***IMDb Top 250 Web Scraper***

This R script scrapes details about movies listed in IMDb's Top 250 chart. The list can be found in <https://www.imdb.com/chart/top/?ref_=tt_awd>.

The Movie Attributes included in the data set are the following:

* Title.
* Year.
* MPAA rating: this is the Motion Picture Association film rating system. The complete guide for what each number or letter means can be found in <https://www.filmratings.com/>. However, variations are found for certain movies not completely supervised by the Motion Picture Association.
* Duration of the film.
* IMDB score.
* Director.
* Main actor.
* Supporting Actor.
* Main writer.
* Second writer.
* Release date.
* Genre.
* Budget: the estimated total used in the creation of the movie. The currency is generally in dollars, however there are some exceptions for the movies that have no relation to the US movie market.
* Gross Worldwide Revenue: the estimated total revenue attained for each movie in dollars.
* Title of featured review: main title of the review featured in the IMDB page at the time of the scraping.
* Text of the featured review: complete text of the featured review found in the IMDB page.
* Awards and nominations: number of awards and nominations given to the movie at the time of the scraping.
* Oscar information: number of Academy Award wins and nominations.

**Prerequisites**

Before running this script, install the following R packages:

* **rvest**: For web scraping
* **httr**: For handling HTTP requests
* **dplyr**: For data manipulation

**Running the Scraper**

1. Open the R script called “**Markdown For IMDB web scraping.Rmd”** in RStudio
2. Run the entire script to scrape data from the IMDb Top 250 chart.
3. The script will create two data frames: **movies\_df**, which contains details from a test sample of 5 movie URLs, and **movies\_df\_final**, which contains details from all movie URLs in the Top 250 chart. The first data frame was created to show a preview of how the data is being collected and not having to wait for an extended period in order to get the full data set.
4. The resulting data frames are printed to the console.

**Notes**

* The scraper includes a user agent string within the set\_config() function to simulate a real browser session. The user agent string is specified as

**"Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/121.0.0.0 Safari/537.36; Sofia Villamil /** [**sofia.v1999@gmail.com**](mailto:sofia.v1999@gmail.com)**"**

in the script. Replace "Sofia Villamil / sofia.v1999@gmail.com" with your actual user agent details if needed.

* The scraper is designed to be polite to the server by including a delay between requests. This is controlled by the **Sys.sleep(runif(1, 1, 3))** function call, which sleeps for a random duration between 1 to 3 seconds between each request.
* The script includes error handling to skip any URLs that might cause an error during the scraping process.
* The structure of web pages can change over time. If IMDb updates the structure of its pages, the scraper may need to be updated as well.