cleaned_arxiv.csv")
```

Lab Exercise 5 Cleaning

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

# Load Arxiv Scraped Dataset
products_reviews <- read_csv("dataset/allProducts.csv")
```

```
## New names:
## Rows: 2500 Columns: 8
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (7): prod_name, title, reviewer, review, date, ratings, type_of_purchase 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`
```

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

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

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

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

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

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

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

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