Упражнение №1

Маркин Артем

21 02 2021

library('rvest')

## Loading required package: xml2

url.1 <- 'https://market.yandex.ru/catalog--kholodilniki/71639/list?cpa=0&hid=15450081&lr=213&onstock=1&local-offers-first=0'  
url.2 <- 'https://market.yandex.ru/catalog--kholodilniki/71639/list?cpa=0&hid=15450081&onstock=1&page=2&local-offers-first=0'  
webpage.1 <- read\_html(url.1)  
webpage.2 <- read\_html(url.2)  
  
## Название холодильника  
names1 <- webpage.1 %>% html\_nodes(".cLo1fZHm2y") %>% html\_text  
names2 <- webpage.2 %>% html\_nodes(".cLo1fZHm2y") %>% html\_text  
names <- c(names1, names2)  
names

## [1] "Холодильник BBK RF-068"   
## [2] "Холодильник BBK RF-049"   
## [3] "Холодильник BBK RF-098"   
## [4] "Холодильник BBK RF-090"   
## [5] "Холодильник LG GA-B509CEUM"   
## [6] "Холодильник Gorenje NRK 6191 ES4"   
## [7] "Холодильник LG DoorCooling+ GA-B509 SVUM"   
## [8] "Холодильник LG GC-L247 CBDC"   
## [9] "Холодильник Weissgauff WRK 2000 XBNF"   
## [10] "Холодильник LG DoorCooling+ GA-B509SEUM"   
## [11] "Холодильник Weissgauff WCD 337 NFX"   
## [12] "Холодильник ATLANT ХМ 4208-000"   
## [13] "Холодильник Weissgauff WRK 2000 WNF"   
## [14] "Холодильник Haier HB25FSSAAA"   
## [15] "Холодильник LG GA-B379 SLUL"   
## [16] "Холодильник Haier C2F636CFRG"   
## [17] "Холодильник Bosch KGN39UL22R"   
## [18] "Холодильник LG DoorCooling+ GA-B509 SAUM"   
## [19] "Холодильник Galaxy GL3103"   
## [20] "Холодильник LG DoorCooling+ GC-Q22 FTBKL"   
## [21] "Холодильник LG DoorCooling+ GA-B 509PBAM"   
## [22] "Холодильник LG DoorCooling+ GA-B509MEUM"   
## [23] "Холодильник LG DoorCooling+ GA-B509CQWL"   
## [24] "Холодильник LG GA-B419 SQGL"   
## [25] "Холодильник Бирюса SBS 587 I"   
## [26] "Холодильник LG DoorCooling+ B509SBUM"   
## [27] "Холодильник LG DoorCooling+ GA-B509MMZL"   
## [28] "Холодильник Weissgauff WRK 2000 XNF"   
## [29] "Холодильник Bosch KGV36NL1AR"   
## [30] "Холодильник Haier CEF537AWD"   
## [31] "Холодильник LG GC-Q22FTAKL"   
## [32] "Холодильник Бирюса 118"   
## [33] "Холодильник ATLANT ХМ 6025-031"   
## [34] "Холодильник Weissgauff WCD 486 NFB"   
## [35] "Холодильник LG GA-B419 SLGL"   
## [36] "Холодильник Indesit EF 16"   
## [37] "Холодильник Hyundai CS4502F"   
## [38] "Холодильник LG DoorCooling+ GA-B509CCUM"   
## [39] "Холодильник LG GA-B509 CCIL"   
## [40] "Холодильник LG DoorCooling+ GA-B509 MCUM"   
## [41] "Холодильник Haier C2F537CWG"   
## [42] "Холодильник Haier C2F537CMSG"   
## [43] "Холодильник Candy CCRN 6200 W"   
## [44] "Холодильник ATLANT Х 1602-100"   
## [45] "Холодильник Indesit ITF 018 W"   
## [46] "Холодильник Weissgauff WCD 486 NFX"   
## [47] "Холодильник Бирюса M90"   
## [48] "Холодильник LG DoorCooling+ GA-B509CLWL"

## Описание холодильника  
description1 <- webpage.1 %>% html\_nodes(".\_2\_oj-OEI-o") %>% html\_text  
description2 <- webpage.2 %>% html\_nodes(".\_2\_oj-OEI-o") %>% html\_text  
description <- c(description1, description2)  
description

