w8_assignment

Frida Hæstrup

1/11/2020

```r
# Loading packages
library(pacman)
pacman::p_load(rvest, dplyr, tidyverse)
```

What top-rated tv-series can you watch?

When looking for a tv-series to watch, one can go to IMDB to look for the top-rated tv-series of all time. However, often, choosing a tv-series to watch also depends on the possibility of streaming it somewhere online. Here, I make a script that filters away the top-rated tv-series that are not available to stream on Netflix. Thus, when browsing the top-rated list, you do not have to manually go into Netflix to see whether each tv-series is available to watch. Instead, you are provided with a top-rated list only featuring those on Netflix to choose from.

Scraping IMDB

I found a list of the 250 top rated tv-series on IMBD (https://www.imdb.com/chart/toptv/?ref_=nv_tvv_250). Using the Google Chrome extension SelectorGadget, I was able to scrape the titles of the tv-series in this list. This is then loaded into R using the rvest library.

```r
# Specifying url to scrape from
imdb_url <- html("https://www.imdb.com/chart/toptv/?ref_=nv_tvv_250")

## Warning: 'html' is deprecated.
## Use 'xml2::read_html' instead.
## See help("Deprecated")

# Scraping titles of all tv-series of the url above and adding them to a dataframe
imdb <- as.data.frame(imdb_url %>%
  html_nodes(".titleColumn a") %>% #as specified from the selectorgadget-tool, the
".titleColumn a" tells R that it is the titles that should be extracted
  html_text())

# Changing column title
colnames(imdb) <- "title"
```

Scraping Netflix

It seems that Netflix does not have an API that makes it posssible for me to directly scrape the titles of tv-series. However, I found a GitHub-repository with information on all tv-series on Netflix (https://github.com/jameskang410/scraping-netflix/blob/master/TV%20Shows.json). I read this json file into R using the rjson-package. As I am only interested in the titles of the tv-series, I used a loop to extract the titles of all the tv-series and add them to a dataframe.

```
# Loading rjson package
pacman::p_load(rjson)

# Reading json file with information on tv-series from netflix
json_file <- "https://raw.githubusercontent.com/jameskang410/scraping-
netflix/master/TV%20Shows.json"
json_data <- fromJSON(paste(readLines(json_file), collapse="")) #function for
reading json-files

## Warning in readLines(json_file): ufuldstændig endelig linje fundet på 'https://
## raw.githubusercontent.com/jameskang410/scraping-netflix/master/TV%20Shows.json'

# Looping through the data to extract titles of tv-series
netflix <- as.data.frame(matrix(0, nrow = 1, ncol = 1))  #new df to be filled with
data from loop

for (i in 1:length(json_data$catalogItems)){ #looping through all tv-series
  current_title <- json_data$catalogItems[[i]]$title #extracting title
    netflix <- rbind(netflix, current_title) #adding title to netflix dataframe
}

# Changing column title
colnames(netflix) <- "title"
```

List of available tv-series

Now I have a dataframe ('netflix') with a column indicating the title of all tv-series on netflix and a dataframe ('imdb') with a column indicating the title of 250 top-rated tv-series. In order to check what series from the top-rated list are available to stream on Netflix, I filter out those that do not appear on both lists.

```
# Filter out those not appearing on both lists
possibilites <- netflix %>%
  filter(as.factor(title) %in% as.factor(imdb$title))
possibilites
```

```
##                                 title
## 1                  Making a Murderer
## 2                             Narcos
## 3  It's Always Sunny in Philadelphia
## 4                    Sons of Anarchy
## 5                     Peaky Blinders
## 6                             Dexter
## 7                               Life
## 8                       Detectorists
## 9                            Mad Men
## 10       Cosmos: A Spacetime Odyssey
## 11                          Sherlock
## 12                          Top Gear
## 13                       Black Books
## 14                      Human Planet
## 15               Friday Night Lights
## 16                           Firefly
## 17                        Twin Peaks
## 18    Star Trek: The Next Generation
## 19                         Peep Show
## 20                         Shameless
## 21         Alfred Hitchcock Presents
## 22                        Still Game
## 23       Mister Rogers' Neighborhood
## 24                    Long Way Round
## 25                              Life
## 26                     North & South
## 27                      Breaking Bad
## 28                            Archer
## 29                    House of Cards
## 30                      The IT Crowd
## 31                      Black Mirror
## 32                   BoJack Horseman
## 33                      Chef's Table
## 34              Parks and Recreation
## 35                             Louie
```

From this, we see the 35 out of the 250 top-rated tv-series that are available to watch on Netflix.