rvest package

Aboubacar HEMA,ITS4

4 juillet 2020

library(rvest)

## Warning: package 'rvest' was built under R version 3.6.3

## Loading required package: xml2

## Warning: package 'xml2' was built under R version 3.6.3

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 3.6.3

## -- Attaching packages --------------------------------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.0 v purrr 0.3.4  
## v tibble 3.0.1 v dplyr 0.8.5  
## v tidyr 1.0.2 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.5.0

## Warning: package 'ggplot2' was built under R version 3.6.3

## Warning: package 'tibble' was built under R version 3.6.3

## Warning: package 'tidyr' was built under R version 3.6.3

## Warning: package 'readr' was built under R version 3.6.3

## Warning: package 'purrr' was built under R version 3.6.3

## Warning: package 'dplyr' was built under R version 3.6.3

## Warning: package 'stringr' was built under R version 3.6.3

## Warning: package 'forcats' was built under R version 3.6.3

## -- 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()

path <- system.file("html-ex", "bad-encoding.html", package = "rvest")  
x <- read\_html(path)  
x %>% html\_nodes("p") %>% html\_text()

## [1] "<c9>migré cause célèbre déjà vu."

#guess\_encoding(x)  
# Two valid encodings, only one of which is correct  
read\_html(path, encoding = "ISO-8859-1") %>% html\_nodes("p") %>% html\_text()

## [1] "Émigré cause célèbre déjà vu."

read\_html(path, encoding = "ISO-8859-2") %>% html\_nodes("p") %>% html\_text()

## [1] "Émigré cause célcbre déjr vu."

# From a url:  
google <- read\_html("http://google.com", encoding = "ISO-8859-1")  
google %>% xml\_structure()

## <html [itemscope, itemtype, lang]>  
## <head>  
## <meta [content, http-equiv]>  
## <meta [content, itemprop]>  
## <title>  
## {text}  
## <script [nonce]>  
## {cdata}  
## <style>  
## {cdata}  
## <style>  
## {cdata}  
## <script [nonce]>  
## <body [bgcolor]>  
## <script [nonce]>  
## {cdata}  
## <div [id]>  
## <div [id]>  
## <nobr>  
## <b [class]>  
## {text}  
## {text}  
## <a [class, href]>  
## {text}  
## {text}  
## <a [class, href]>  
## {text}  
## {text}  
## <a [class, href]>  
## {text}  
## {text}  
## <a [class, href]>  
## {text}  
## {text}  
## <a [class, href]>  
## {text}  
## {text}  
## <a [class, href]>  
## {text}  
## {text}  
## <a [class, href]>  
## {text}  
## {text}  
## <a [class, style, href]>  
## <u>  
## {text}  
## {text}  
## <div [id, width]>  
## <nobr>  
## <span [id, class]>  
## <span [id, class]>  
## <span [id]>  
## <a [href, class]>  
## {text}  
## {text}  
## <a [href, class]>  
## {text}  
## {text}  
## <a [target, id, href, class]>  
## {text}  
## <div [class, style]>  
## <div [class, style]>  
## <center>  
## <br [clear, id]>  
## <div [id]>  
## <img [alt, height, src, style, width, id]>  
## <br>  
## <br>  
## <form [action, name]>  
## <table [cellpadding, cellspacing]>  
## <tr [valign]>  
## <td [width]>  
## {text}  
## <td [align, nowrap]>  
## <input [name, value, type]>  
## <input [value, name, type]>  
## <input [name, type, value]>  
## <input [name, type]>  
## <input [name, type]>  
## <div [class, style]>  
## <input [class, style, autocomplete, value, title, maxlength, name, size]>  
## <br [style]>  
## <span [class]>  
## <span [class]>  
## <input [class, value, name, type]>  
## <span [class]>  
## <span [class]>  
## <input [class, id, value, name, type]>  
## <script [nonce]>  
## {cdata}  
## <input [value, name, type]>  
## <td [class, align, nowrap, width]>  
## <a [href]>  
## {text}  
## <input [id, name, type, value]>  
## <script [nonce]>  
## {cdata}  
## <div [id]>  
## <div [style]>  
## <br>  
## <div [id]>  
## <style>  
## {cdata}  
## <div [id]>  
## {text}  
## <a [href]>  
## {text}  
## {text}  
## <span [id]>  
## <div [style]>  
## <div [style, id]>  
## <a [href]>  
## {text}  
## <a [href]>  
## {text}  
## <a [href]>  
## {text}  
## <a [href]>  
## {text}  
## <p [style]>  
## {text}  
## <a [href]>  
## {text}  
## {text}  
## <a [href]>  
## {text}  
## <script [nonce]>  
## {cdata}

