Web scrapping using rvest: case study

```
#
# rvest: case study
#
# reference:
#
#   https://blog.rstudio.org/2014/11/24/rvest-easy-web-scraping-with-r/
#   https://cran.r-project.org/web/packages/rvest/vignettes/selectorgadget.html
#


# install.packages("rvest")
library(rvest)
```

```
## Warning: package 'rvest' was built under R version 3.3.3
```

```
## Loading required package: xml2
```

```
library(stringr)
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.3.3
```

```
library(plyr)

### Example: extract info from datacamp
url = 'https://www.datacamp.com/courses/all'
datacamp = read_html(url)

#   get course titles
course = html_nodes(datacamp, '.course-block__title') # css element
courseName = html_text(course)

#   get course hours
hours = html_nodes(datacamp, '.course-block__length')  # css element
hours
```

```
## {xml_nodeset (73)}
##  [1] <span class="course-block__length"> 4 hours</span>
##  [2] <span class="course-block__length"> 4 hours</span>
##  [3] <span class="course-block__length"> 4 hours</span>
##  [4] <span class="course-block__length"> 6 hours</span>
##  [5] <span class="course-block__length"> 4 hours</span>
##  [6] <span class="course-block__length"> 4 hours</span>
##  [7] <span class="course-block__length"> 4 hours</span>
##  [8] <span class="course-block__length"> 5 hours</span>
##  [9] <span class="course-block__length"> 6 hours</span>
## [10] <span class="course-block__length"> 3 hours</span>
## [11] <span class="course-block__length"> 3 hours</span>
## [12] <span class="course-block__length"> 3 hours</span>
## [13] <span class="course-block__length"> 4 hours</span>
## [14] <span class="course-block__length"> 4 hours</span>
## [15] <span class="course-block__length"> 4 hours</span>
## [16] <span class="course-block__length"> 4 hours</span>
```

```
## [17] <span class="course-block__length"> 4 hours</span>
## [18] <span class="course-block__length"> 4 hours</span>
## [19] <span class="course-block__length"> 4 hours</span>
## [20] <span class="course-block__length"> 2 hours</span>
## ...
```

```
courseHours = hours %>%
  html_text() %>%
  str_sub(2, 2) %>%
  as.numeric()

#   get categories
pattern_r = '([R]{1}|ggplot2)'
pattern_py = '(Python|pandas|Machine Learning|scikit-learn)'
pattern_sql = 'SQL'

str_count(courseName, pattern_r)
```

```
##  [1] 0 1 0 1 0 0 0 1 0 0 0 1 1 0 1 1 0 1 0 0 0 1 0 1 0 1 0 1 0 0 0 1 1 1 0
## [36] 0 0 1 1 1 0 0 1 2 1 1 0 0 0 1 1 2 2 1 1 0 0 1 1 1 1 1 1 1 1 2 1 2 1 1
## [71] 1 1 1
```

```
str_subset(courseName, pattern_r)
```

```
##  [1] "Introduction to R"
##  [2] "Intermediate R"
##  [3] "Data Visualization with ggplot2 (Part 1)"
##  [4] "Importing Data in R (Part 1)"
##  [5] "Cleaning Data in R"
##  [6] "Intermediate R - Practice"
##  [7] "Data Manipulation in R with dplyr"
##  [8] "Writing Functions in R"
##  [9] "Forecasting Using R"
## [10] "Data Visualization in R"
## [11] "Importing Data in R (Part 2)"
## [12] "Correlation and Regression"
## [13] "Introduction to R for Finance"
## [14] "Visualizing Time Series Data in R"
## [15] "Data Analysis in R, the data.table Way"
## [16] "Importing & Cleaning Data in R: Case Studies"
## [17] "Joining Data in R with dplyr"
## [18] "Data Visualization in R with lattice"
## [19] "Manipulating Time Series Data in R with xts & zoo"
## [20] "Reporting with R Markdown"
## [21] "Data Visualization with ggplot2 (Part 2)"
## [22] "Working with the RStudio IDE (Part 1)"
## [23] "Statistical Modeling in R (Part 1)"
## [24] "Exploratory Data Analysis in R: Case Study"
## [25] "Credit Risk Modeling in R"
## [26] "ARIMA Modeling with R"
## [27] "Unsupervised Learning in R"
## [28] "String Manipulation in R with stringr"
## [29] "Intermediate R for Finance"
## [30] "Importing and Managing Financial Data in R"
## [31] "Financial Trading in R"
```

