Chapter 10: Làm việc với tập tin XML, JSON

Exercise 1: Đọc và hiển thị tập tin XML

- Link: http://www.w3schools.com/xml/plant_catalog.xml (http://www.w3schools.com/xml/plant_catalog.xml)
- Download nội dung xml từ link trên vào tập tin plant_catalog.xml
- Đọc nội dung tập tin và chuyển thành dataframe
- Xuất nội dung dataframe
- Liệt kê những cây trồng quanh năm (Annual) và cần có nắng (Sunny)

```
[1] "Plants grow in sunny and annual:"
```

	COMMON	BOTANICAL	ZONE	LIGHT	PRICE	AVAILABILITY
23	Black-Eyed Susan	Rudbeckia hirta	Annual	Sunny	$9.80	061899
26	Butterfly Weed	Asclepias tuberosa	Annual	Sunny	$2.78	063099

Exercise 2: Ghi nội dung vào file XML

- Sử dụng dữ liệu mtcars
- Tạo tài liệu xml từ dữ liệu này nhưng chỉ lấy thông tin: names, mpg, wt, gear
- Ghi tài liệu xml này vào tập tin mtcars.xml
- Đọc tập tin vừa ghi để xem kết quả

	names	mpg	wt	gear
Mazda RX4	Mazda RX4	21	2.62	4
Mazda RX4 Wag	Mazda RX4 Wag	21	2.875	4
Datsun 710	Datsun 710	22.8	2.32	4
Hornet 4 Drive	Hornet 4 Drive	21.4	3.215	3
Hornet Sportabout	Hornet Sportabout	18.7	3.44	3
Valiant	Valiant	18.1	3.46	3

Exercise 3: Đọc, xử lý và ghi nội dung JSON

- Cung cấp tập tin orange.json
- Đọc nội dung tập tin này => kiểm tra kiểu dữ liệu => đưa vào data.frame
- Cho biết cây cam có tuổi thọ cao nhất, thấp nhất
- Cho biết các cây cam có circumference >100 và age >1000. Có tất cả bao nhiêu cây cam?
- Chuyển dữ liệu những cây cam này thành json
- Ghi vào tập tin json
- Đọc nội dung tập tin vừa ghi để kiểm tra kết quả

	Tree	age	circumference
1	1	1004	115
2	1	1231	120
3	1	1372	142
4	1	1582	145
5	2	1004	156
6	2	1231	172
7	2	1372	203
8	2	1582	203
9	3	1004	108
10	3	1231	115
11	3	1372	139
12	3	1582	140
13	4	1004	167
14	4	1231	179
15	4	1372	209
16	4	1582	214
17	5	1004	125
18	5	1231	142
19	5	1372	174
20	5	1582	177

Exercise 4: Đọc nội dung từ URL và ghi nội dung JSON

- Cung cấp URL: http://phuong13021982.pythonanywhere.com/mystore/product_service/ (http://phuong13021982.pythonanywhere.com/mystore/product_service/)
- Đọc nội dung từ URL này => đọc JSON => chuyển thành data.frame tên là Tivis
- Bỏ cột description trong Tivis
- Chuyển Tivis thành json

	pk	name	fee	image
1	2	Asanzo 50 inch	9000000	images/asanzo_4k_50_11690000.jpg
2	11	Dell Vostro V3568 XF6C61	11999000	images/Dell_vostro.jpg
3	12	Macbook Air 2017 MQD32	18990000	images/MacbooAir.jpg
4	3	Panasonic 40 inch	6590000	images/panansonic_40_6590000.jpg
5	5	Samsung 32 inch	5999000	images/samsung_32_5990000.jpg
6	9	Samsung Galaxy J2 Prime	2690000	images/SamsungPrime.jpg
7	6	Sharp 45 inch	7490000	images/sharp_45_7490000.jpg
8	7	Sony 48 inch	11599000	images/sony_48_11599000.png
9	4	Sony 55 inch	20590000	images/sony_55_20590000.jpg
10	8	TCL 55 inch	9900000	images/tcl_55_9900000.jpg
11	1	Toshiba 32 inch	4590000	images/toshiba32_4590000.jpg
12	10	iPad WiFi 32GB New 2018	8390000	images/iPad_Samsung_Wifi.jpg

- Ghi vào tập tin tivis.json
- Đọc nội dung của tập tin vừa ghi và xem kết quả

Gợi ý:

Exercise 1: Đọc và hiển thị tập tin XML

In [2]:
```
# Load the package required to read XML files.
library("XML")
# Also load the other required package.
library("methods")
```

In [3]:
```
# download file xml
fileUrl = "http://www.w3schools.com/xml/plant_catalog.xml"
# tao file co ten la plant_catalog.xml
download.file(fileUrl, destfile = "Du_lieu/plant_catalog.xml")
```

In [14]:
```
# chuyen noi dung file sang data frame
xmldataframe <- xmlToDataFrame("Du_lieu/plant_catalog.xml")
print("Plants data frame:")
head(xmldataframe)
```

```
[1] "Plants data frame:"
```

COMMON	BOTANICAL	ZONE	LIGHT	PRICE	AVAILABILITY
Bloodroot	Sanguinaria canadensis	4	Mostly Shady	$2.44	031599
Columbine	Aquilegia canadensis	3	Mostly Shady	$9.37	030699
Marsh Marigold	Caltha palustris	4	Mostly Sunny	$6.81	051799
Cowslip	Caltha palustris	4	Mostly Shady	$9.90	030699
Dutchman's-Breeches	Dicentra cucullaria	3	Mostly Shady	$6.44	012099
Ginger, Wild	Asarum canadense	3	Mostly Shady	$9.03	041899