## [1] "ШхВхГ: 44.50х62х46.50 смобщий объем: 68 лкласс энергопотребления: A+ручное размораживаниеобъем холодильной камеры: 62 л"   
## [2] "ШхВхГ: 45х51х45 смобщий объем: 45 лкласс энергопотребления: Aручное размораживаниеобъем холодильной камеры: 42 л"   
## [3] "ШхВхГ: 45.50х83.20х54 смобщий объем: 98 лкласс энергопотребления: A+ручное размораживаниеобъем холодильной камеры: 65 л"   
## [4] "ШхВхГ: 44.50х83х46.50 смобщий объем: 90 лкласс энергопотребления: A+ручное размораживаниеобъем холодильной камеры: 82 л"   
## [5] "ШхВхГ: 59.50х203х68.20 смобщий объем: 384 лкласс энергопотребления: A++No Frostобъем холодильной камеры: 277 л"   
## [6] "ШхВхГ: 60х185х59.20 смобщий объем: 302 лкласс энергопотребления: A+No Frostобъем холодильной камеры: 203 л"   
## [7] "ШхВхГ: 59.50х203х68.20 смобщий объем: 384 лкласс энергопотребления: A++No Frostобъем холодильной камеры: 277 л"   
## [8] "ШхВхГ: 91.20х179х73.80 смобщий объем: 601 лNo Frostобъем холодильной камеры: 405 лобъем морозильной камеры: 196 л"   
## [9] "ШхВхГ: 60х193.50х67 смобщий объем: 342 лкласс энергопотребления: A+No Frostобъем холодильной камеры: 245 л"   
## [10] "ШхВхГ: 59.50х203х68.20 смобщий объем: 384 лкласс энергопотребления: A+No Frostобъем холодильной камеры: 277 л"   
## [11] "ШхВхГ: 79.50х180х61.20 смобщий объем: 404 лкласс энергопотребления: A+No Frostобъем холодильной камеры: 225 л"   
## [12] "линейка: COMPACT 42 SerieШхВхГ: 54.50х142.50х57.20 смобщий объем: 173 лкласс энергопотребления: Aкапельная система разморозки"   
## [13] "ШхВхГ: 60х193.50х67 смобщий объем: 342 лкласс энергопотребления: A+No Frostобъем холодильной камеры: 245 л"   
## [14] "ШхВхГ: 100.50х190х76 смобщий объем: 655 лкласс энергопотребления: A++No Frostобъем холодильной камеры: 426 л"   
## [15] "ШхВхГ: 59.50х173.70х65.50 смобщий объем: 261 лкласс энергопотребления: A+No Frostобъем холодильной камеры: 182 л"   
## [16] "ШхВхГ: 59.50х190х67.20 смобщий объем: 364 лкласс энергопотребления: A+No Frostобъем холодильной камеры: 235 л"   
## [17] "линейка: Serie 4ШхВхГ: 60х203х66 смобщий объем: 419 лкласс энергопотребления: A+No Frost"   
## [18] "ШхВхГ: 59.50х203х68.20 смобщий объем: 384 лкласс энергопотребления: A++No Frostобъем холодильной камеры: 277 л"   
## [19] "ШхВхГ: 43х49х45 смобщий объем: 45 лкласс энергопотребления: Aручное размораживаниеобъем холодильной камеры: 42 л"   
## [20] "ШхВхГ: 83.50х178.70х73.40 смобщий объем: 458 лкласс энергопотребления: A+No Frostобъем холодильной камеры: 315 л"   
## [21] "ШхВхГ: 59.50х203х68.20 смобщий объем: 384 лкласс энергопотребления: A++No Frostобъем холодильной камеры: 277 л"   
## [22] "ШхВхГ: 59.50х203х68.20 смобщий объем: 384 лкласс энергопотребления: A++No Frostобъем холодильной камеры: 277 л"   
## [23] "ШхВхГ: 59.50х203х68.20 смобщий объем: 384 лкласс энергопотребления: A+No Frostобъем холодильной камеры: 277 л"   
## [24] "ШхВхГ: 59.50х190.70х65.50 смобщий объем: 302 лкласс энергопотребления: A+No Frostобъем холодильной камеры: 223 л"   
## [25] "ШхВхГ: 