```
## [32] "Data Visualization with ggplot2 (Part 3)"
## [33] "Introduction to Spark in R using sparklyr"
## [34] "Introduction to Portfolio Analysis in R"
## [35] "Bond Valuation and Analysis in R"
## [36] "Data Visualization in R with ggvis"
## [37] "Object-Oriented Programming in R: S3 and R6"
## [38] "Working with Geospatial Data in R"
## [39] "Quantitative Risk Management in R"
## [40] "Manipulating Time Series Data in R: Case Studies"
## [41] "Intermediate Portfolio Analysis in R"
## [42] "Working with the RStudio IDE (Part 2)"
## [43] "Statistical Modeling in R (Part 2)"
## [44] "Exploring Pitch Data with R"
```

```
sum(str_detect(courseName, pattern_r))
```

```
## [1] 44
```

```
str_count(courseName, pattern_py)
```

```
##  [1] 1 0 0 0 1 1 1 0 1 1 1 0 0 1 0 0 1 0 1 1 1 0 1 0 1 0 1 0 0 1 1 0 0 0 1
## [36] 0 0 0 0 0 1 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0
## [71] 0 0 0
```

```
str_subset(courseName, pattern_py)
```

```
##  [1] "Intro to Python for Data Science"
##  [2] "Intermediate Python for Data Science"
##  [3] "Deep Learning in Python"
##  [4] "Introduction to Data Visualization with Python"
##  [5] "Introduction to Machine Learning"
##  [6] "Python Data Science Toolbox (Part 1)"
##  [7] "Importing Data in Python (Part 1)"
##  [8] "pandas Foundations"
##  [9] "Python Data Science Toolbox (Part 2)"
## [10] "Cleaning Data in Python"
## [11] "Importing Data in Python (Part 2)"
## [12] "Supervised Learning with scikit-learn"
## [13] "Machine Learning Toolbox"
## [14] "Statistical Thinking in Python (Part 1)"
## [15] "Introduction to Databases in Python"
## [16] "Network Analysis in Python (Part 1)"
## [17] "Manipulating DataFrames with pandas"
## [18] "Network Analysis in Python (Part 2)"
## [19] "Unsupervised Learning in Python"
## [20] "Statistical Thinking in Python (Part 2)"
## [21] "Merging DataFrames with pandas"
## [22] "Machine Learning with the Experts: School Budgets"
```

```
sum(str_detect(courseName, pattern_py))
```

```
## [1] 22
```

```
str_subset(courseName, pattern_sql)
```

```
## [1] "Intro to SQL for Data Science"
```

```r
sum(str_detect(courseName, pattern_sql))
```

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

```r
#   get instructors
instr = html_nodes(datacamp, '.course-block__author-name')
instrName = html_text(instr)

df = data.frame(title = courseName,
                instructor = instrName,
                hours = courseHours)

#    add category to df
for (i in 1:nrow(df)) {
  if (str_detect(df$title[i], pattern_r))
    df$type[i] = 'R'
  else if (str_detect(df$title[i], pattern_py))
    df$type[i] = 'Python'
  else if (str_detect(df$title[i], pattern_sql))
    df$type[i] = 'SQL'
  else
    df$type[i] = 'Mixture'
}

head(df)
```

```
