

# Movie Reviews

Jasper Conlu

2024-02-27

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
library(rvest)
library(httr)
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(polite)
library(stringr)
library(ggplot2)
library(readr)
```

```
##
## Attaching package: 'readr'

## The following object is masked from 'package:rvest':
##
##   guess_encoding
```

```
polite::use_manners(save_as = 'polite_scrape.R')
```

```
## v Setting active project to '/cloud/project'
```

```
session <- bow(url = 'https://www.imdb.com/title/tt4779682/reviews?ref_=tt_urv' , user_agent = "Educational")
session
```

```
## <polite session> https://www.imdb.com/title/tt4779682/reviews?ref_=tt_urv
##   User-agent: Educational
##   robots.txt: 35 rules are defined for 3 bots
##   Crawl delay: 5 sec
##   The path is scrapable for this user-agent
```



```

session_scrape <- scrape(session)

Saltburn_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()

  sb_df= data.frame(
    Movie_Name=c('Saltburn'),
    Reviewer_Name = names[1:25],
    Content_Review = content_reviews[1:25],
    Date = dates[1:25],
    User_Rating = ratings[1:25],
    Reviews = reviews[1:25]
  )
}

urls<-c('https://www.imdb.com/title/tt17351924/reviews?ref_=tt_urv',
'https://www.imdb.com/title/tt17351924/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjnrvrqx',
'https://www.imdb.com/title/tt17351924/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjnztrdr',
'https://www.imdb.com/title/tt17351924/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjnbwrlq',
'https://www.imdb.com/title/tt17351924/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjnbqrhq',
'https://www.imdb.com/title/tt17351924/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjnjyqps',
'https://www.imdb.com/title/tt17351924/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjnjzqlu',
'https://www.imdb.com/title/tt17351924/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjnjwrlu',
'https://www.imdb.com/title/tt17351924/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjnjvrds',
'https://www.imdb.com/title/tt17351924/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjnjqqps',
'https://www.imdb.com/title/tt17351924/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjmrzrhq',
'https://www.imdb.com/title/tt17351924/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjmrzrhq'
)

S_reviews <- lapply(urls, Saltburn_reviews)
Saltburn_Review <- do.call(rbind, S_reviews)

#3rd Movie

session <- bow(url = 'https://www.imdb.com/title/tt9362722/reviews/?ref_=tt_ov_rt', user_agent = "Educational")

session

## <polite session> https://www.imdb.com/title/tt9362722/reviews/?ref_=tt_ov_rt
## User-agent: Educational
## robots.txt: 35 rules are defined for 3 bots
## Crawl delay: 5 sec
## The path is scrapable for this user-agent

session_scrape <- scrape(session)

```

```

Spidey_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()

  sp_df= data.frame(
    Movie_Name=c('Spiderman'),
    Reviewer_Name = names[1:25],
    Content_Review = content_reviews[1:25],
    Date = dates[1:25],
    User_Rating = ratings[1:25],
    Reviews = reviews[1:25]
  )
}

spidey_urls<-c('https://www.imdb.com/title/tt9362722/reviews/?ref_=tt_ov_rt',
  'https://www.imdb.com/title/tt17351924/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxw',
  'https://www.imdb.com/title/tt17351924/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxw',
  'https://www.imdb.com/title/tt17351924/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxw',
  'https://www.imdb.com/title/tt17351924/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxw',
  'https://www.imdb.com/title/tt17351924/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxw',
  'https://www.imdb.com/title/tt17351924/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxw',
  'https://www.imdb.com/title/tt17351924/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxw',
  'https://www.imdb.com/title/tt17351924/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxw',
  'https://www.imdb.com/title/tt17351924/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxw'
)

sp_reviews <- lapply(spidey_urls, Spidey_reviews)
Spiderman_Review <- do.call(rbind, sp_reviews)

#4th Movie

session <- bow(url = 'https://www.imdb.com/title/tt9362722/reviews/?ref_=tt_ov_rt', user_agent = "Educational")
session

## <polite session> https://www.imdb.com/title/tt9362722/reviews/?ref_=tt_ov_rt
## User-agent: Educational
## robots.txt: 35 rules are defined for 3 bots
## Crawl delay: 5 sec
## The path is scrapable for this user-agent

session_scrape <- scrape(session)

PussB_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()

sp_df= data.frame(
  Movie_Name=c('Puss N Boots'),
  Reviewer_Name = names[1:25],
  Content_Review = content_reviews[1:25],
  Date = dates[1:25],
  User_Rating = ratings[1:25],
  Reviews = reviews[1:25]
)
}

puss_urls<-c('https://www.imdb.com/title/tt3915174/reviews?ref_=tt_urv',
  'https://www.imdb.com/title/tt3915174/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjnb',
  'https://www.imdb.com/title/tt3915174/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjnb',
  'https://www.imdb.com/title/tt3915174/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjnb',
  'https://www.imdb.com/title/tt3915174/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjnb',
  'https://www.imdb.com/title/tt3915174/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjnb',
  'https://www.imdb.com/title/tt3915174/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjnb',
  'https://www.imdb.com/title/tt17351924/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjnb',
  'https://www.imdb.com/title/tt17351924/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjnb',
  'https://www.imdb.com/title/tt17351924/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjnb',
  'https://www.imdb.com/title/tt17351924/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjnb',
  'https://www.imdb.com/title/tt17351924/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjnb'
)

pnb_reviews <- lapply(puss_urls, PussB_reviews)
Puss_N_Boots_Review <- do.call(rbind, pnb_reviews)