89.50х178.80х69 смобщий объем: 510 лкласс энергопотребления: A+No Frostобъем холодильной камеры: 335 л"   
## [26] "ШхВхГ: 59.50х203х68.20 смобщий объем: 384 лкласс энергопотребления: A++No Frostобъем холодильной камеры: 277 л"   
## [27] "ШхВхГ: 59.50х203х68.20 смобщий объем: 384 лкласс энергопотребления: A++No Frostобъем холодильной камеры: 277 л"   
## [28] "ШхВхГ: 60х193.50х67 смобщий объем: 342 лкласс энергопотребления: A+No Frostобъем холодильной камеры: 245 л"   
## [29] "ШхВхГ: 60х185х65 смобщий объем: 317 лкласс энергопотребления: Aкапельная система разморозкиобъем холодильной камеры: 223 л"   
## [30] "ШхВхГ: 59.50х200х65 смобщий объем: 368 лкласс энергопотребления: ANo Frostобъем холодильной камеры: 263 л"   
## [31] "ШхВхГ: 83.50х178.70х73.40 смобщий объем: 458 лкласс энергопотребления: A+No Frostобъем холодильной камеры: 315 л"   
## [32] "ШхВхГ: 48х145х60.50 смобщий объем: 180 лкласс энергопотребления: Aкапельная система разморозкиобъем холодильной камеры: 125 л"   
## [33] "линейка: SOFT LINE 60 SerieШхВхГ: 60х205х63 смобщий объем: 364 лкласс энергопотребления: Aкапельная система разморозки"   
## [34] "ШхВхГ: 79.50х180х69.20 смобщий объем: 401 лкласс энергопотребления: A+No Frostобъем холодильной камеры: 268 л"   
## [35] "ШхВхГ: 59.50х190.70х65.50 смобщий объем: 302 лкласс энергопотребления: A+No Frostобъем холодильной камеры: 223 л"   
## [36] "ШхВхГ: 60х167х64 смобщий объем: 256 лкласс энергопотребления: ANo Frostобъем холодильной камеры: 181 л"   
## [37] "ШхВхГ: 83.60х178х63.60 смобщий объем: 476 лкласс энергопотребления: A+No Frostобъем холодильной камеры: 271 л"   
## [38] "ШхВхГ: 59.50х203х68.20 смобщий объем: 384 лкласс энергопотребления: A++No Frostобъем холодильной камеры: 277 л"   
## [39] "ШхВхГ: 59.50х203х68.20 смобщий объем: 384 лкласс энергопотребления: A+No Frostобъем холодильной камеры: 277 л"   
## [40] "ШхВхГ: 59.50х203х68.20 смобщий объем: 384 лкласс энергопотребления: A++No Frostобъем холодильной камеры: 277 л"   
## [41] "ШхВхГ: 59.50х199.80х67.20 смобщий объем: 386 лкласс энергопотребления: ANo Frostобъем холодильной камеры: 278 л"   
## [42] "ШхВхГ: 59.50х199.80х67.20 смобщий объем: 386 лкласс энергопотребления: ANo Frostобъем холодильной камеры: 278 л"   
## [43] "ШхВхГ: 59.50х200х65.70 смобщий объем: 370 лкласс энергопотребления: ANo Frostобъем холодильной камеры: 264 л"   
## [44] "ШхВхГ: 59.50х186.80х62.90 смобщий объем: 371 лкласс энергопотребления: A+капельная система разморозкиобъем холодильной камеры: 370 л"   
## [45] "ШхВхГ: 60х185х64 смобщий объем: 298 лкласс энергопотребления: ANo Frostобъем холодильной камеры: 223 л"   
## [46] "ШхВхГ: 79.50х180х69.20 смобщий объем: 472 лкласс энергопотребления: A+No Frostобъем холодильной камеры: 268 л"   
## [47] "ШхВхГ: 47х85х45 смобщий объем: 93 лкласс энергопотребления: A+ручное размораживаниеколичество компрессоров: один"   
## [48] "ШхВхГ: 59.50х203х68.20 смобщий объем: 384 лкласс энергопотребления: A+No Frostобъем холодильной камеры: 277 л"

