Assignment 8

Vaishnavi Venkatesh

July 10, 2017

```
#question1
library(rvest)

## Loading required package: xml2

library(stringr)
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(purrr)

##
## Attaching package: 'purrr'

## The following objects are masked from 'package:dplyr':
##
##     contains, order_by

page <- read_html("https://www.yelp.com/search?find_desc=burgers&start=0&l=Bo
ston,MA")
# list the children of the <html> element (the whole page)
html_children(page)

## {xml_nodeset (2)}
## [1] <head>\n<script>            window.yPageStart = new Date().getTime() .
..
## [2] <body id="yelp_main_body" class="jquery country-us logged-out">\n\n  .
..

## {xml_nodeset (2)}
## [1] <head>\n<script> window.yPageStart = new Date().getTime() ...
## [2] <body id="yelp_main_body" class="jquery country-us logged-out">\n\n ..
.
# get the root of the actual html body
```

```
root <- html_node(page, 'body')

children <- html_children(html_children(page)[2])
children

## {xml_nodeset (28)}
##  [1] <script>(function (d, w) {\n    var supportsSVG = (\n        !!d.cr .
..
##  [2] <noscript>\n    <link rel="stylesheet" href="https://s3-media2.fl.y .
..
##  [3] <div id="fb-root"></div>
##  [4] <div id="wrap" class="lang-en">\n            <div class="page-heade .
..
##  [5] <script>            yConfig = {"bingMapsUrl": "https://www.bing.com .
..
##  [6] <noscript><img src="https://sb.scorecardresearch.com/p?cj=1&amp;c15 .
..
##  [7] <script>\n            (function() {\n                var main = nul .
..
##  [8] <noscript>\n        <img style="display: none;" src="https://pixel. .
..
##  [9] <script src="https://s3-media2.fl.yelpcdn.com/assets/2/www/js/a59b2 .
..
## [10] <script src="//ajax.googleapis.com/ajax/libs/jquery/1.8.3/jquery.mi .
..
## [11] <script>if(document.readyState === 'interactive') jQuery.ready();\n .
..
## [12] <script src="https://s3-media1.fl.yelpcdn.com/assets/2/www/js/ca292 .
..
## [13] <script src="https://s3-media2.fl.yelpcdn.com/assets/2/www/js/dd4e2 .
..
## [14] <script src="https://s3-media4.fl.yelpcdn.com/assets/2/www/js/07523 .
..
## [15] <script>\n                yConfig.vendorExternalURLs["plugin-detect .
..
## [16] <script src="https://s3-media1.fl.yelpcdn.com/assets/2/www/js/f45e7 .
..
## [17] <script>yelp.www.init.search.Controller({"adVisibilityURI": "/ad_vi .
..
## [18] <script src="//sb.scorecardresearch.com/beacon.js"></script>
## [19] <script>\n                if (window.COMSCORE && window.COMSCORE['b .
..
## [20] <script>\n            (function() {\n                var main = nul .
..
## ...

name<- html_name(children)
name
```

```
##  [1] "script"   "noscript" "div"      "div"      "script"   "noscript"
##  [7] "script"   "noscript" "script"   "script"   "script"   "script"
## [13] "script"   "script"   "script"   "script"   "script"   "script"
## [19] "script"   "script"   "script"   "script"   "noscript" "script"
## [25] "div"      "noscript" "script"   "noscript"
```

```
length(children)
```

```
## [1] 28
```

```
html_attr(children, 'id')
```

```
##  [1] NA
##  [2] NA
##  [3] "fb-root"
##  [4] "wrap"
##  [5] NA
##  [6] NA
##  [7] NA
##  [8] NA
##  [9] NA
## [10] NA
## [11] NA
## [12] NA
## [13] NA
## [14] NA
## [15] NA
## [16] NA
## [17] NA
## [18] NA
## [19] NA
## [20] NA
## [21] NA
## [22] NA
## [23] NA
## [24] NA
## [25] "ttdUniversalPixelTag290e816a69e9439f960a9588bc2ffb54"
## [26] NA
## [27] NA
## [28] NA
```

```
#The css selector to select restaurants that are advertisements
#.yloca-search-result:nth-child(1) span

#question2
library(rvest)
library(stringr)
library(dplyr)
library(purrr)
get_yelp_sr_one_page <- function(keyword, loc, page =x) {
  yelp_url <- 'https://www.yelp.com/search?find_desc=%s&find_loc=%s&start=%s'
```

```
  page1 = (page-1)*10;
  yelp_url <- sprintf(yelp_url, URLencode(keyword), URLencode(loc), page1)
  yelpsr <- read_html(yelp_url)
  items <- yelpsr %>%
    html_nodes("li.regular-search-result")
  links <- items %>% html_nodes("a.biz-name")
  names <- links %>% html_text(trim=T)
  urls <- links %>%
    html_attr("href") %>%
    str_replace("\\?osq=.*", "")
  pricelevels <- items %>%
    html_nodes(".business-attribute.price-range") %>%
    html_text(trim=T) %>%
    str_count()
  secondary_attrs <- items %>%
    html_nodes('.secondary-attributes') %>%
    purrr::map(function(item) {
      tibble(
        neighborhood = item %>%
          html_node('.neighborhood-str-list') %>%
          html_text(trim=T),
        address = item %>%
          html_node('address') %>%
          html_text(trim=T),
        phone = item %>%
          html_node('.biz-phone') %>%
          html_text(trim=T)
      )
    }) %>%
    bind_rows()
  tibble(
    name = names,
    url = urls,
    price = pricelevels
  ) %>%
    cbind(secondary_attrs)
}
get_yelp_sr_one_page("Burgers", "Boston, MA", page = 1) %>%
  head(5) %>%
  select(name, url, price, phone, address) %>%
  knitr::kable()
```

