Construindo bases de dados via Webscraping

Benilton Carvalho & Guilherme Ludwig

Webscraping

- É possível "raspar" (*scrape*) informações de páginas da internet, e guardálas em um banco de dados. A prática é chamada de *webscraping*.
- Nós iremos usar o pacote rvest, que simplifica algumas operações dos pacotes xml2 e httr.

```
library(tidyverse)
library(RSQLite)
library(httr)
library(rvest)
```

Idéias

Uma página da web é um documento que pode ser exibido por um navegador. Estes documentos normalmente exibem resultados de consultas à bancos de dados, que são nosso principal interesse nesta disciplina. Em geral:

- Páginas simples podem ser acessadas através do R com o pacote rvest.
- Páginas dinâmicas podem exigir alguma autenticação do usuário, na forma de *cookies*. Para acessar essas páginas, podemos precisar do pacote httr.
- Nosso objetivo é coletar dados com o rvest (e talvez httr) e armazenálos em um banco de dados.

Alguns recursos:

- http://material.curso-r.com/scrape/: Material organizado pelo pessoal do Curso-R sobre webscraping (com mais exemplos).
- https://blog.rstudio.com/2014/11/24/rvest-easy-web-scraping-with-r/: Tutorial do Rvest (pode estar desatualizado).

HTML

Inevitavelmente, vocês deverão ter alguma idéia de HTML (pelo menos como funciona). Em geral, páginas html são texto estruturado, interpretado pelo navegador. Veja exemplos em:

https://www.w3schools.com/html/html_basic.asp

```
<!DOCTYPE html>
<html>
<body>
<h1>My First Heading</h1>
My first paragraph.
</body>
</html>
```

Note: "body" na linguagem do rvest é um node. Dentro desse node, há um node tipo h1 (cabeçalho) e outro tipo p (parágrafo). Nodes interessantes incluem table, a (links) e img.

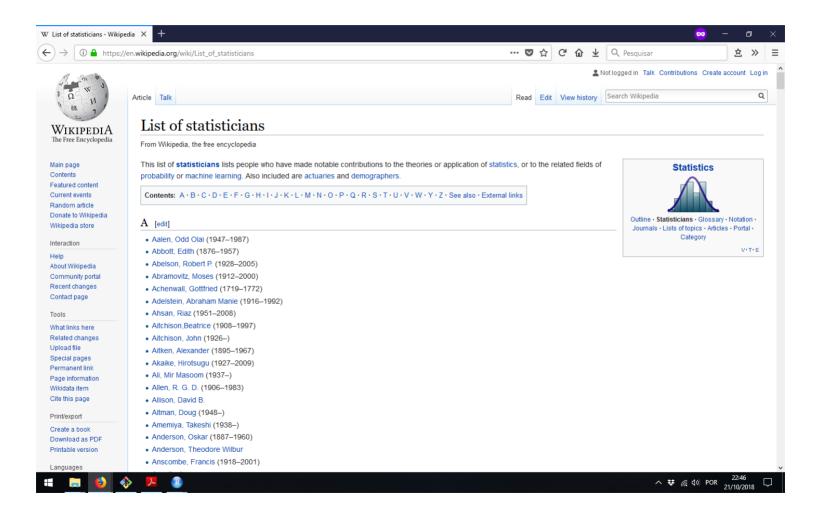
Exemplo: wikipedia

A *wikipedia* é particularmente interessante para scraping, pois ela possui muitas páginas com listas, de onde podemos começar nossas buscas. Por exemplo,

https://en.wikipedia.org/wiki/List_of_statisticians

Podemos estar interessados em compilar uma lista com nome, *alma mater*, data de nascimento (e local), e data de falecimento (caso já tenha falecido) de estatísticos famosos.