## Цена холодильника  
price1 <- webpage.1 %>% html\_nodes(".\_3f2ZtYT7NH") %>% html\_text  
price2 <- webpage.2 %>% html\_nodes(".\_3f2ZtYT7NH") %>% html\_text  
price <- c(price1, price2)  
price

## [1] "7 400 <U+20BD>" "6 815 <U+20BD>" "12 319 <U+20BD>" "9 498 <U+20BD>" "49 990 <U+20BD>" "25 550 <U+20BD>"  
## [7] "34 990 <U+20BD>" "119 690 <U+20BD>" "43 990 <U+20BD>" "44 990 <U+20BD>" "59 990 <U+20BD>" "18 132 <U+20BD>"  
## [13] "40 990 <U+20BD>" "193 238 <U+20BD>" "32 499 <U+20BD>" "43 990 <U+20BD>" "49 990 <U+20BD>" "53 700 <U+20BD>"  
## [19] "8 423 <U+20BD>" "123 000 <U+20BD>" "60 400 <U+20BD>" "44 400 <U+20BD>" "33 485 <U+20BD>" "29 490 <U+20BD>"  
## [25] "36 240 <U+20BD>" "59 100 <U+20BD>" "32 760 <U+20BD>" "42 990 <U+20BD>" "29 180 <U+20BD>" "37 990 <U+20BD>"  
## [31] "117 990 <U+20BD>" "13 299 <U+20BD>" "25 990 <U+20BD>" "69 990 <U+20BD>" "32 490 <U+20BD>" "20 899 <U+20BD>"  
## [37] "35 690 <U+20BD>" "43 000 <U+20BD>" "56 990 <U+20BD>" "49 990 <U+20BD>" "40 000 <U+20BD>" "40 100 <U+20BD>"  
## [43] "23 190 <U+20BD>" "21 489 <U+20BD>" "20 070 <U+20BD>" "69 990 <U+20BD>" "7 899 <U+20BD>" "29 999 <U+20BD>"

# Оставляем только числа  
price <- gsub("[^[:digit:]]", "", price)  
# Меняем тип со строки в числа  
price <- as.numeric(price)  
price

## [1] 7400 6815 12319 9498 49990 25550 34990 119690 43990 44990  
## [11] 59990 18132 40990 193238 32499 43990 49990 53700 8423 123000  
## [21] 60400 44400 33485 29490 36240 59100 32760 42990 29180 37990  
## [31] 117990 13299 25990 69990 32490 20899 35690 43000 56990 49990  
## [41] 40000 40100 23190 21489 20070 69990 7899 29999

## Рейтинг холодильника  
# функция перебора тегов внутри тегов более высокого уровня  
get\_tags <- function(node){  
 # найти все теги с рейтингом  
 raw\_data <- html\_nodes(node, selector) %>% html\_text  
 # значения нулевой длины меняем на пропуски  
 data\_NAs <- ifelse(length(raw\_data) == 0, NA, raw\_data)  
}  
  
selector <- '.\_1iKHblnc3a'  
  
# Парсим рейтинг с первой страницы  
doc1 <- html\_nodes(webpage.1, '.\_1B9w\_GzQuM')  
rating1 <- sapply(doc1, get\_tags)  
# Парсим рейтинг со второй страницы  
doc2 <- html\_nodes(webpage.2, '.\_1B9w\_GzQuM')  
rating2 <- sapply(doc2, get\_tags)  
  
rating <- c(rating1, rating2)  
# Превращаем строку в число  
rating <- as.numeric(rating)  
rating

## [1] 4.3 4.4 4.6 4.2 5.0 4.7 4.3 5.0 4.9 4.8 4.6 4.5 4.6 4.8 4.6 4.7 4.8 4.4 NA  
## [20] 4.6 5.0 5.0 4.7 4.5 4.7 4.3 NA 4.7 4.5 NA 4.5 4.5 4.6 4.9 4.5 4.6 4.8 NA  
## [39] 4.3 NA 4.5 4.7 4.6 4.6 4.6 4.8 4.8 4.7

## Количество отзывов  
selector <- '.KdrkCVDrVm'  
doc1 <- html\_nodes(webpage.1, '.\_1B9w\_GzQuM')  
feedback1 <- sapply(doc1, get\_tags)  
doc2 <- html\_nodes(webpage.2, '.\_1B9w\_GzQuM')  
feedback2 <- sapply(doc2, get\_tags)  
  
feedback <- c(feedback1, feedback2)  
# Оставляем только числа  
feedback <- gsub("[^[:digit:]]", "", feedback)  
# Строку в число  
feedback <- as.numeric(feedback)  
feedback

## [1] 8 4 3 7 4 19 7 5 17 4 23 195 29 4 82 20 20 10 NA  
## [20] 19 3 2 48 53 16 9 NA 37 27 NA NA 77 240 24 34 64 11 NA  
## [39] 6 NA 19 19 36 26 36 21 1 75

data.dir <- './data'  
  
# Создаём директорию для данных, если она ещё не существует:  
if (!file.exists(data.dir)) {  
 dir.create(data.dir)  
}  
  
# Создаём файл с логом загрузок, если он ещё не существует:  
log.filename <- './data/download.log'  
if (!file.exists(log.filename)) file.create(log.filename)  
  
df <- data.frame(Name = names, Description = description,   
 Price = price, Rating = rating, Feedback = feedback)  
  
# Записываем в .csv  
write.csv(df, file = './data//yandex\_market.csv', row.names = F)  
# запись в лог  
write(paste('Файл "yandex\_market.csv" создан:', Sys.time()),   
 file = log.filename, append = T)