Exercise 3 (Web Scraping)

Georg Pretzner (wi19b013)

26 9 2021

Contents

1 Task

Choose a web site of your choice that contains tabular data worth being harvested. You could look at sites like Wikipedia, or official sites like Statistik Austria offering information. Retrieve the data using tidiyverse and rvest packages so it becomes a "tidy" tibble. If the data is in "wide" format, transform into long format. If it is in long format, transform into wide format. Document your steps in a Notebook.

2 Datamanagement

2.1 Required packages

```
library(tidyverse)
library(magrittr)
library(rvest)
```

2.2 WebScraping

The "Details table" from the website `https://en.wikipedia.org/wiki/List_of_Wikipedias` is scraped and then converted into the correct format.

The page source is written to the variable `wiki`.

```
wiki <- read_html("https://en.wikipedia.org/wiki/List_of_Wikipedias")
```

Brief overview of the page source:

```
wiki
```

```
## {html_document}
## <html class="client-nojs" lang="en" dir="ltr">
## [1] <head>\n<meta http-equiv="Content-Type" content="text/html; charset=UTF-8 ...
## [2] <body class="mediawiki ltr sitedir-ltr mw-hide-empty-elt ns-0 ns-subject  ...
```

```
html_nodes(wiki, "*")
```

```
## {xml_nodeset (14970)}
##  [1] <head>\n<meta http-equiv="Content-Type" content="text/html; charset=UTF- ...
##  [2] <meta http-equiv="Content-Type" content="text/html; charset=UTF-8">\n
##  [3] <meta charset="UTF-8">\n
##  [4] <title>List of Wikipedias - Wikipedia</title>\n
##  [5] <script>document.documentElement.className="client-js";RLCONF={"wgBreakF ...
##  [6] <script>(RLQ=window.RLQ||[]).push(function(){mw.loader.implement("user.o ...
##  [7] <link rel="stylesheet" href="/w/load.php?lang=en&amp;modules=ext.cite.st ...
##  [8] <script async="" src="/w/load.php?lang=en&amp;modules=startup&amp;only=s ...
##  [9] <meta name="ResourceLoaderDynamicStyles" content="">\n
## [10] <link rel="stylesheet" href="/w/load.php?lang=en&amp;modules=site.styles ...
## [11] <meta name="generator" content="MediaWiki 1.38.0-wmf.1">\n
## [12] <meta name="referrer" content="origin">\n
## [13] <meta name="referrer" content="origin-when-crossorigin">\n
## [14] <meta name="referrer" content="origin-when-cross-origin">\n
## [15] <meta name="format-detection" content="telephone=no">\n
## [16] <meta property="og:image" content="https://upload.wikimedia.org/wikipedi ...
## [17] <meta property="og:title" content="List of Wikipedias - Wikipedia">\n
## [18] <meta property="og:type" content="website">\n
## [19] <link rel="preconnect" href="//upload.wikimedia.org">\n
## [20] <link rel="alternate" media="only screen and (max-width: 720px)" href="/ ...
## ...
```

3 Tiding

The source of the website is filtered to tables and the third table is selected. Furthermore,

```
tmp <-
  wiki %>%
  html_nodes("table") %>%
  nth(3) %>%
  html_table()
tmp
```

```
## # A tibble: 323 x 11
##    Language `Language (local)` Wiki  Articles  `Total pages` Edits  Admins Users
##    <chr>    <chr>              <chr> <chr>     <chr>         <chr>  <chr>  <chr>
##  1 English  English            en    6,383,962 54,265,386    1,042~ 1,082  42,2~
##  2 Cebuano  Cebuano            ceb   5,950,573 10,704,160    33,10~ 6      82,8~
##  3 Swedish  svenska            sv    2,912,535 6,745,382     49,59~ 64     791,~
##  4 German   Deutsch            de    2,618,057 7,250,735     214,5~ 189    3,78~
##  5 French   français           fr    2,363,066 11,542,213    186,2~ 157    4,19~
##  6 Dutch    Nederlands         nl    2,067,480 4,375,041     59,83~ 37     1,16~
##  7 Russian  <U+0440><U+0443><U+0441><U+0441><U+043A><U+0438><U+0439>            ru    1,757,033 6,69
##  8 Italian  italiano           it    1,718,551 7,208,670     122,8~ 114    2,16~
##  9 Spanish  español            es    1,717,541 7,502,291     138,1~ 65     6,33~
## 10 Polish   polski             pl    1,490,809 3,442,880     64,45~ 105    1,12~
## # ... with 313 more rows, and 3 more variables: Active users <chr>,
## #   Images <chr>, Depth <chr>
```

The columns `Language`, `Articles` and `Users` are extracted from the Details Table.

```
NrArticlesCountry <- tmp %>% select(1,4,8)  %>% mutate_all(str_replace_all, ",", "")
NrArticlesCountry
```

```
## # A tibble: 323 x 3
##    Language Articles Users
##    <chr>    <chr>    <chr>
##  1 English  6383962  42280446
##  2 Cebuano  5950573  82807
##  3 Swedish  2912535  791019
##  4 German   2618057  3780580
##  5 French   2363066  4191659
##  6 Dutch    2067480  1165874
##  7 Russian  1757033  3041819
##  8 Italian  1718551  2168354
##  9 Spanish  1717541  6338732
## 10 Polish   1490809  1124924
## # ... with 313 more rows
```

4 Transformation

Convert char fields with numeric value to integer:

```
NrArticlesCountry %<>%
  mutate(
    Articles = parse_integer(Articles),
    Users = parse_integer(Users)
  )
NrArticlesCountry
```

```
## # A tibble: 323 x 3
##    Language Articles    Users
##    <chr>       <int>    <int>
##  1 English   6383962 42280446
```

```
##  2 Cebuano   5950573    82807
##  3 Swedish   2912535   791019
##  4 German    2618057  3780580
##  5 French    2363066  4191659
##  6 Dutch     2067480  1165874
##  7 Russian   1757033  3041819
##  8 Italian   1718551  2168354
##  9 Spanish   1717541  6338732
## 10 Polish    1490809  1124924
## # ... with 313 more rows
```

4.1 Convert to long format

The columns Articles and Users are grouped into one column and supplemented with the column Count.

```
NrArticlesCountry_long <- NrArticlesCountry %>%
gather(key = "Type", value = "Count",
Articles, Users)
head(NrArticlesCountry_long)
```

```
## # A tibble: 6 x 3
##   Language Type       Count
##   <chr>    <chr>      <int>
## 1 English  Articles 6383962
## 2 Cebuano  Articles 5950573
## 3 Swedish  Articles 2912535
## 4 German   Articles 2618057
## 5 French   Articles 2363066
## 6 Dutch    Articles 2067480
```

4.2 Convert to to wide format

The tibble changes again from the long format to the wide format.

```
NrArticlesCountry_long %>% spread(Type,Count) %>% head()
```

```
## # A tibble: 6 x 3
##   Language  Articles  Users
##   <chr>        <int>  <int>
## 1 Abkhazian     5806  16393
## 2 Acehnese     12513  22502
## 3 Adyghe         446   5243
## 4 Afar             1   4059
## 5 Afrikaans   100290 140380
## 6 Akan          1426  11128
```