Лабораторная работа №1

Галиченко Евгений Александрович

15 02 2021

C помощью пакета rvest или парсинга XML с помощью xpath запростов соберите данные с сайта согласно своему варианту. В итоговой таблице должно быть не менее 50 записей и не менее 5 признаков, из которых как минимум два количественныхю Снабдите файл справочником в формате Markdawn. Результаты: \* .csv-файл с данными, .md-файл со справочником, .Rmd-файл с кодом загрузки данных разместить в репозитории github. \* Файл .Rmd должен содержать постановку задачи и комментарии по ходу сбора данных.

library('rvest')

## Loading required package: xml2

# Ссылка на топ 20 фильмов 2018 года  
url <- 'https://www.kinopoisk.ru/top/y/2018/'  
  
webpage <- read\_html(url)  
  
# Получаем названия всех фильмов на русском языке  
names\_data <- webpage %>% html\_nodes(".all") %>% html\_text  
new\_names\_data <- names\_data[3:22]  
length(new\_names\_data)

## [1] 20

new\_names\_data

## [1] "Зеленая книга (2018)"   
## [2] "Мстители: Война бесконечности (2018)"   
## [3] "Капернаум (2018)"   
## [4] "Богемская рапсодия (2018)"   
## [5] "Остров собак (2018)"   
## [6] "Человек-паук: Через вселенные (2018)"   
## [7] "Они никогда не станут старше (2018)"   
## [8] "Фри-соло (2018)"   
## [9] "Отель Мумбаи: Противостояние (2018)"   
## [10] "Работа без авторства (2018)"   
## [11] "Укрась прощальное утро цветами обещания (2018)"  
## [12] "Апгрейд (2018)"   
## [13] "Первому игроку приготовиться (2018)"   
## [14] "Дэдпул 2 (2018)"   
## [15] "Поиск (2018)"   
## [16] "Лето (2018)"   
## [17] "Суперсемейка 2 (2018)"   
## [18] "Сорняки (2018)"   
## [19] "Магазинные воришки (2018)"   
## [20] "Фаворитка (2018)"

# Функция перебора тегов внутри тегов более высокого уровня  
get\_tags <- function(node){  
 raw\_data <- html\_nodes(node, selector) %>% html\_text  
 data\_NAs <- ifelse(length(raw\_data) == 0, NA, raw\_data)  
}  
  
selector <- '.text-grey'  
  
en\_names\_data = array()  
  
# Получаем названия всех фильмов на английском языке  
for(i in 1:length(new\_names\_data)){  
 tag <- paste0('#top250\_place\_', toString(i))  
 doc <- html\_nodes(webpage, tag)  
 en\_names <- sapply(doc, get\_tags)  
 en\_names\_data <- append(en\_names\_data, en\_names)  
}  
  
en\_names\_data = en\_names\_data[2:length(en\_names\_data)]  
en\_names\_data

## [1] "Green Book"   
## [2] "Avengers: Infinity War"   
## [3] "Capharnaum"   
## [4] "Bohemian Rhapsody"   
## [5] "Isle of Dogs"   
## [6] "Spider-Man: Into the Spider-Verse"   
## [7] "They Shall Not Grow Old"   
## [8] "Free Solo"   
## [9] "Hotel Mumbai"   
## [10] "Werk ohne Autor"   
## [11] "Sayonara no asa ni yakusoku no hana o kazaro"  
## [12] "Upgrade"   
## [13] "Ready Player One"   
## [14] "Deadpool 2"   
## [15] "Searching"   
## [16] NA   
## [17] "Incredibles 2"   
## [18] "Mauvaises herbes"   
## [19] "Manbiki kazoku"   
## [20] "The Favourite"

# Получаем рейтинг фильмов  
rating\_data <- webpage %>% html\_nodes(".continue") %>% html\_text  
length(rating\_data)

## [1] 20

# Превращаем строку в число  
rating\_data <- as.numeric(rating\_data)  
rating\_data

## [1] 8.231 7.956 7.941 7.926 7.908 7.905 7.753 7.739 7.572 7.562 7.462 7.432  
## [13] 7.416 7.388 7.351 7.344 7.329 7.297 7.293 7.287

# Получаем количество проголосовавших пользователей  
users\_data <- webpage %>% html\_nodes(".js-rum-hero div span") %>% html\_text  
# Избавляемся от скобок "()"  
users\_data <- gsub("[[:punct:]]", "", users\_data)  
# Избавляемся от пробелов  
users\_data <- gsub(pattern = "\\s", replacement = "", x=users\_data)  
length(users\_data)

## [1] 20

# Переводим строку в число  
users\_data <- as.numeric(users\_data)  
users\_data

## [1] 420514 514086 110371 319986 82326 253649 4969 4279 95079 46442  
## [11] 33896 205247 313852 330005 91221 62650 165580 4745 22742 78787

# Получаем ранк фильма в списке  
place\_data <- webpage %>% html\_nodes(".js-rum-hero td a") %>% html\_attr("name")  
bad <- is.na(place\_data)  
place\_data[!bad]

## [1] "1" "2" "3" "4" "5" "6" "7" "8" "9" "10" "11" "12" "13" "14" "15"  
## [16] "16" "17" "18" "19" "20"

# Оформляем все в дата фрейм  
DF\_top\_movies <- data.frame(Place = place\_data[!bad], Ru\_name = new\_names\_data,  
 En\_name = en\_names\_data,  
 Rating = rating\_data,  
 Vote = users\_data)  
  
data.dir <- './data'  
  
# Создаем директорию для данных  
if (!file.exists(data.dir)) {  
 dir.create(data.dir)  
}  
  
# Создаём файл с логом загрузок  
log.filename <- './data/download.log'  
if (!file.exists(log.filename)) file.create(log.filename)

## [1] TRUE

# Загружаем данные в .csv файл  
write.csv(DF\_top\_movies, file = './data/DF\_top\_movie.csv', row.names = FALSE)  
write(paste('Файл "DF\_top\_movies.csv" записан!', Sys.time()), file = log.filename, append = TRUE)