Lista de Estatísticos



Página: George Box

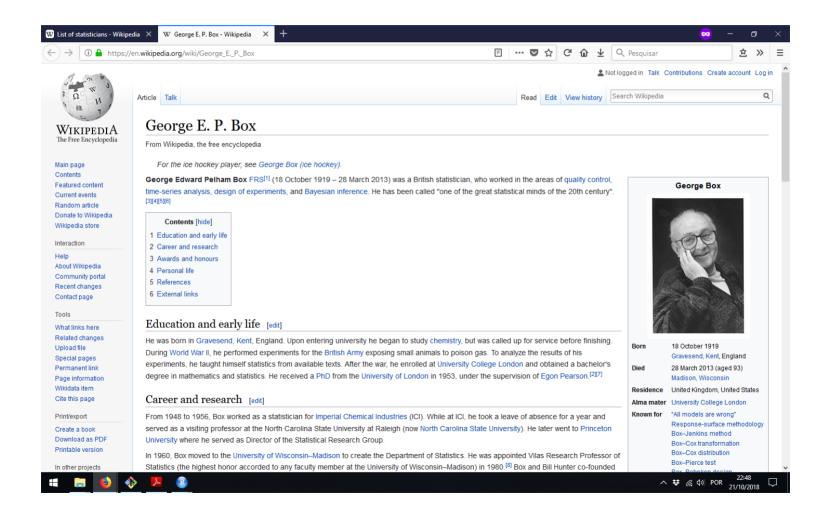
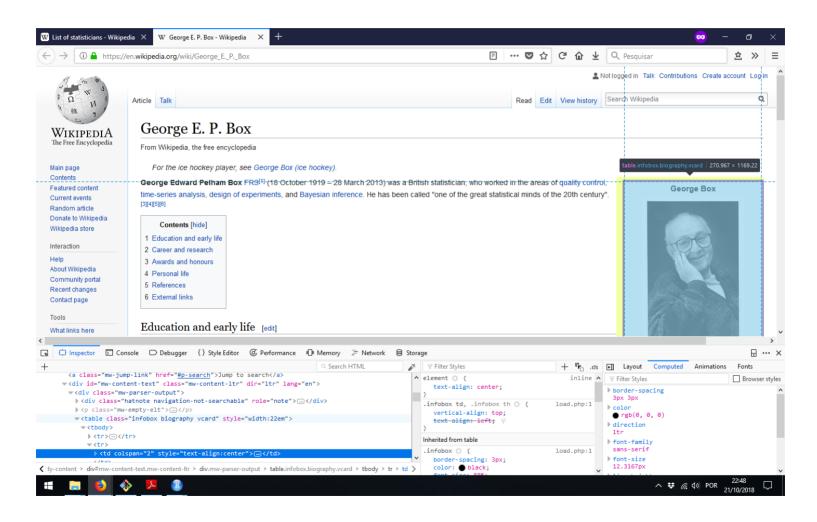


Tabela de Interesse



SelectorGadget

Uma ferramenta recomendada pelo rvest é o chamado SelectorGadget (https://selectorgadget.com/), que mostra o nome de um "selector" em CSS. Há uma extensão para o navegador Chrome que permite que você use o SelectorGadget em qualquer página.

Com o selector correto, você pode acessá-lo usando html_nodes(). Selectors interessantes incluem "tables.<nome>" e "li". É preciso inspecionar as páginas de interesse caso a caso.

Usando SelectorGadget (Chrome)

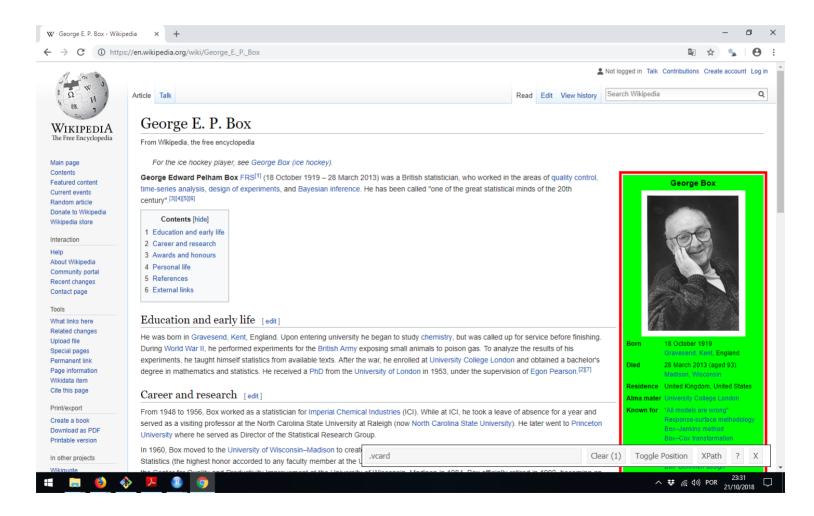


Tabela de interesse

```
url = "https://en.wikipedia.org/wiki/George_E._P._Box"
webpage <- read_html(url)

table <- webpage %>%
html_nodes("table.vcard") %>% # SG
html_table(header = FALSE)

# Retorna uma lista de tabelas... eu só quero a primeira
table <- table[[1]]</pre>
```

