How to scrape JUMIA.com using R programming

Henrys kasereka

January 7 2021

Introduction

Jumia is an online marketplace for electronics and fashion, among others, targeting several African countries, but headquartered and incorporated in Germany. The company is also a logistics service, which enables the shipment and delivery of packages from sellers to consumers, and a payment service, which facilitates transactions between active participants and the platform of Jumia in selected markets. It has established partnerships with more than 50,000 local African businesses.

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     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 ------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(rvest)
```

```
## Loading required package: xml2

##
## Attaching package: 'rvest'

## The following object is masked from 'package:purrr':
##
##     pluck

## The following object is masked from 'package:readr':
##
##     guess_encoding
```

```
library(stringi)
library(stringr)
library(dplyr)

#Jumia websites
url_jumia <- 'https://group.jumia.com/'
url_jumia_site <- url_jumia %>% read_html() %>% html_nodes('.n2mu-single-client > a') %>% html_attr('hr
url_jumia_site[1] <- url_jumia
url_jumia_site
```

```
##  [1] "https://group.jumia.com/"  "https://www.jumia.com.eg/"
##  [3] "https://www.jumia.ma"      "https://www.jumia.co.ke"
##  [5] "https://www.jumia.ci"      "https://www.zando.co.za"
##  [7] "https://www.jumia.com.tn/" "https://www.jumia.dz/"
##  [9] "https://www.jumia.com.gh/" "https://www.jumia.sn/"
## [11] "https://www.jumia.ug/"
```

Liste of jumia website in Africa.

```
url_jumia_site
```

```
##  [1] "https://group.jumia.com/"  "https://www.jumia.com.eg/"
##  [3] "https://www.jumia.ma"      "https://www.jumia.co.ke"
##  [5] "https://www.jumia.ci"      "https://www.zando.co.za"
##  [7] "https://www.jumia.com.tn/" "https://www.jumia.dz/"
##  [9] "https://www.jumia.com.gh/" "https://www.jumia.sn/"
## [11] "https://www.jumia.ug/"
```

Web scraping

Web scraping, also known as data mining, is the process of collecting large amounts of data from the web and then placing it in databases for future analysis and later use.

The algorithm we will provide can scrape all data on Jumia.

Check if jumia website allow us to be scrape

```
library(robotstxt)
paths_allowed("https://group.jumia.com/")
```

```
##  group.jumia.com
```

```
## [1] TRUE
```

Let's go

Specify the url of the category to scrape

```
#urlbas <- "https://www.jumia.co.ke/smartphones/"
#urlbas <- 'https://www.jumia.co.ke/laptops/'
urlbas <- "https://www.jumia.com.ng/smartphones/"
```

Get the number of dataset to scrape

```
products_found <- urlbas %>% read_html() %>% html_nodes(".-fs14.-gy5.-phs") %>% html_text()
products_found
```

```
## [1] "15563 products found"
```

Split products found

```
products_found_number <- str_split_fixed(products_found, " ", 2)
products_found_number[1]
```

```
## [1] "15563"
```

Generate pagination link

```
products_by_page <- 48
page_number <- round(as.numeric(products_found_number[1]) / products_by_page)
page_number
```

```
## [1] 324
```

Get the all links

```
list_of_pages <- str_c(urlbas,'?page=', 1:page_number ,'#catalog-listing')
head(list_of_pages, n = 10)
```

```
##  [1] "https://www.jumia.com.ng/smartphones/?page=1#catalog-listing"
##  [2] "https://www.jumia.com.ng/smartphones/?page=2#catalog-listing"
##  [3] "https://www.jumia.com.ng/smartphones/?page=3#catalog-listing"
##  [4] "https://www.jumia.com.ng/smartphones/?page=4#catalog-listing"
##  [5] "https://www.jumia.com.ng/smartphones/?page=5#catalog-listing"
##  [6] "https://www.jumia.com.ng/smartphones/?page=6#catalog-listing"
##  [7] "https://www.jumia.com.ng/smartphones/?page=7#catalog-listing"
##  [8] "https://www.jumia.com.ng/smartphones/?page=8#catalog-listing"
##  [9] "https://www.jumia.com.ng/smartphones/?page=9#catalog-listing"
## [10] "https://www.jumia.com.ng/smartphones/?page=10#catalog-listing"
```

Create an empty table to store results

```
result_table <- tibble()
```

Start scraping

```
for(page in list_of_pages){
  page_source <- read_html(page)
  title <- html_nodes(page_source,'.name') %>% html_text()
  price <- html_nodes(page_source,'.prc') %>% html_text()
  temp_table <- tibble(title = title, price = price)
  result_table <- bind_rows(result_table, temp_table)
}
```

Remove empty values

```
result_table <- result_table %>% filter(price != '')
head(result_table, n = 10)
```

```
## # A tibble: 10 x 2
##    title                                                              price
##    <chr>                                                              <chr>
##  1 "Gionee S11 Lite 5.7-Inch HD (4GB,64GB ROM) Android 7.1 (13MP + 2MP~ ₦ 42,510
##  2 "Samsung Samsung Galaxy A31 (128GB, 4GB RAM) 6.4\" FHD + 4 Rear Cam~ ₦ 120,0~
##  3 "Oukitel C15 Pro 6.1-Inch Android 9.0 Pie(3GB RAM 32GB ROM), 8.0MP ~ ₦ 36,900
##  4 "Samsung Galaxy A11 6.4\" 13MP+5MP+2MP Camera, 2/32GB Memory, 4000m~ ₦ 80,000
##  5 "Cubot Note 7, 5.5 Inches,4G LTE,2GB + 16GB,3100mAh  (Dual SIM),Tri~ ₦ 31,500
##  6 "Cubot Note 7, 5.5 Inches,4G LTE,2GB + 16GB,3100mAh  (Dual SIM),Tri~ ₦ 31,500
##  7 "Cubot Note 20, 6.5\", 3GB+64GB, (Dual SIM) Quad Camera 12MP Androi~ ₦ 49,050
##  8 "Samsung Galaxy S20 FE 6.5-Inch (6GB 128GB ROM) (12MP + 12MP + 8MP)~ ₦ 350,0~
##  9 "Oppo A93 Dual SIM, 128GB, 8GB RAM, 4G LTE, 48 MP + 8 MP + 2 MP + 2~ ₦ 129,9~
## 10 "Oppo A93 6.43''  8GB RAM 128GB ROM Andriod 10, (48MP + 8MP + 2MP +~ ₦ 135,0~
```

Export data

```
currentTime <- Sys.time()
csvFileName <- paste("resultatdata", currentTime,".csv", sep = ",")
write.csv(result_table, file = csvFileName, fileEncoding = "UTF-16LE")
```

Contributing

Pull requests are welcome. For major changes, please open an issue first to discuss what you would like to change.

Please make sure to update tests as appropriate.

License

Email

arisjokov@gmail.com