Web Scraping Movie Data

Jonathan Che

8 April 2017

Overview

The kids-in-mind website has an alphabetical listing of all of their rated movies, but I cannot view all of them on one page. My goal here is to simply get all of the movies into one dataset.

Method

First, I make an empty matrix that's large enough to hold all 4503 movies on kids-in-mind.com.

```
mat_master <- matrix(NA, nrow=4507, ncol=6)    # 4503 films on kids-in-mind.com
```

Then, I construct a function that can iteratively scrape the information from each page.

```
add_movies <- function(x){
  url <- str_c(url1, x, url2, sep="")
  webpage <- read_html(url)
  string <- webpage %>%
    html_nodes(".t11normal+ p") %>%
    html_text()
  split_string <- str_split(string, "\n\n")[[1]]
  split_string <- split_string[2:(length(split_string)-1)]

  KIM_rating <- word(split_string, -1) %>%
    str_split("[.]")
  MPAA_rating <- word(split_string, -3) %>%
    str_sub(2,-2)
  Year <- word(split_string, -4) %>%
    str_sub(2,-2)
  Title <- word(split_string, 1, -5)

  temp <- cbind(Title, Year, MPAA_rating, do.call(rbind, KIM_rating))
  for (n in 1:dim(temp)[1]){
    mat_master[(n+x),] <<- temp[n,]    # Assign value to global variable
  }
}
```

Now, I loop through all the relevant pages of kids-in-mind.com

```
vec <- seq(from=0, to=4488, by=34)    # 4488 is label of last page
url1 <- "http://www.kids-in-mind.com/cgi-bin/listbyrating/search.pl?query=&stpos="
url2 <- "&stype=AND&s1=0&s2=10&v1=0&v2=10&p1=0&p2=10&m=1&m=2&m=3&m=4"

for(i in vec){
  add_movies(i)
}
```

Finaly, I output the master data frame

```r
movie_ratings <- data.frame(mat_master)
names(movie_ratings) <- c("Title", "Year", "MPAA_Rating", "Sex", "Violence", "Profanity")
write_csv(movie_ratings, path="movie_ratings.csv")
saveRDS(movie_ratings, "movie_ratings.rds")
```