#5th Movie

session <- bow(url = 'https://www.imdb.com/title/tt1160419/reviews/?ref_=tt_ov_rt', user_agent = "Educational")

session

## <polite session> https://www.imdb.com/title/tt1160419/reviews/?ref_=tt_ov_rt
## User-agent: Educational
## robots.txt: 35 rules are defined for 3 bots
## Crawl delay: 5 sec
## The path is scrapable for this user-agent

session_scrape <- scrape(session)

Dune_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()

dune_df= data.frame(
  Movie_Name=c('Dune'),
  Reviewer_Name = names[1:25],
  Content_Review = content_reviews[1:25],
  Date = dates[1:25],
  User_Rating = ratings[1:25],
  Reviews = reviews[1:25]
)
}

dune_urls<-c('https://www.imdb.com/title/tt1160419/reviews/?ref_=tt_ov_rt',
  'https://www.imdb.com/title/tt3915174/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjnr',
  'https://www.imdb.com/title/tt3915174/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjnr',
  'https://www.imdb.com/title/tt3915174/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjnr',
  'https://www.imdb.com/title/tt3915174/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjnr',
  'https://www.imdb.com/title/tt3915174/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjnr',
  'https://www.imdb.com/title/tt17351924/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjnr',
  'https://www.imdb.com/title/tt17351924/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjnr',
  'https://www.imdb.com/title/tt17351924/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjnr',
  'https://www.imdb.com/title/tt17351924/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjnr'
)

dune_reviews <- lapply(dune_urls, Dune_reviews)
Dune_Review <- do.call(rbind, dune_reviews)

#6th Movie

session <- bow(url = 'https://www.imdb.com/title/tt0816692/reviews/?ref_=tt_ov_rt', user_agent = "Educational")

session

## <polite session> https://www.imdb.com/title/tt0816692/reviews/?ref_=tt_ov_rt
## User-agent: Educational
## robots.txt: 35 rules are defined for 3 bots
## Crawl delay: 5 sec
## The path is scrapable for this user-agent

session_scrape <- scrape(session)

Interstellar_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()

```

```

Inter_df= data.frame(
  Movie_Name=c('Interstellar'),
  Reviewer_Name = names[1:25],
  Content_Review = content_reviews[1:25],
  Date = dates[1:25],
  User_Rating = ratings[1:25],
  Reviews = reviews[1:25]
)
}

Inter_urls<-c('https://www.imdb.com/title/tt0816692/reviews/?ref_=tt_ov_rt',
  'https://www.imdb.com/title/tt3915174/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjn',
  'https://www.imdb.com/title/tt3915174/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjn',
  'https://www.imdb.com/title/tt3915174/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjn',
  'https://www.imdb.com/title/tt3915174/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjn',
  'https://www.imdb.com/title/tt3915174/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjn',
  'https://www.imdb.com/title/tt3915174/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjn',
  'https://www.imdb.com/title/tt17351924/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjn',
  'https://www.imdb.com/title/tt17351924/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjn',
  'https://www.imdb.com/title/tt17351924/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjn',
  'https://www.imdb.com/title/tt17351924/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjn',
  'https://www.imdb.com/title/tt17351924/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjn'
)

Int_reviews <- lapply(Inter_urls, Interstellar_reviews)
Interstellar_Review <- do.call(rbind, Int_reviews)

#7th Movie

session <- bow(url = 'https://www.imdb.com/title/tt0468569/reviews/?ref_=tt_ov_rt', user_agent = "Educational")

session

## <polite session> https://www.imdb.com/title/tt0468569/reviews/?ref_=tt_ov_rt
## User-agent: Educational
## robots.txt: 35 rules are defined for 3 bots
## Crawl delay: 5 sec
## The path is scrapable for this user-agent

session_scrape <- scrape(session)

Batman_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()

  Inter_df= data.frame(
    Movie_Name=c('The Dark Knight'),

```







```

    Date = dates[1:25],
    User_Rating = ratings[1:25],
    Reviews = reviews[1:25]
  )
}

GoT_urls<-c('https://www.imdb.com/title/tt0944947/reviews/?ref_=tt_ov_rt',
            'https://www.imdb.com/title/tt3915174/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjnrz',
            'https://www.imdb.com/title/tt3915174/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjnrz',
            'https://www.imdb.com/title/tt3915174/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjnrz',
            'https://www.imdb.com/title/tt3915174/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjnrz',
            'https://www.imdb.com/title/tt3915174/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjnrz',
            'https://www.imdb.com/title/tt17351924/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjnrz',
            'https://www.imdb.com/title/tt17351924/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjnrz',
            'https://www.imdb.com/title/tt17351924/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjnrz',
            'https://www.imdb.com/title/tt17351924/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwjnrz')

