Home

PUBLIC

🌐 Stack Overflow

  Tags

  Users

  Jobs

Teams
Q&A for work

Learn More

Web Scraping in R from Google Images

Ask Question

I have used "rvest" package to web scrape for
different purposes. Now I need to use it to get
source of an image object (png) from google
images. I have tried the solution on this link: Web
scraping of image. It does exactly what I want to do.
So I come up with the code below but my
html_nodes function gets empty object.

```
library("rvest")
page <- read_html("https://www.google.com.tr/search?
q=manitou&espv=2&biw=1366&bih=662&source=lnms&tbm=isch&sa=X&ved=0ahUKEwjCnJ6H2ITRAhWCQBoKH

node <- html_nodes(page,xpath='//*[@id="rg_s"]/div[1]/a/img')
src <-  html_attr(node,"src")
```

I also tried css selector and the name of the image
as it is done on the link I gave above. My node
object is empty in any ways. I should also point out
that I want to scrape the source of very first image
on the link, which has the xpath that I wrote above.
Thank you in advance.

html   r   image   xpath   web-scraping

edited Jun 26 '17 at 15:16

Mike Wise
13.1k   4   38   64

asked Dec 21 '16 at 7:51

Barsssk
48   6

1 Answer

I think it is working fine, you just don't understand
the makeup of that file well enough yet, i.e. there is
probably no node corresponding to the xpath
selector you wrote.

Here for example I select all the `<img>` nodes and
print them out:

```
library("rvest")
page <- read_html("https://www.google.com.tr/search?
q=manitou&espv=2&biw=1366&bih=662&source=lnms&tbm=is

node <- html_nodes(page,xpath = '//img')
node
```

```
  [2] <img height="113" src="https://encrypted-tbn2.g
q=tbn:ANd9GcRg92_01ZbpYpV_agaHP4M3GoRoaCsZW5Sym8eqcX
alt="manitou ile ilgili görsel s ...
  [3] <img height="98" src="https://encrypted-tbn1.gs
q=tbn:ANd9GcSbJOecoEPbrJjZ-TjJMgMwlulXRMPLBWZX45vwUJ
alt="manitou ile ilgili görsel so ...
  [4] <img height="79" src="https://encrypted-tbn3.gs
-9B7R3O3OZJFrDsuOUuP94HwwNw-av9tUyjziG3sCl6M9s7G4" w
görsel so ...
  [5] <img height="95" src="https://encrypted-tbn0.gs
q=tbn:ANd9GcTkibMqBWEifcyw_d-vrNob6UqYP-hDFPoQG2pkzV
alt="manitou ile ilgili görsel so ...
  [6] <img height="91" src="https://encrypted-tbn2.gs
-7QrQwovNBUHIpDFHe8Zwwad3UIvnwppv74GRIrsI1XYNPkFOg"
görsel s ...
  [7] <img height="112" src="https://encrypted-tbn0.g
q=tbn:ANd9GcS1gpUEBucliP4WK2_22K4wElI2lIrDs2PZT7sRCL
alt="manitou ile ilgili görsel  ...
  [8] <img height="69" src="https://encrypted-tbn2.gs
q=tbn:ANd9GcSssiUhuZe_1YmQ9dwmYHdKoFXyQBj9IQPGX_LU8m
alt="manitou ile ilgili görsel s ...
  [9] <img height="113" src="https://encrypted-tbn0.g
q=tbn:ANd9GcTCM9Mu6K63QpzNk20HFrHkybi--dw3JPu5JDd4LS
alt="manitou ile ilgili görsel s ...
 [10] <img height="95" src="https://encrypted-tbn1.gs
q=tbn:ANd9GcS8sI3fBSJmjftqC9Rx2bhXh_xgP3-nS2WuD2as9U
alt="manitou ile ilgili görsel so ...
 [11] <img height="83" src="https://encrypted-tbn0.gs
gf45JbC4Q4lD3hioj_CP6imrO5RUWBeW6IuygNaN8LM1qydX56l5
ilgili görsel s ...
 [12] <img height="84" src="https://encrypted-tbn0.gs
q=tbn:ANd9GcS6tnxPJYeS48IoNALN0D52U5TNjmq7Ta-GcPNifM
alt="manitou ile ilgili görsel s ...
 [13] <img height="140" src="https://encrypted-tbn1.g
q=tbn:ANd9GcTwmI9PxfLBT2dCPnR04I9pXmK8V9whAI2yEv4dX5
alt="manitou ile ilgili görse ...
 [14] <img height="71" src="https://encrypted-tbn1.gs
q=tbn:ANd9GcQNx2Pe1AZtT-0XQ44HSurWO6O2syXrXG6YPfggtZ
alt="manitou ile ilgili görsel so ...
 [15] <img height="130" src="https://encrypted-tbn0.g
q=tbn:ANd9GcRGACLfeRm6U0xwSeYncSUDQtcd4noTewVF4aGnQc
alt="manitou ile ilgili görsel  ...
 [16] <img height="107" src="https://encrypted-tbn3.g
q=tbn:ANd9GcQ1RwAscQpzVXfquuAoPaLE9hFMuZSOpo6ckOzdpk
alt="manitou ile ilgili görsel s ...
 [17] <img height="98" src="https://encrypted-tbn2.gs
q=tbn:ANd9GcTE5sLf71TxAYla6nlfLRgXwL1IC-gXzXQRq1ZcnE
alt="manitou ile ilgili görsel so ...
 [18] <img height="91" src="https://encrypted-tbn1.gs
Hc0Muy6Vjw5OLQZocflSCqR3oz0GBRu3Bs7_JCoNyjr5vjNP7KZ4
görsel s ...
 [19] <img height="68" src="https://encrypted-tbn2.gs
q=tbn:ANd9GcR8R_39V3bxWJUDdNhrsAS6YOYEg6U-QpaLEV0MQ5
alt="manitou ile ilgili görsel so ...
 [20] <img height="99" src="https://encrypted-tbn0.gs
q=tbn:ANd9GcTIrnwcUbo9WYT-gyvrLb5g4JFEc27odkzzU6Swzx
alt="manitou ile ilgili görsel so ...
 ...
>
```

And here is the first node:

```
>node[[1]]
{xml_node} <img style="padding-top:2px"
src="/textinputassistant/tia.png" onclick="(function
   text_input_assistant_js='/textinputassistant/11/t
  document.createElement('script');s.src =
  text_input_assistant_js;(document.getElementById('
  document.body).appendChild(s);})();"
  alt="" height="23" width="27">
```

answered Dec 21 '16 at 9:41

Thank you Mike, taking all the images and going the
elements individually from that element solved it. I
know how to handle objects but I had empty object

1   Err, could you mark this correct? — Mike Wise Dec 22
     '16 at 11:41

to learn, share knowledge, and build your career.