WorkSheet #5

Malayas, Pauchano, Madayag BSIT2A

2024-11-09

```
# libraries
library(polite)
library(httr)
library(rvest)
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
```

```
# Set URL and establish session
imdb_url <- "https://www.imdb.com/chart/toptv/?sort=rank%2Casc"
imdb_session <- bow(imdb_url, user_agent = "Educational")
imdb_session
```

```
## <polite session> https://www.imdb.com/chart/toptv/?sort=rank%2Casc
##     User-agent: Educational
##     robots.txt: 35 rules are defined for 3 bots
##    Crawl delay: 5 sec
##   The path is scrapable for this user-agent
```

```
# Extract TV show titles and ranks
tv_titles <- read_html(imdb_url) %>%
  html_nodes('.ipc-title__text') %>%
  html_text()
```

```
# transform extracted titles
tv_titles_df <- as.data.frame(tv_titles[3:27], stringsAsFactors = FALSE)
colnames(tv_titles_df) <- "ranked_titles"
```

```r
# Rename and delete columns
split_rank_title <- strsplit(as.character(tv_titles_df$ranked_titles), "\\.", fixed = FALSE)
split_rank_title_df <- data.frame(do.call(rbind, split_rank_title), stringsAsFactors = FALSE)
colnames(split_rank_title_df) <- c("Rank", "Title")
split_rank_title_df$Title <- trimws(split_rank_title_df$Title)

ranked_titles_df <- split_rank_title_df
```

```r
# Extract ratings, number of votes, episodes, and release years
tv_ratings <- read_html(imdb_url) %>%
  html_nodes('.ipc-rating-star--rating') %>%
  html_text()
```

```r
tv_votes <- read_html(imdb_url) %>%
  html_nodes('.ipc-rating-star--voteCount') %>%
  html_text()
cleaned_votes <- gsub('[()]', '', tv_votes)
```

```r
# Extract episode counts
episode_counts <- read_html(imdb_url) %>%
  html_nodes('span.sc-5bc66c50-6.OOdsw.cli-title-metadata-item:nth-of-type(2)') %>%
  html_text()
cleaned_episodes <- gsub('[eps]', '', episode_counts)
episode_counts_num <- as.numeric(cleaned_episodes)
```

```r
# Extract release years
release_years <- read_html(imdb_url) %>%
  html_nodes('span.sc-5bc66c50-6.OOdsw.cli-title-metadata-item:nth-of-type(1)') %>%
  html_text()
```