ACTIVITY2_SADSAD

BSIT-2B

2024-02-06

```
library(rvest)
```

```
## Warning: package 'rvest' was built under R version 4.3.2
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.3.2
```

```
##
## 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(writexl)
```

```
## Warning: package 'writexl' was built under R version 4.3.2
```

```
library(polite)
```

```
## Warning: package 'polite' was built under R version 4.3.2
```

```
session <- bow(url = 'https://www.imdb.com/title/tt5776858/reviews?ref_=tt_urv',
               user_agent = "Student's Demo Educational")
session
```

```
## <polite session> https://www.imdb.com/title/tt5776858/reviews?ref_=tt_urv
##     User-agent: Student's Demo Educational
##     robots.txt: 35 rules are defined for 3 bots
##    Crawl delay: 5 sec
##   The path is scrapable for this user-agent
```

```r
session_scrape <- scrape(session)

scrape_reviews <- function(page_url) {
  page <- read_html(page_url)

  names <- page %>% html_nodes(".display-name-link") %>% html_text()
  dates <- page %>% html_nodes("span.review-date") %>% html_text()
  ratings <- page %>% html_nodes("span.rating-other-user-rating") %>% html_text()
  content_reviews <- page %>% html_nodes("a.title") %>% html_text()
  reviews <- page %>% html_nodes(".text.show-more__control") %>% html_text()

  phantom_thread_df = data.frame(
    Name = names[1:25],
    Date = dates[1:25],
    User_Rating = ratings[1:25],
    Content_Review = content_reviews[1:25],
    Reviews = reviews[1:25]
  )
}

phantom_thread_urls <- c(
  'https://www.imdb.com/title/tt5776858/reviews?ref_=tt_urv',
  'https://www.imdb.com/title/tt5776858/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjnbzr3smmarz3u
  'https://www.imdb.com/title/tt5776858/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjnbqr3t4ycjt3q
  'https://www.imdb.com/title/tt5776858/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjnjzqpsmscb23u
  'https://www.imdb.com/title/tt5776858/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjnjvrtrmmbr43q
  'https://www.imdb.com/title/tt5776858/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjnjtrdumsab53q
  'https://www.imdb.com/title/tt5776858/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjmryqlrmscj52m
  'https://www.imdb.com/title/tt5776858/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjmrwqlrmsaj43a
  'https://www.imdb.com/title/tt5776858/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjmruqpsmoaby2i
  'https://www.imdb.com/title/tt5776858/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjmrvrdu44az73u
  'https://www.imdb.com/title/tt5776858/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjmrtrdumsab53q
  'https://www.imdb.com/title/tt5776858/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjmrrqlqm6cby3i
)


all_reviews <- lapply(phantom_thread_urls, scrape_reviews)

final_all_reviews <- do.call(rbind, all_reviews)

View(final_all_reviews)
```