Scraping HTML

Carolina Alves de Lima Salge

2/1/2021

Web Scraping

In this document I will scrape data about the LEGO movie from the IMDB website

```r
library(rvest)
```

```
## Loading required package: xml2
```

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter()         masks stats::filter()
## x readr::guess_encoding() masks rvest::guess_encoding()
## x dplyr::lag()            masks stats::lag()
## x purrr::pluck()          masks rvest::pluck()
```

```r
lego_movie <- read_html("http://www.imdb.com/title/tt1490017/")

rating <- lego_movie %>%
  html_nodes("strong span") %>%
  html_text() %>%
  as.numeric()
rating
```

```
## [1] 7.7
```

```r
cast <- lego_movie %>%
  html_nodes(".primary_photo+ td a") %>%
  html_text() %>%
  trimws()
cast
```

```
##  [1] "Will Arnett"     "Elizabeth Banks" "Craig Berry"     "Alison Brie"
##  [5] "David Burrows"   "Anthony Daniels" "Charlie Day"     "Amanda Farinos"
##  [9] "Keith Ferguson"  "Will Ferrell"    "Will Forte"      "Dave Franco"
## [13] "Morgan Freeman"  "Todd Hansen"     "Jonah Hill"
```

```
poster <- lego_movie %>%
  html_nodes(".poster img") %>%
  html_attr("src")
poster
```

```
## [1] "https://m.media-amazon.com/images/M/MV5BMTg4MDk1ODExN15BMl5BanBnXkFtZTgwNzIyNjg3MDE@._V1_UX182_
```

```
lego <- tibble(rating = rating,
       cast = cast,
       poster = poster)
lego
```

```
## # A tibble: 15 x 3
##    rating cast           poster
##     <dbl> <chr>          <chr>
##  1    7.7 Will Arnett    https://m.media-amazon.com/images/M/MV5BMTg4MDk1ODExN15~
##  2    7.7 Elizabeth Ba~ https://m.media-amazon.com/images/M/MV5BMTg4MDk1ODExN15~
##  3    7.7 Craig Berry    https://m.media-amazon.com/images/M/MV5BMTg4MDk1ODExN15~
##  4    7.7 Alison Brie    https://m.media-amazon.com/images/M/MV5BMTg4MDk1ODExN15~
##  5    7.7 David Burrows https://m.media-amazon.com/images/M/MV5BMTg4MDk1ODExN15~
##  6    7.7 Anthony Dani~ https://m.media-amazon.com/images/M/MV5BMTg4MDk1ODExN15~
##  7    7.7 Charlie Day    https://m.media-amazon.com/images/M/MV5BMTg4MDk1ODExN15~
##  8    7.7 Amanda Farin~ https://m.media-amazon.com/images/M/MV5BMTg4MDk1ODExN15~
##  9    7.7 Keith Fergus~ https://m.media-amazon.com/images/M/MV5BMTg4MDk1ODExN15~
## 10    7.7 Will Ferrell   https://m.media-amazon.com/images/M/MV5BMTg4MDk1ODExN15~
## 11    7.7 Will Forte     https://m.media-amazon.com/images/M/MV5BMTg4MDk1ODExN15~
## 12    7.7 Dave Franco    https://m.media-amazon.com/images/M/MV5BMTg4MDk1ODExN15~
## 13    7.7 Morgan Freem~ https://m.media-amazon.com/images/M/MV5BMTg4MDk1ODExN15~
## 14    7.7 Todd Hansen    https://m.media-amazon.com/images/M/MV5BMTg4MDk1ODExN15~
## 15    7.7 Jonah Hill     https://m.media-amazon.com/images/M/MV5BMTg4MDk1ODExN15~
```

```
write_csv(lego, "lego.csv")
```