Conteúdo

```
table %>% as.tibble
## Warning: `as.tibble()` is deprecated, use `as_tibble()` (but mind the new
## This warning is displayed once per session.
## # A tibble: 17 x 2
##
     X1
                      X2
                      <chr>
##
      <chr>
##
    1 George Box
                      George Box
   2 ""
##
                      (1919-10-18)18 October 1919Gravesend, Kent, England
##
   3 Born
##
   4 Died
                      28 March 2013(2013-03-28) (aged 93)Madison, Wiscor
   5 Residence
                      United Kingdom, United States
##
##
   6 Alma mater
                           University College London
   7 Known  for
                           ""All models are wrong"\nResponse-surface methodol
##
   8 Awards
                      "Shewhart Medal (1968)\nWilks Memorial Award (1972)\nR.
##
   9 Scientific car... Scientific career
## 10 Fields
                      "Statistics\nDesign of experiments\nBayesian statistics
## 11 Institutions
                      "ICI\nPrinceton University\nUniversity of Wisconsin-Mad
## 12 Thesis
                      Departures from Independence and Homoskedasticity in th
## 13 Doctoral advis... "Egon Pearson\nH. 0. Hartley[2]"
## 14 Doctoral stude... John F. MacGregor[2]Greta M. Ljung
                                                                     12 / 24
```

Conteúdo (limpeza com regex)

```
## # A tibble: 17 x 2
##
      X1
                      X2
##
      <chr>
                      <chr>>
                      George Box
##
    1 George Box
    2 ""
##
##
   3 Born
                      (1919-10-18)18 October 1919Gravesend, Kent, England
   4 Died
                      28 March 2013(2013-03-28) (aged 93) Madison, Wisconsin
##
## 5 Residence
                      United Kingdom, United States
## 6 Alma mater
                      University College London
## 7 Known for
                      "All models are wrong" Response-surface methodology EVO
## 8 Awards
                      Shewhart Medal (1968) Wilks Memorial Award (1972) R. A.
   9 Scientific car... Scientific career
## 10 Fields
                      Statistics Design of experiments Bayesian statistics Ti
## 11 Institutions
                      ICI Princeton University University of Wisconsin-Madiso
## 12 Thesis
                      Departures from Independence and Homoskedasticity in the
## 13 Doctoral advis... "Egon Pearson H. O. Hartley "
## 14 Doctoral stude... John F. MacGregor Greta M. Ljung
## 15 Influences
                      Ronald Fisher
## 16 Influenced
                      Norman Draper George C. Tiao
## 17 ""
                       11 11
```

Procurando Links

Inspecionando a página no navegador, eu descobri que dentro de body #content (o conteúdo da página) os links estão guardados no node "li".

```
url = "https://en.wikipedia.org/wiki/List_of_statisticians"
listPages <- read_html(url)
links <- listPages %>%
  html_nodes("body #content") %>% # Inspect Object...
html_nodes("li") # All links
```

Procurando Links

links

```
## {xml nodeset (691)}
   [1] <a href="/wiki/Outline_of_statistics" title="Outline of statist .
##
   [2] <a class="mw-selflink selflink">Statisticians</a>
##
   [3] <a href="/wiki/Glossary_of_probability_and_statistics" title="G .
##
   [4] <a href="/wiki/Notation_in_probability_and_statistics" title="N .
##
##
   [5] <a href="/wiki/List_of_statistics_journals" title="List of stat .
   [6] <a href="/wiki/Lists_of_statistics_topics" title="Lists of stat .
##
   [7] <a href="/wiki/List_of_statistics_articles" title="List of stat .
##
   [8] \n<a href="/wiki/File:Nuvola_apps_edu_mathematics_blue-p.svg" c .
##
   [9] <a href="/wiki/Category:Statistics" title="Category:Statistics" .
##
## [10] <a href="/wiki/Template:Statistics_topics_sideb .
## [11] <a href="/wiki/Template_talk:Statistics_topics_ .
## [12] <a class="external text" href="https://en.wikip .
## [13] <a href="#A">A</a>
## [14] <a href="#B">B</a>
## [15] <a href="#C">C</a>
## [16] <a href="#D">D</a>
## [17] <a href="#E">E</a>
## [18] <a href="#F">F</a>
## [19] <a href="#G">G</a>
                                                             15 / 24
```

"Sajid Ali Khan, Rawalakot" até "Zipf, George Kingsley"

```
estat1 = links %>%
  as.character %>%
  grep("Sajid Ali Khan, Rawalakot", .)
estatN = links %>%
  as.character %>%
  grep("Zipf, George Kingsley", .)
estat1
## [1] 40
estatN
## [1] 679
links <- links[estat1:estatN]</pre>
```