##                                 title           instructor hours   type
## 1     Intro to Python for Data Science   Filip Schouwenaars     4 Python
## 2                   Introduction to R Jonathan Cornelissen     4      R
## 3        Intro to SQL for Data Science        Nick Carchedi     4    SQL
## 4                       Intermediate R   Filip Schouwenaars     6      R
## 5 Intermediate Python for Data Science   Filip Schouwenaars     4 Python
## 6              Deep Learning in Python          Dan Becker     4 Python
```

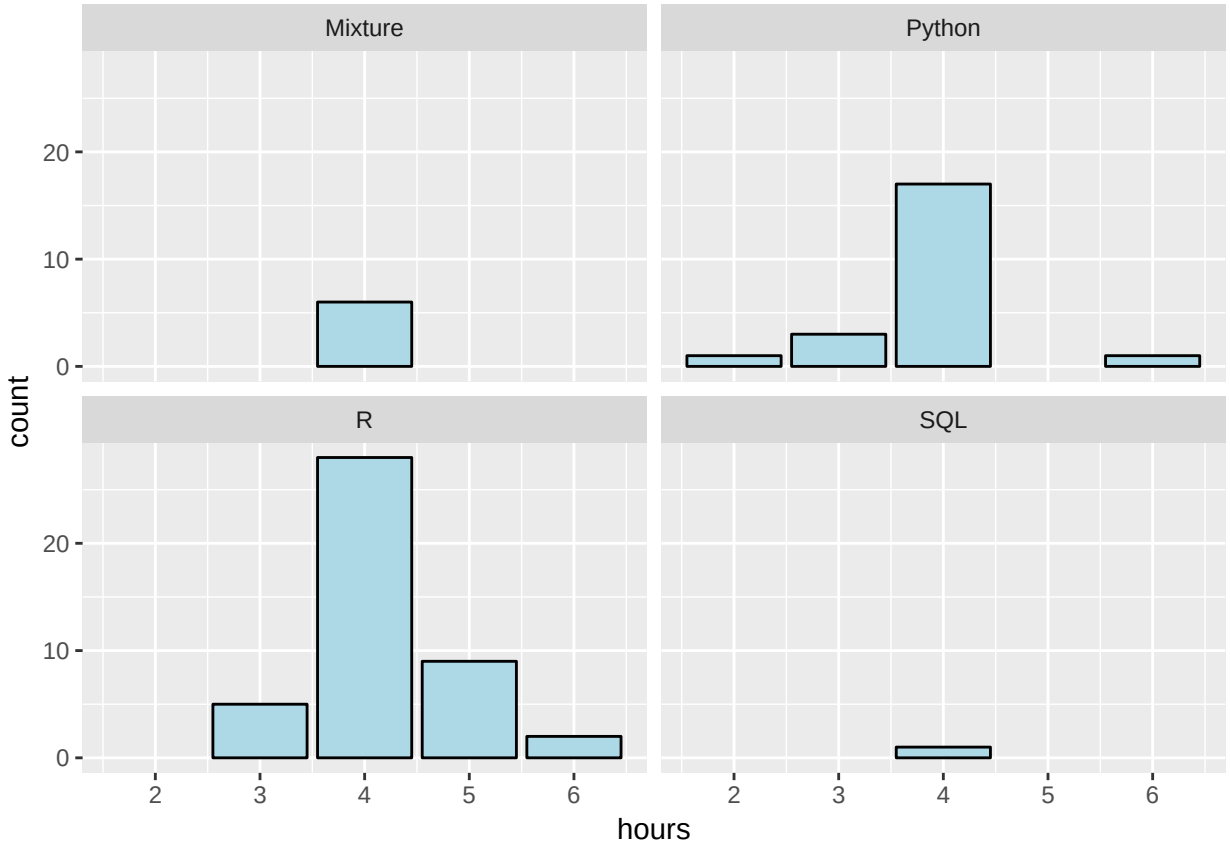
```r
### summary statistics
#   mean hour of each type of course
tapply(df$hours, df$type, mean)
```

```
##  Mixture   Python        R      SQL
## 4.000000 3.863636 4.181818 4.000000
```

```r
#   number of each type of course
table(df$type)
```

```
##
## Mixture  Python       R     SQL
##       6      22      44       1
```

```r
#   distribution of hours of each type
ggplot(df, aes(x = hours)) +
  geom_bar(color = 'black', fill = 'lightblue') +
  facet_wrap(~type, nrow = 2)
```




```
#   number of courses that each instructor teaches
table(df$instructor, df$type)
```

```
##
##                          Mixture Python R SQL
##   Alexander J. McNeil          0      0 1   0
##   Andreas Müller               0      1 0   0
##   Andrew Bray                  1      0 0   0
##   Arnaud Amsellem              0      0 1   0
##   Ben Baumer                   0      0 1   0
##   Benjamin Wilson              0      1 0   0
##   Brian M. Mills               0      0 1   0
##   Bryan Van de Ven             1      1 0   0
##   Charlotte Wickham            0      0 2   0
##   Clifford Ang                 0      0 1   0
##   Dan Becker                   0      1 0   0
##   Daniel Chen                  0      1 0   0
##   Daniel Kaplan                0      0 2   0
##   David Robinson               0      0 1   0
##   David S. Matteson            1      0 0   0
##   David Stoffer                0      0 1   0
##   Deepayan Sarkar              0      0 1   0
##   Dhavide Aruliah              0      3 0   0
##   Eric Ma                      0      2 0   0
##   Filip Schouwenaars           0      2 4   0
##   Garrett Grolemund            0      0 6   0
```

```
##   Hadley Wickham              0      0 1   0
##   Hank Roark                  0      0 1   0
##   Hugo Bowne-Anderson         0      4 0   0
##   Ilya  Kipnis                0      0 1   0
##   Jason Myers                 0      1 0   0
##   Jeffrey Ryan                0      0 1   0
##   Jo Hardin                   1      0 0   0
##   Jonathan Cornelissen        0      0 1   0
##   Joshua Ulrich               0      0 1   0
##   Justin Bois                 0      2 0   0
##   Kris Boudt                  0      0 1   0
##   Lore Dirick                 0      0 4   0
##   Matt Dowle                  0      0 1   0
##   Mine Cetinkaya-Rundel       1      0 0   0
##   Nick Carchedi               0      0 2   1
##   Peter Bull                  0      1 0   0
##   Richie Cotton               0      0 2   0
##   Rick Scavetta               0      0 3   0
##   Rob J. Hyndman              0      0 1   0
##   Ronald Pearson              0      0 1   0
##   Ross Bennett                0      0 1   0
##   Ted Kwartler                1      0 0   0
##   Vincent Vankrunkelsven      0      1 0   0
##   Zachary Deane-Mayer         0      1 0   0
```