google %>% html\_nodes("div")

## {xml\_nodeset (13)}  
## [1] <div id="mngb">\n<div id="gbar"><nobr><b class="gb1">Recherche</b> <a cl ...  
## [2] <div id="gbar"><nobr><b class="gb1">Recherche</b> <a class="gb1" href="h ...  
## [3] <div id="guser" width="100%"><nobr><span id="gbn" class="gbi"></span><sp ...  
## [4] <div class="gbh" style="left:0"></div>\n  
## [5] <div class="gbh" style="right:0"></div>  
## [6] <div id="lga">\n<img alt="Google" height="92" src="/images/branding/goog ...  
## [7] <div class="ds" style="height:32px;margin:4px 0"><input class="lst" styl ...  
## [8] <div id="gac\_scont"></div>\n  
## [9] <div style="font-size:83%;min-height:3.5em">\n<br><div id="gws-output-pa ...  
## [10] <div id="gws-output-pages-elements-homepage\_additional\_languages\_\_als">\ ...  
## [11] <div id="SIvCob">Google disponible en : <a href="http://www.google.com/ ...  
## [12] <div style="font-size:10pt"><div style="margin:19px auto;text-align:cent ...  
## [13] <div style="margin:19px auto;text-align:center" id="fll">\n<a href="/int ...

# From an httr request  
google2 <- read\_html(httr::GET("http://google.com"))

box\_office <- read\_html("http://www.boxofficemojo.com/movies/?id=ateam.htm")  
box\_office %>% html\_node("form") %>% html\_form()

## <form> '<unnamed>' (GET /search)  
## <input search> 'q':   
## <input submit> '':

# CSS selectors ----------------------------------------------  
url <- paste0(  
"https://web.archive.org/web/20190202054736/",  
"https://www.boxofficemojo.com/movies/?id=ateam.htm"  
)  
ateam <- read\_html(url)  
html\_nodes(ateam, "center")

## {xml\_nodeset (1)}  
## [1] <center><table border="0" cellspacing="1" cellpadding="4" bgcolor="#dcdcd ...

html\_nodes(ateam, "center font")

## {xml\_nodeset (1)}  
## [1] <font size="4">Domestic Total Gross: <b>$77,222,099</b></font>

html\_nodes(ateam, "center font b")

## {xml\_nodeset (1)}  
## [1] <b>$77,222,099</b>

# But html\_node is best used in conjunction with %>% from magrittr  
# You can chain subsetting:  
ateam %>% html\_nodes("center") %>% html\_nodes("td")

## {xml\_nodeset (7)}  
## [1] <td align="center" colspan="2"><font size="4">Domestic Total Gross: <b>$7 ...  
## [2] <td valign="top">Distributor: <b><a href="/web/20190202054736/https://www ...  
## [3] <td valign="top">Release Date: <b><nobr><a href="/web/20190202054736/http ...  
## [4] <td valign="top">Genre: <b>Action</b>\n</td>\n  
## [5] <td valign="top">Runtime: <b>1 hrs. 57 min.</b>\n</td>  
## [6] <td valign="top">MPAA Rating: <b>PG-13</b>\n</td>\n  
## [7] <td valign="top">Production Budget: <b>$110 million</b>\n</td>

ateam %>% html\_nodes("center") %>% html\_nodes("font")