GOT_reviews <- lapply(GoT_urls, GoT_reviews)
Game_Of_Thrones_Review <- do.call(rbind, GOT_reviews)

#9th Movie

session <- bow(url = 'https://www.imdb.com/title/tt0133093/reviews/?ref_=tt_ov_rt', user_agent = "Educational")

session

## <polite session> https://www.imdb.com/title/tt0133093/reviews/?ref_=tt_ov_rt
##   User-agent: Educational
##   robots.txt: 35 rules are defined for 3 bots
##   Crawl delay: 5 sec
##   The path is scrapable for this user-agent

session_scrape <- scrape(session)

TheMAT_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()

  TheMat_df= data.frame(
    Movie_Name=c('The Matrix'),
    Reviewer_Name = names[1:25],
    Content_Review = content_reviews[1:25],
    Date = dates[1:25],
    User_Rating = ratings[1:25],
    Reviews = reviews[1:25]
  )
}

```

```

)
}

TheMat_urls<-c('https://www.imdb.com/title/tt0133093/reviews/?ref_=tt_ov_rt',
               'https://www.imdb.com/title/tt3915174/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwj',
               'https://www.imdb.com/title/tt3915174/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwj',
               'https://www.imdb.com/title/tt3915174/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwj',
               'https://www.imdb.com/title/tt3915174/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwj',
               'https://www.imdb.com/title/tt3915174/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxwj',
               'https://www.imdb.com/title/tt17351924/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxw',
               'https://www.imdb.com/title/tt17351924/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxw',
               'https://www.imdb.com/title/tt17351924/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxw',
               'https://www.imdb.com/title/tt17351924/reviews/_ajax?&paginationKey=g4w6ddbmqyzdo6ic4oxw')

)

Mat_reviews <- lapply(TheMat_urls, TheMAT_reviews)
The_Matrix_Review <- do.call(rbind, Mat_reviews)

#10th Movie

session <- bow(url = 'https://www.imdb.com/title/tt7286456/reviews/?ref_=tt_ov_rt', user_agent = "Educa

session

## <polite session> https://www.imdb.com/title/tt7286456/reviews/?ref_=tt_ov_rt
##   User-agent: Educational
##   robots.txt: 35 rules are defined for 3 bots
##   Crawl delay: 5 sec
##   The path is scrapable for this user-agent

session_scrape <- scrape(session)

J_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()

  J_df= data.frame(
    Movie_Name=c('Joker'),
    Reviewer_Name = names[1:25],
    Content_Review = content_reviews[1:25],
    Date = dates[1:25],
    User_Rating = ratings[1:25],
    Reviews = reviews[1:25]
  )
}

```

```

Joker_urls<-c('https://www.imdb.com/title/tt7286456/reviews/?ref_=tt_ov_rt',
              'https://www.imdb.com/title/tt3915174/reviews/_ajax?&paginationKey=g4xojermtizcsyab7gthhn',
              'https://www.imdb.com/title/tt3915174/reviews/_ajax?&paginationKey=g4u6dermtizcsyql7svxtnl',
              'https://www.imdb.com/title/tt3915174/reviews/_ajax?&paginationKey=g4v6jermtizcsyqe72th7n',
              'https://www.imdb.com/title/tt3915174/reviews/_ajax?&paginationKey=g4s6rermtizcsyqg7kxxhml',
              'https://www.imdb.com/title/tt3915174/reviews/_ajax?&paginationKey=g4w6hcbsqyxdo6ih7svxfm',
              'https://www.imdb.com/title/tt3915174/reviews/_ajax?&paginationKey=g4w6lbjsqyxdo6ih7suxvm',
              'https://www.imdb.com/title/tt17351924/reviews/_ajax?&paginationKey=g4w6ncbsqyxdo6ih7wth5',
              'https://www.imdb.com/title/tt17351924/reviews/_ajax?&paginationKey=g4w6rbjsqyxdo6ih7wuhv',
              'https://www.imdb.com/title/tt17351924/reviews/_ajax?&paginationKey=g4xohcbsqyxdo6ih7wvxh',
              'https://www.imdb.com/title/tt17351924/reviews/_ajax?&paginationKey=g4xolbjsqyxdo6ih7wwxh',
              'https://www.imdb.com/title/tt17351924/reviews/_ajax?&paginationKey=g4xoncbsqyxdo6ih7wxxx')

Joker_reviews <- lapply(Joker_urls, J_reviews)
Joker_Review <- do.call(rbind, Joker_reviews)

Movie_Reviews_3000 <- rbind ( TheMeg2_Review , Saltburn_Review , Spiderman_Review , Puss_N_Boots_Review

csv_file <- "Movie_Reviews_3000.csv"
write.csv(Movie_Reviews_3000, file = csv_file)
HouseholdData <- read.csv("Movie_Reviews_3000.csv")

```