CS102_5th_Act

2024-03-24

```
knitr::opts_chunk$set(echo = TRUE)
```

```
knitr::opts_chunk$set(echo = TRUE)
```

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

```
droga <- read_csv("drugs.csv")
view(droga)
```

```
extdate <- str_extract(droga$meta, "\\d+\\s[A-Za-z]+\\s\\d+")

extdatetype <- as.Date(extdate, format = "%d %b %Y")
head(extdatetype)

cleaned_droga <- droga %>%
  mutate(date = extdatetype) %>%
  mutate(subject = gsub("\\s\\(.*\\)", "", subject),
         across(where(is.character), tolower)) %>%
         select(-meta, -...1)

write.csv(cleaned_droga, "cleanedfiles/cleaned_arxivDrugs_Papers.csv")
```

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

```
prod_rev <- read_csv("ReviewedAmazonProducts.csv")
```

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

```
revratings_int<- as.integer(str_extract(prod_rev$ratings, "\\d+\\.\\d+"))
```

```
    prod_rev$title <- gsub("\\p{So}", "", prod_rev$title, perl = TRUE)
    prod_rev$reviewer <- gsub("\\p{So}", "", prod_rev$reviewer, perl = TRUE)
    prod_rev$review <- gsub("\\p{So}", "", prod_rev$review, perl = TRUE)


    prod_rev$title <- gsub("[^a-zA-Z ]", "", prod_rev$title)
    prod_rev$reviewer <- gsub("[^a-zA-Z ]", "", prod_rev$reviewer)
    prod_rev$review <- gsub("[^a-zA-Z ]", "", prod_rev$review)


    prod_rev$title <- na_if(prod_rev$title, "")
    prod_rev$reviewer <- na_if(prod_rev$reviewer, "")
    prod_rev$review <- na_if(prod_rev$review, "")

prods_revs <- prod_rev %>%
  mutate(across(where(is.character), tolower)) %>%
  select(-...1)

cleaned_prodrev <- prods_revs %>%
  mutate(date = revdatetype, ratings = revratings_int)



write.csv(cleaned_prodrev, "cleanedfiles/cleaned_ReviewedAmazonProducts.csv")
```