Páginas individuais

links

```
## {xml nodeset (640)}
   [1] \n<a href="/w/index.php?title=Sajid_Ali_Khan&amp;action=edit&am .
##
   [2] \n<a href="/wiki/Odd Aalen" title="Odd Aalen">Aalen, Odd Olai</ .
##
##
   [3] \n<a href="/wiki/Edith Abbott" title="Edith Abbott">Abbott, Edi .
   [4] \n<a href="/wiki/Robert_P._Abelson" class="mw-redirect" title=" .
##
##
   [5] \n<a href="/wiki/Moses Abramovitz" title="Moses Abramovitz">Abr .
   [6] \n<a href="/wiki/Gottfried_Achenwall" title="Gottfried Achenwal .
##
   [7] \n<a href="/wiki/Abraham Manie Adelstein" title="Abraham Manie .
##
   [8] \n<a href="/wiki/Riaz_Ahsan" title="Riaz Ahsan">Ahsan, Riaz</a> .
##
   [9] \n<a href="/wiki/Beatrice Aitchison" title="Beatrice Aitchison".
##
  [10] \n<a href="/wiki/John_Aitchison" title="John Aitchison">Aitchis .
## [11] \n<a href="/wiki/Alexander Aitken" title="Alexander Aitken">Ait .
  [12] \n<a href="/wiki/Hirotsugu_Akaike" class="mw-redirect" title="H .
  [13] \n<a href="/wiki/Mir Masoom Ali" title="Mir Masoom Ali">Ali, Mi .
## [14] \n<a href="/wiki/R._G._D._Allen" title="R. G. D. Allen">Allen, .
## [15] <a href="/wiki/David_B._Allison" title="David B. Allison">Allis .
  [16] \n<a href="/wiki/Doug_Altman" title="Doug Altman">Altman, Doug< .
  [17] \n<a href="/wiki/Takeshi_Amemiya" title="Takeshi Amemiya">Amemi .
## [18] \n<a href="/wiki/Oskar_Anderson" title="Oskar Anderson">Anderso .
## [19] <a href="/wiki/Theodore_Wilbur_Anderson" title="Theodore Wilbur .
```

Páginas individuais

```
links %>%
  html nodes("a")
## {xml nodeset (640)}
    [1] <a href="/w/index.php?title=Sajid_Ali_Khan&amp;action=edit&amp;redl .
##
    [2] <a href="/wiki/Odd Aalen" title="Odd Aalen">Aalen, Odd Olai</a>
##
    [3] <a href="/wiki/Edith_Abbott" title="Edith Abbott">Abbott">Abbott, Edith</a>
##
##
    [4] <a href="/wiki/Robert P. Abelson" class="mw-redirect" title="Robert .
    [5] <a href="/wiki/Moses_Abramovitz" title="Moses Abramovitz">Abramovit .
##
    [6] <a href="/wiki/Gottfried Achenwall" title="Gottfried Achenwall">Ach .
##
    [7] <a href="/wiki/Abraham_Manie_Adelstein" title="Abraham Manie Adelst .
##
    [8] <a href="/wiki/Riaz Ahsan" title="Riaz Ahsan">Ahsan, Riaz</a>
##
    [9] <a href="/wiki/Beatrice_Aitchison" title="Beatrice Aitchison">Aitch .
##
## [10] <a href="/wiki/John Aitchison" title="John Aitchison">Aitchison, Jo .
   [11] <a href="/wiki/Alexander_Aitken" title="Alexander Aitken">Aitken, A .
  [12] <a href="/wiki/Hirotsugu_Akaike" class="mw-redirect" title="Hirotsu .
## [13] <a href="/wiki/Mir_Masoom_Ali" title="Mir Masoom Ali">Ali, Mir Maso .
## [14] <a href="/wiki/R._G._D._Allen" title="R. G. D. Allen">Allen, R. G. .
## [15] <a href="/wiki/David_B._Allison" title="David B. Allison">Allison,
## [16] <a href="/wiki/Doug_Altman" title="Doug Altman">Altman, Doug</a>
## [17] <a href="/wiki/Takeshi_Amemiya" title="Takeshi Amemiya">Amemiya, Ta .
## [18] <a href="/wiki/Oskar_Anderson" title="Oskar Anderson">Anderson'\& Os4k .
```