## {xml\_nodeset (1)}  
## [1] <font size="4">Domestic Total Gross: <b>$77,222,099</b></font>

td <- ateam %>% html\_nodes("center") %>% html\_nodes("td")  
td

## {xml\_nodeset (7)}  
## [1] <td align="center" colspan="2"><font size="4">Domestic Total Gross: <b>$7 ...  
## [2] <td valign="top">Distributor: <b><a href="/web/20190202054736/https://www ...  
## [3] <td valign="top">Release Date: <b><nobr><a href="/web/20190202054736/http ...  
## [4] <td valign="top">Genre: <b>Action</b>\n</td>\n  
## [5] <td valign="top">Runtime: <b>1 hrs. 57 min.</b>\n</td>  
## [6] <td valign="top">MPAA Rating: <b>PG-13</b>\n</td>\n  
## [7] <td valign="top">Production Budget: <b>$110 million</b>\n</td>

# When applied to a list of nodes, html\_nodes() returns all nodes,  
# collapsing results into a new nodelist.  
td %>% html\_nodes("font")

## {xml\_nodeset (1)}  
## [1] <font size="4">Domestic Total Gross: <b>$77,222,099</b></font>

# html\_node() returns the first matching node. If there are no matching  
# nodes, it returns a "missing" node  
if (utils::packageVersion("xml2") > "0.1.2") {  
td %>% html\_node("font")  
}

## {xml\_nodeset (7)}  
## [1] <font size="4">Domestic Total Gross: <b>$77,222,099</b></font>  
## [2] <NA>  
## [3] <NA>  
## [4] <NA>  
## [5] <NA>  
## [6] <NA>  
## [7] <NA>

# To pick out an element at specified position, use magrittr::extract2  
# which is an alias for [[  
library(magrittr)

## Warning: package 'magrittr' was built under R version 3.6.3

##   
## Attaching package: 'magrittr'

## The following object is masked from 'package:purrr':  
##   
## set\_names

## The following object is masked from 'package:tidyr':  
##   
## extract

ateam %>% html\_nodes("table") %>% extract2(1) %>% html\_nodes("img")

## {xml\_nodeset (6)}  
## [1] <img src="https://web.archive.org/web/20190202054736im\_/https://m.media-a ...  
## [2] <img src="//web.archive.org/web/20190202054736im\_/https://www.assoc-amazo ...  
## [3] <img src="//web.archive.org/web/20190202054736im\_/https://www.assoc-amazo ...  
## [4] <img src="/web/20190202054736im\_/https://www.boxofficemojo.com/img/misc/b ...  
## [5] <img src="/web/20190202054736im\_/https://www.boxofficemojo.com/img/misc/I ...  
## [6] <img src="https://web.archive.org/web/20190202054736im\_/http://b.scorecar ...

ateam %>% html\_nodes("table") %>% `[[`(1) %>% html\_nodes("img")

## {xml\_nodeset (6)}  
## [1] <img src="https://web.archive.org/web/20190202054736im\_/https://m.media-a ...  
## [2] <img src="//web.archive.org/web/20190202054736im\_/https://www.assoc-amazo ...  
## [3] <img src="//web.archive.org/web/20190202054736im\_/https://www.assoc-amazo ...  
## [4] <img src="/web/20190202054736im\_/https://www.boxofficemojo.com/img/misc/b ...  
## [5] <img src="/web/20190202054736im\_/https://www.boxofficemojo.com/img/misc/I ...  
## [6] <img src="https://web.archive.org/web/20190202054736im\_/http://b.scorecar ...

