Lab 5

Charles Daniel Songaling

2024-03-30

```
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
```

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

#Amazon_Reviews

```
ama_revs <- read.csv("50AmazonProducts_Reviews.csv")

ama_reviews <- ama_revs

# Removing emojis
ama_reviews$Review <- gsub("\\p{So}", "", ama_reviews$Review, perl = TRUE)

# Removing non-English language
ama_reviews$Review <- gsub("[^a-zA-Z ]", "",ama_reviews$Review)

amazon_reviews <- ama_reviews %>%
  mutate(across(where(is.character), tolower)) %>%
  select(-Review)

# Adding the review date
review_date <- ama_revs$DateOfReview

ama_reviews$DateOfReview <- review_date

review_date <- as.Date(str_extract(ama_reviews$date, "\\d+\\s[A-Za-z]+\\s\\d+"), format = "%d %b %Y")

# Filtering the ratings
review_rating<- as.integer(str_extract(ama_reviews$Rating, "\\d+\\.\\d+"))


filtered_amazonreviews <- amazon_reviews %>%
  mutate(DateOfReview = DateOfReview, Rating = review_rating)
```

```r
#Writing cleaned data to CSV file
write.csv(filtered_amazonreviews,file = "AmazonReviews_filtered&cleaned.csv")
```

#Arxiv

```r
arxiv_paper <- read_csv("Arxiv papers on Sports.csv")
```

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

```r
date_extract <- str_extract(arxiv_paper$meta, "\\d+\\s[A-Za-z]+\\s\\d+")

datex <- as.Date(date_extract, format = "%d %b %Y")
head(datex)
```

```
## [1] "2024-03-13" "2024-03-11" "2024-03-07" "2024-03-06" "2024-03-06"
## [6] "2024-03-03"
```

```r
filtered_Sportsarxivpapers <- arxiv_paper %>%
  mutate(date = datex) %>%
  mutate(subject = gsub("\\s\\(.*\\)", "", subject),
         across(where(is.character), tolower)) %>%
         select(-meta, -...1)

#Writing cleaned data to CSV file
write.csv(filtered_Sportsarxivpapers,file = "Sportsarxivpapers_filtered&cleaned.csv")
```