Lab Exercise 5

Ace Bagilidad

2024-03-14

#Cleaning arxiv articles-Machine Learning

```
library(readr)
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(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

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

```
Machine_Learning <- read_csv("Data Sets/Machine Learning.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.
```

```
#cleaning of dates
cleaned_date <- Machine_Learning$meta %>% str_extract("\\w+, \\d+ \\w+ \\d{4}") %>%
  as.POSIXct(format = " %a, %d %b %Y", tz = "UTC") %>%
  as.Date() %>%
 format("%Y-%m-%d")

#mutating cleaned date,cleaning of subjects and converting dataset to lowercase.
```

```
clMachine_Learning <- Machine_Learning %>% mutate(meta=cleaned_date,
                                                  subject= gsub("\\s\\(.*\\)", "", subject)) %>% mu
head(clMachine_Learning)
```

```
## # A tibble: 6 x 6
##   ...1  title                                    author subject abstract meta
##   <chr> <chr>                                    <chr>  <chr>   <chr>    <chr>
## 1 1     telemoma: a modular and versatile teleope~ shivi~ roboti~ "a crit~ 2024~
## 2 2     exploring safety generalization challenge~ qibin~ comput~ "the ra~ 2024~
## 3 3     low coordinate degree algorithms i: unive~ dmitr~ statis~ "we stu~ 2024~
## 4 4     fairness feedback loops: training on synt~ sierr~ machin~ "model-~ 2024~
## 5 5     quantum support vector machine for prosta~ walid~ machin~ "this s~ 2024~
## 6 6     distilling the knowledge in data pruning   emanu~ comput~ "with t~ 2024~
```

```
#writing to csv
#write.csv(clMachine_Learning,"Cleaned Data Set-Machine Learning")
```

#Cleaning of amazon product reviews

```
library(stringi)

FiftyProductsReview <- read_csv("Data Sets/FiftyProductsReview.csv")
```

```
## New names:
## Rows: 2500 Columns: 8
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (7): Category, ProductName, Reviewer, Review, TypeOfReviewer, ReviewDate... 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`
```

```
#convert dataset to lower
tl_FiftyProductsReview <- FiftyProductsReview %>% mutate(across(where(is.character), tolower))

#cleaning of dates
cleaned_date <- tl_FiftyProductsReview$ReviewDate %>%
  str_extract("\\d{1,2} \\w+ \\d{4}") %>%
  as.Date(format = " %d %B %Y") %>%
  format("%Y-%m-%d")

# Extracting the rating from the rating column and changing to integer
products_reviews_integer <-  as.integer(str_extract(tl_FiftyProductsReview$Rating, "\\d+\\.\\d+"))




# Converting dataset into lowercase and cleaning emoji,non char, and newlines
cleaned_FiftyProductsReview <- tl_FiftyProductsReview %>% mutate(ReviewDate=cleaned_date,
                                                  Rating=products_reviews_integer,
                                                  Review = gsub("\\n", "", Review),
                                                  Review = stri_replace_all_regex(Review, pattern =
                                                  Review = gsub("[^[:graph:][:space:]]", "", Revie
                                                  )
head(cleaned_FiftyProductsReview)
```

```
## # A tibble: 6 x 8
##    ...1 Category    ProductName Reviewer Review TypeOfReviewer ReviewDate Rating
##   <dbl> <chr>       <chr>       <chr>    <chr>  <chr>          <chr>       <int>
## 1     1 mens jacke~ columbia    customer "    ~ verified purc~ 2023-12-06      5
## 2     2 mens jacke~ columbia    rt       "    ~ verified purc~ 2023-11-14      5
## 3     3 mens jacke~ columbia    crystal  "    ~ verified purc~ 2023-07-28      5
## 4     4 mens jacke~ columbia    abc 99   "    ~ verified purc~ 2021-07-15      4
## 5     5 mens jacke~ columbia    james i~ "    ~ verified purc~ 2021-07-13      5
## 6     6 mens jacke~ columbia    jeffrey~ "    ~ verified purc~ 2021-06-28      5
```

```
#view(cleaned_FiftyProductsReview)
write.csv(cleaned_FiftyProductsReview,"Cleaned Data Sets/Cleaned Data Set - FiftyProductsReviews.csv")
```