# Find all images contained in the first two tables  
ateam %>% html\_nodes("table") %>% `[`(1:2) %>% html\_nodes("img")

## {xml\_nodeset (6)}  
## [1] <img src="https://web.archive.org/web/20190202054736im\_/https://m.media-a ...  
## [2] <img src="//web.archive.org/web/20190202054736im\_/https://www.assoc-amazo ...  
## [3] <img src="//web.archive.org/web/20190202054736im\_/https://www.assoc-amazo ...  
## [4] <img src="/web/20190202054736im\_/https://www.boxofficemojo.com/img/misc/b ...  
## [5] <img src="/web/20190202054736im\_/https://www.boxofficemojo.com/img/misc/I ...  
## [6] <img src="https://web.archive.org/web/20190202054736im\_/http://b.scorecar ...

ateam %>% html\_nodes("table") %>% extract(1:2) %>% html\_nodes("img")

## {xml\_nodeset (6)}  
## [1] <img src="https://web.archive.org/web/20190202054736im\_/https://m.media-a ...  
## [2] <img src="//web.archive.org/web/20190202054736im\_/https://www.assoc-amazo ...  
## [3] <img src="//web.archive.org/web/20190202054736im\_/https://www.assoc-amazo ...  
## [4] <img src="/web/20190202054736im\_/https://www.boxofficemojo.com/img/misc/b ...  
## [5] <img src="/web/20190202054736im\_/https://www.boxofficemojo.com/img/misc/I ...  
## [6] <img src="https://web.archive.org/web/20190202054736im\_/http://b.scorecar ...

# XPath selectors ---------------------------------------------  
# chaining with XPath is a little trickier - you may need to vary  
# the prefix you're using - // always selects from the root node  
# regardless of where you currently are in the doc  
ateam %>%  
html\_nodes(xpath = "//center//font//b") %>%  
html\_nodes(xpath = "//b")

## {xml\_nodeset (21)}  
## [1] <b>Adjuster:</b>  
## [2] <b>The A-Team</b>  
## [3] <b>$77,222,099</b>  
## [4] <b><a href="/web/20190202054736/https://www.boxofficemojo.com/studio/cha ...  
## [5] <b><nobr><a href="/web/20190202054736/https://www.boxofficemojo.com/sche ...  
## [6] <b>Action</b>  
## [7] <b>1 hrs. 57 min.</b>  
## [8] <b>PG-13</b>  
## [9] <b>$110 million</b>  
## [10] <b>Domestic:</b>  
## [11] <b>$77,222,099</b>  
## [12] <b>43.6%</b>  
## [13] <b>Worldwide:</b>  
## [14] <b>$177,238,796</b>  
## [15] <b>&gt; View All 14 Weekends</b>  
## [16] <b>Showdown: 'Men-on-a-Mission'</b>  
## [17] <b>4</b>  
## [18] <b>Chart</b>  
## [19] <b>Rank</b>  
## [20] <b>Charts (Premier Pass Users Only)</b>  
## ...

sample1 <- minimal\_html("<table>  
<tr><th>Col A</th><th>Col B</th></tr>  
<tr><td>1</td><td>x</td></tr>  
<tr><td>4</td><td>y</td></tr>  
<tr><td>10</td><td>z</td></tr>  
</table>")  
sample1 %>%  
html\_node("table") %>%  
html\_table()

## Col A Col B  
## 1 1 x  
## 2 4 y  
## 3 10 z

# Values in merged cells will be duplicated  
sample2 <- minimal\_html("<table>  
<tr><th>A</th><th>B</th><th>C</th></tr>  
<tr><td>1</td><td>2</td><td>3</td></tr>  
<tr><td colspan='2'>4</td><td>5</td></tr>  
<tr><td>6</td><td colspan='2'>7</td></tr>  
</table>")  
sample2 %>%  
html\_node("table") %>%  
html\_table()

## A B C  
## 1 1 2 3  
## 2 4 4 5  
## 3 6 7 7

# If the table is badly formed, and has different number of columns  
# in each row, use `fill = TRUE` to fill in the missing values  
sample3 <- minimal\_html("<table>  
<tr><th>A</th><th>B</th><th>C</th></tr>  
<tr><td colspan='2'>1</td><td>2</td></tr>  
<tr><td colspan='2'>3</td></tr>  
<tr><td>4</td></tr>  
</table>")  
sample3 %>%  
html\_node("table") %>%  
html\_table(fill = TRUE)

## A B C  
## 1 1 1 2  
## 2 3 3 NA  
## 3 4 NA NA