HW6

Chao Chen Yu

3/22/2021

```
library(rvest)
```

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

```
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 3.3.3     v purrr   0.3.4
## v tibble  3.0.5     v stringr 1.4.0
## v tidyr   1.1.2     v forcats 0.5.0
## v readr   1.4.0
```

```
## -- 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()
```

```
library(repurrrsive)
```

1.

```
movie100 = read_html("100.html" , encoding = "UTF-8" )
movie100 %>% html_nodes(".lister-item-header a") %>% html_text() -> Title
movie100 %>% html_nodes(".ipl-rating-star.small .ipl-rating-star__rating") %>% html_text() -> Ratings
movie100 %>% html_nodes(".runtime") %>% html_text() -> Runtime
Runtime <- as.numeric(gsub("min","",Runtime))

top100 = data.frame(Title, Ratings, Runtime)
head(top100)
```

```
##                      Title Ratings Runtime
## 1 The Shawshank Redemption     9.3     142
## 2            The Godfather     9.2     175
## 3   The Godfather: Part II       9     202
## 4          The Dark Knight       9     152
## 5             12 Angry Men       9      96
## 6         Schindler's List     8.9     195
```

2-a List within another list either order list or unordered list is called nested list. If a list A is the list item of another list B, then list A would be called a nested list.

2-b

```
#listviewer::jsonedit(gh_repos)
```

There are six main lists in this data and in each main lists have 30 or 26 second sub-lists. Also, in each second sub-list have 68 variables. Lastly, in some of variable also have some lists.

2-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]>
```

After running this code, we can see that there are six lists but, in each list there are still have some list. So we still need to un-nest it.

2-d

```
#listviewer::jsonedit(gh_repos)
```

As I mentioned above, the nested list configuration can let us to arrange the data methodically. Since sometime we will have the large amount data, so that we can use the nested list configuration to display the content neatly.

The 30 means that it's a first layer and 68 means that it's still have a second layer. Go into the detailed configuration "rephos" data -> 6 lists -> 30 lists -> 68 variables -> list (some variable)

2-e

```
tibble(repo = gh_repos) %>%
  unnest_auto(repo) %>%
  unnest_auto(repo)
```

```
## Using ‘unnest_longer(repo)‘; no element has names
```

```
## Using ‘unnest_wider(repo)‘; elements have 68 names in common
```

```
## # 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>
```