Web Scraping and Rectangling

William Cohen

3/27/2021

library(rvest)

## Warning: package 'rvest' was built under R version 4.0.4

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(tidyverse)

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

## v ggplot2 3.3.2 v purrr 0.3.4  
## v tibble 3.0.4 v stringr 1.4.0  
## v tidyr 1.1.2 v forcats 0.5.0  
## v readr 1.4.0

## Warning: package 'tibble' was built under R version 4.0.3

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x readr::guess\_encoding() masks rvest::guess\_encoding()  
## x dplyr::lag() masks stats::lag()

library(repurrrsive)

## Warning: package 'repurrrsive' was built under R version 4.0.4

#1.  
Link <- "https://www.imdb.com/list/ls091520106/?sort=list\_order,asc&st\_dt=&mode=detail&page=1"  
page = read\_html(Link)  
Title = page%>% html\_nodes(".lister-item-header a")%>%  
 html\_text()  
#Title  
  
Link <- "https://www.imdb.com/list/ls091520106/?sort=list\_order,asc&st\_dt=&mode=detail&page=1"  
page = read\_html(Link)  
Runtime = page%>% html\_nodes(".runtime")%>%  
 html\_text()  
#Runtime  
  
  
Link <- "https://www.imdb.com/list/ls091520106/?sort=list\_order,asc&st\_dt=&mode=detail&page=1"  
page = read\_html(Link)  
Ratings = page%>% html\_nodes(".ipl-rating-star.small .ipl-rating-star\_\_rating")%>%  
 html\_text()  
#Ratings  
parse\_number(Ratings) -> Ratings  
  
moviesdataframe = data.frame(Title,Ratings,Runtime)  
#moviesdataframe  
  
as\_tibble(moviesdataframe)

## # A tibble: 100 x 3  
## Title Ratings Runtime  
## <chr> <dbl> <chr>   
## 1 The Shawshank Redemption 9.3 142 min  
## 2 The Godfather 9.2 175 min  
## 3 The Godfather: Part II 9 202 min  
## 4 The Dark Knight 9 152 min  
## 5 12 Angry Men 9 96 min   
## 6 Schindler's List 8.9 195 min  
## 7 The Lord of the Rings: The Return of the King 8.9 201 min  
## 8 Pulp Fiction 8.9 154 min  
## 9 The Good, the Bad and the Ugly 8.8 178 min  
## 10 Fight Club 8.8 139 min  
## # ... with 90 more rows

#2.  
#a.  
#Nested lists are lists which contain lists.  
#For examples, X = list(5,7.5,"hello", y = (2.5,10,"goodbye",NA))  
#Here list y is nested within list X.  
  
#b.  
  
#gh\_repos  
  
#The data is unstructured and hard to work with.  
#There appear to be a number of lists within the 6 lists that make up gh\_repos.  
  
#c.  
repos <- tibble(repo = gh\_repos)  
repos

## # A tibble: 6 x 1  
## repo   
## <list>   
## 1 <list [30]>  
## 2 <list [30]>  
## 3 <list [30]>  
## 4 <list [26]>  
## 5 <list [30]>  
## 6 <list [30]>

#This code is showing me a tibble of the larger list which contains the 6 nested sub lists. This tibble shows that there are 6 lists, 5 of which have 30 elements, 1 with 26 elements.  
  
#d.  
  
#listviewer::jsonedit(gh\_repos)  
  
#The data is represented in a interactive display. The data is an array of 6 lists, each corresponding to a github user, each containing a number of lists representing repos. The nested list configuration is in place to better sort the data between users. The list within users setup is used to seperate and identify individual repos. 30 represents the number of repo lists in 5 of the 6 user lists. 68 represents the number of elements within each repo list the user list has.  
#e.  
  
repos %>%  
 unnest\_longer(repo) -> repos\_long  
  
repos\_long %>%  
 unnest\_wider(repo)

## # A tibble: 176 x 67  
## id name full\_name owner private html\_url description fork url   
## <int> <chr> <chr> <lis> <lgl> <chr> <chr> <lgl> <chr>  
## 1 6.12e7 after gaborcsa~ <nam~ FALSE https:/~ Run Code i~ FALSE http~  
## 2 4.05e7 argu~ gaborcsa~ <nam~ FALSE https:/~ Declarativ~ FALSE http~  
## 3 3.64e7 ask gaborcsa~ <nam~ FALSE https:/~ Friendly C~ FALSE http~  
## 4 3.49e7 base~ gaborcsa~ <nam~ FALSE https:/~ Do we get ~ FALSE http~  
## 5 6.16e7 cite~ gaborcsa~ <nam~ FALSE https:/~ Test R pac~ TRUE http~  
## 6 3.39e7 clis~ gaborcsa~ <nam~ FALSE https:/~ Unicode sy~ FALSE http~  
## 7 3.72e7 cmak~ gaborcsa~ <nam~ FALSE https:/~ port of cm~ TRUE http~  
## 8 6.80e7 cmark gaborcsa~ <nam~ FALSE https:/~ CommonMark~ TRUE http~  
## 9 6.32e7 cond~ gaborcsa~ <nam~ FALSE https:/~ <NA> TRUE http~  
## 10 2.43e7 cray~ gaborcsa~ <nam~ FALSE https:/~ R package ~ FALSE http~  
## # ... with 166 more rows, and 58 more variables: forks\_url <chr>,  
## # keys\_url <chr>, collaborators\_url <chr>, teams\_url <chr>, hooks\_url <chr>,  
## # issue\_events\_url <chr>, events\_url <chr>, assignees\_url <chr>,  
## # branches\_url <chr>, tags\_url <chr>, blobs\_url <chr>, git\_tags\_url <chr>,  
## # git\_refs\_url <chr>, trees\_url <chr>, statuses\_url <chr>,  
## # languages\_url <chr>, stargazers\_url <chr>, contributors\_url <chr>,  
## # subscribers\_url <chr>, subscription\_url <chr>, commits\_url <chr>,  
## # git\_commits\_url <chr>, comments\_url <chr>, issue\_comment\_url <chr>,  
## # contents\_url <chr>, compare\_url <chr>, merges\_url <chr>, archive\_url <chr>,  
## # downloads\_url <chr>, issues\_url <chr>, pulls\_url <chr>,  
## # milestones\_url <chr>, notifications\_url <chr>, labels\_url <chr>,  
## # releases\_url <chr>, deployments\_url <chr>, created\_at <chr>,  
## # updated\_at <chr>, pushed\_at <chr>, git\_url <chr>, ssh\_url <chr>,  
## # clone\_url <chr>, svn\_url <chr>, size <int>, stargazers\_count <int>,  
## # watchers\_count <int>, language <chr>, has\_issues <lgl>,  
## # has\_downloads <lgl>, has\_wiki <lgl>, has\_pages <lgl>, forks\_count <int>,  
## # open\_issues\_count <int>, forks <int>, open\_issues <int>, watchers <int>,  
## # default\_branch <chr>, homepage <chr>