name	url	price	phone	address
Boston Burger Company - Boston	/biz/boston-burger-company-boston-boston	2	(857) 233-4560	1100 Boylston StBoston, MA 02215
Wheelhouse	/biz/wheelhouse-boston-3	1	(617) 422-0082	63 Broad StBoston, MA 02109

Gate Bar & Resturant	/biz/gate-bar-and-resturant-boston	2	(617) 942-7262	3171 Washington StBoston, MA 02130
Tasty Burger	/biz/tasty-burger-boston	1	(617) 425-4444	1301 Boylston StBoston, MA 02215
The Gallows	/biz/the-gallows-boston	2	(617) 425-0200	1395 Washington StBoston, MA 02118

```
#question3
library(rvest)
library(stringr)
library(dplyr)
library(purrr)
get_yelp_sr <- function(keyword, loc, page =x) {
  yelp_url <- 'https://www.yelp.com/search?find_desc=%s&find_loc=%s&start=%s'
  page1 = (page-1)*10;
  yelp_url <- sprintf(yelp_url, URLencode(keyword), URLencode(loc), page1)
  yelpsr <- read_html(yelp_url)
  items <- yelpsr %>%
    html_nodes("li.regular-search-result")
  links <- items %>% html_nodes("a.biz-name")
  names <- links %>% html_text(trim=T)
  urls <- links %>%
    html_attr("href") %>%
    str_replace("\\?osq=.*", "")
  pricelevels <- items %>%
    html_nodes(".business-attribute.price-range") %>%
    html_text(trim=T) %>%
    str_count()
  secondary_attrs <- items %>%
    html_nodes('.secondary-attributes') %>%
    purrr::map(function(item) {
      tibble(
        neighborhood = item %>%
          html_node('.neighborhood-str-list') %>%
          html_text(trim=T),
        address = item %>%
          html_node('address') %>%
          html_text(trim=T),
        phone = item %>%
          html_node('.biz-phone') %>%
          html_text(trim=T)
      )
    }) %>%
    bind_rows()
  tibble(
    name = names,
```

```
    url = urls,
    price = pricelevels
  ) %>%
    cbind(secondary_attrs)
}
x <- NULL
page = 10
for( i in 1:page){
  final<- get_yelp_sr("Burgers", "Boston, MA", page =i)
 x <-  rbind(x, final)
}

#question 4
library(rvest)
library(stringr)
library(dplyr)
library(purrr)
get_yelp_sr <- function(keyword, loc, page =x) {
  yelp_url <- 'https://www.yelp.com/search?find_desc=%s&find_loc=%s&start=%s'
  page1 = (page-1)*10;
  yelp_url <- sprintf(yelp_url, URLencode(keyword), URLencode(loc), page1)
  yelpsr <- read_html(yelp_url)
  items <- yelpsr %>%
    html_nodes("li.regular-search-result")
  links <- items %>% html_nodes("a.biz-name")
  names <- links %>% html_text(trim=T)
  urls <- links %>%
    html_attr("href") %>%
    str_replace("\\?osq=.*", "")
  pricelevels <- items %>%
    html_nodes(".business-attribute.price-range") %>%
    html_text(trim=T) %>%
    str_count()
  secondary_attrs <- items %>%
    html_nodes('.secondary-attributes') %>%
    purrr::map(function(item) {
      tibble(
        neighborhood = item %>%
          html_node('.neighborhood-str-list') %>%
          html_text(trim=T),
        address = item %>%
          html_node('address') %>%
          html_text(trim=T),
        phone = item %>%
          html_node('.biz-phone') %>%
          html_text(trim=T)
      )
    }) %>%
    bind_rows()
  tibble(
```

```
    name = names,
    url = urls,
    price = pricelevels
  ) %>%
    cbind(secondary_attrs)
}
x <- NULL
page = 10
for( i in 1:page){
  Sys.sleep(1)
  final<- get_yelp_sr("Burgers", "Boston, MA", page =i)
 x <-  rbind(x, final)
}
```