Páginas Individuais

```
links %>%
  html nodes("a") %>%
  html attr("href") # Salvar title também!
     [1] "/w/index.php?title=Sajid_Ali_Khan&action=edit&redlink=1"
##
##
     [2] "/wiki/Odd Aalen"
     [3] "/wiki/Edith Abbott"
##
##
     [4] "/wiki/Robert P. Abelson"
    [5] "/wiki/Moses Abramovitz"
##
    [6] "/wiki/Gottfried Achenwall"
##
    [7] "/wiki/Abraham_Manie_Adelstein"
##
##
    [8] "/wiki/Riaz Ahsan"
     [9] "/wiki/Beatrice Aitchison"
##
##
    [10] "/wiki/John Aitchison"
    [11] "/wiki/Alexander_Aitken"
##
    [12] "/wiki/Hirotsugu Akaike"
##
    [13] "/wiki/Mir_Masoom_Ali"
##
    [14] "/wiki/R._G._D._Allen"
##
    [15] "/wiki/David_B._Allison"
##
    [16] "/wiki/Doug_Altman"
##
##
    [17] "/wiki/Takeshi_Amemiya"
##
    [18] "/wiki/Oskar_Anderson"
```

DB

A melhor maneira de armazenar o conteúdo das páginas é através de um banco de dados.

```
li <- links %>% html nodes("a") %>% html attr("href")
li <- paste0("https://en.wikipedia.org", li)</pre>
names <- links %>% html_nodes("a") %>% html attr("title")
db = dbConnect(SQLite(), "estatisticos.db")
bad = c("page does not exist", "Florence Nightingale", "Harold Wilson
bad1 = unlist(sapply(bad, grep, names))
bad2 = unlist(sapply(c("mshkhan", "redlink", "orghttp"), grep, li))
names = names[-c(bad1, bad2)]
li = li[-c(bad1, bad2)]
dbWriteTable(db, "person", data.frame(id = seq_along(names),
                                       names = names,
                                       links = li))
dbExecute(db, "CREATE TABLE info
               (id INTEGER, Born TEXT, Died TEXT,
                AlmaMater TEXT)")
```

[1] 0

Conferindo...

Moses Abramovitz

3

4

3 Robert P. Abelson https://en.wikipedia.org/wiki/Robert_P._Abelson

https://en.wikipedia.org/wiki/Moses_Abramovitz

Extraindo tabelas (demora alguns minutos...)

```
library(doMC) ## se windows library(doParallel)
## Loading required package: foreach
##
## Attaching package: 'foreach'
## The following objects are masked from 'package:purrr':
##
##
       accumulate, when
## Loading required package: iterators
## Loading required package: parallel
registerDoMC(4) ## se windows registerDoParallel(nproc)
f <- function(x) if(length(x) == 0) return(NA_character_) else return</pre>
out = foreach(i=seq_along(li), .combine=rbind) %dopar% {
  webpage <- read_html(li[i])</pre>
  table <- webpage %>% html_nodes("table.vcard") %>% html_table(head
  if(length(table) == 0) return(NULL)
```

Consultando nossa tabela

O código quebrou com *Florence Nightingale* e *Harold Wilson*, porque tinham muitos campos diferentes do esperado, ficando inconsistente com as tabelas anteriores. Estes podem ser trabalhados manualmente. Fora isso, o código de scraping funciona bem:

```
## # A tibble: 8 x 4
##
                                    Died
                                                        AlmaMater
     names
                Born
     <chr> <chr>
                                    <chr>
                                                        <chr>
##
## 1 Odd Aalen (1947-05-06) May ... <NA>
                                                        University of Oslo
## 2 Edith Ab... (1876-09-26) Septe... July 28, 1957(195... <NA>
## 3 Moses Ab... (1912-01-01) Janua... December 1, 2000 (... Harvard University and C
## 4 Gottfrie... (1719-10-20)20 Oc... 1 May 1772(1772-0... <NA>
## 5 Riaz Ahs... 1951, December 25... November 8, 2008(... University of KarachiAda
## 6 Beatrice... (1908-07-18)18 Ju... 22 September 1997... Goucher College (BA)Johr
## 7 John Ait... 22 July 1926East ... 23 December 2016 ... University of Edinburgh
## 8 Alexande... (1895-04-01)1 Apr... 3 November 1967(1... University of Edinburghl
```

Crawlers + Cron

- *Crawlers* são programas de scraping que interpretam conexões nas páginas e continuam buscando novas páginas. Estão bem além do escopo do nosso curso.
- cron é um programa de Linux que permite executar outros programas em intervalos fixos de tempo. Existe um pacote, cronR (https://cran.r-project.org/web/packages/cronR/README.html), que permite executar scripts de R em períodos fixos de tempo. Em conjunto com uma base de dados, é possível coletar dados periodicamente (por exemplo, com o tweetR ou em páginas de notícias).

Fechando base de dados...

dbDisconnect(db)