Web Scraping Movie Genre Data

Jonathan Che

11 April 2017

Overview

I use information from the-numbers.com to extract, for the top 10 movies of each year, their genre and a text description of their rating.

Method

First, I do some variable setup.

```
url1 <- "http://www.the-numbers.com/market/"
url2 <- "/summary"
years <- seq(from=1998, to=2004)
master_list <- list()   # For the years
movie_info_list <- list()   # For the 10 movies in a year
#movie_info_df <- data.frame()
```

Next, I make a function that does the following: 1) Go to the "summary" page for a given year 2) Visit each of the top 10 movies for that year 3) Pull genre/MPAA Rating Description information for each of those movies

```
year <- 2000

extract_year <- function(year){
  url <- str_c(url1, year, url2, sep="")
  webpage <- read_html(url)
  movie_names <- webpage %>%
    html_nodes("table:nth-child(1) b a") %>%
    html_text()
  movie_urls <- webpage %>%
    html_nodes("table:nth-child(1) b a") %>%
    html_attr("href")
  movie_urls <- str_c("http://www.the-numbers.com", movie_urls)

  for(i in 1:10){
    moviepage <- movie_urls[i]
    info_table <- moviepage %>%
      read_html() %>%
      # html_nodes(xpath=`//*[(@id = "box_office_chart") and (((count(preceding-sibling::*) + 1) = 8) a
      # html_nodes(xpath=`//*[@id="box_office_chart"]/table`) %>%
      # html_nodes("#box_office_chart:nth-child(8)") %>%
      html_nodes("h2+ table") %>%
      html_table()
    info_table <- info_table[[1]]
    info_table <- info_table %>%
      filter(str_detect(X1, "MPAA")|str_detect(X1, "Genre"))
    movie_info_list[[i]] <<- data.table(Name=movie_names[i], Genre=info_table$X2[2], Info=info_table$X2
    #movie_info_list[[i]] <<- list(movie_names[i], info_table$X2[2], info_table$X2[1])
```

```
    #movie_info_df <<- rbind(movie_info_df, c(movie_names[i], info_table$X2[2], info_table$X2[1]))
  }
  count <- year - years[1] + 1
  master_list[[count]] <<- rbindlist(movie_info_list, use.names=TRUE)
  #master_list[[count]] <<- movie_info_df
}
```

Then, I run the function on years 1998-2004.

```
for (i in years){
  extract_year(i)
  # print(str_c("Finished ", i, sep=""))
}
```

```
## Warning in FUN(X[[i]], ...): failed to assign NativeSymbolInfo for lhs
## since lhs is already defined in the 'lazyeval' namespace
```

```
## Warning in FUN(X[[i]], ...): failed to assign NativeSymbolInfo for rhs
## since rhs is already defined in the 'lazyeval' namespace
```

```
movie_genres <- rbindlist(master_list)
```

At this point, the data are consolidated. I export a csv and a rda file.

```
write_csv(movie_genres, path="movie_genres.csv")
saveRDS(movie_genres, "movie_genres.rds")
```