In [8]:
```
#loc bo du lieu cac cay trong quanh nam va can anh nang
data_annual_sunny <- subset(xmldataframe, xmldataframe$ZONE=="Annual" &
                            xmldataframe$LIGHT=="Sunny")
print("Plants grow in sunny and annual:")
data_annual_sunny
```

```
[1] "Plants grow in sunny and annual:"
```

	COMMON	BOTANICAL	ZONE	LIGHT	PRICE	AVAILABILITY
23	Black-Eyed Susan	Rudbeckia hirta	Annual	Sunny	$9.80	061899
26	Butterfly Weed	Asclepias tuberosa	Annual	Sunny	$2.78	063099

Exercise 2: Ghi nội dung vào file XML

In [9]:
```
# Load the packages required to read XML files.
library("XML")
library("methods")

df <- mtcars
df <- cbind(names = rownames(df), df)
rownames(df) <- c()
print(head(df))
```

```
              names  mpg cyl disp  hp drat    wt  qsec vs am gear carb
1         Mazda RX4 21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
2     Mazda RX4 Wag 21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
3        Datsun 710 22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
4    Hornet 4 Drive 21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
5 Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
6           Valiant 18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

In [10]:
```
print("Create xml...")
```

```
[1] "Create xml..."
```

In [11]:
```
doc = newXMLDoc()
# Simple creation of an XML tree using these functions
top = newXMLNode("cars", doc = doc)
for(row in 1:nrow(df)){
  carnode = newXMLNode("car", parent=top)
  newXMLNode("names", df[row, "names"],parent=carnode)
  newXMLNode("mpg", df[row, "mpg"],parent=carnode)
  newXMLNode("wt", df[row, "wt"],parent=carnode)
  newXMLNode("gear", df[row, "gear"],parent=carnode)
}
```

In [12]:
```
# save file
print("Saving file...")
cat(saveXML(doc,
            indent = TRUE,
            prefix = "<?xml version=\"1.0\" encoding=\"utf-8\"
                              standalone=\"no\"?>\n"),
    file="Du_lieu/mtcars_new.xml")
print("Complete!")
```

```
[1] "Saving file..."
[1] "Complete!"
```

In [15]:
```
# chuyen noi dung file sang data frame
xmldataframe <- xmlToDataFrame("Du_lieu/mtcars_new.xml")
head(xmldataframe)
```

names	mpg	wt	gear
Mazda RX4	21	2.62	4
Mazda RX4 Wag	21	2.875	4
Datsun 710	22.8	2.32	4
Hornet 4 Drive	21.4	3.215	3
Hornet Sportabout	18.7	3.44	3
Valiant	18.1	3.46	3

Exercise 3: Đọc, xử lý và ghi nội dung JSON

In [16]:
```
# Load the package required to read JSON files.
library("rjson")
```

In [24]:
```
# Give the input file name to the function.
result <- fromJSON(file = "Du_lieu/orange.json")
# cho biet kieu du lieu cua orange
print(paste("Data type:",class(result)))
#neu khong phai la data frame thi doi thanh data frame
#in ket qua
data <- data.frame(result)
print(head(data))
```

```
[1] "Data type: list"
  Tree  age circumference
1    1  118            30
2    1  484            58
3    1  664            87
4    1 1004           115
5    1 1231           120
6    1 1372           142
```

In [18]:
```
# cho biet trong nhung cay cam nay cay nao co tuoi tho cao nhat
data_max_year <- data[which.max(data$age ),]
print("Max year:")
print(data_max_year)
```

```
[1] "Max year:"
  Tree  age circumference
7    1 1582           145
```

In [19]:
```
# cho biet trong nhung cay cam nay cay nao co tuoi tho thap nhat'
data_min_year <- data[which.min(data$age ),]
print("Min year:")
print(data_min_year)
```

```
[1] "Min year:"
  Tree age circumference
1    1 118            30
```

In [25]:
```
# danh sach cac cay trong co circumference >100 va age >1000
large_100_age_1000 <- subset(data, data$age>1000 &
                               data$circumference>100)
print(head(large_100_age_1000))
print(class(large_100_age_1000))
print(paste("Number of rows:", nrow(large_100_age_1000)))
```

```
   Tree  age circumference
4     1 1004           115
5     1 1231           120
6     1 1372           142
7     1 1582           145
11    2 1004           156
12    2 1231           172
[1] "data.frame"
[1] "Number of rows: 20"
```

In [21]:
```
# ghi vao file
# doc file de kiem tra ket qua
large_100_age_1000_json <- toJSON(large_100_age_1000)
write(large_100_age_1000_json, file="Du_lieu/large_100_age_1000_new.json")
```

In [26]:
```
# Give the input file name to the function.
result <- fromJSON(file = "Du_lieu/large_100_age_1000_new.json")
#in ket qua
data <- data.frame(result)
print(head(data))
```

```
  Tree  age circumference
1    1 1004           115
2    1 1231           120
3    1 1372           142
4    1 1582           145
5    2 1004           156
6    2 1231           172
```

Exercise 4: Đọc nội dung từ URL và ghi nội dung JSON

In [27]:
```
library(httr)
library("jsonlite")
```

```
Attaching package: 'jsonlite'


The following objects are masked from 'package:rjson':

    fromJSON, toJSON
```

In [28]:
```
# doc noi dung tu internet
URL <- "http://phuong13021982.pythonanywhere.com/mystore/product_service/"
getURL <- GET(URL)
content <- rawToChar(getURL$content)
json <- fromJSON(content)
Tivis <- data.frame(json)
print(class(Tivis))
#bo cot description
Tivis$description <-NULL
print(Tivis)
```

```
[1] "data.frame"
   pk                       name      fee                                image
1   2             Asanzo 50 inch  9000000 images/asanzo_4k_50_11690000.jpg
2  11 Dell Vostro V3568 XF6C61 11999000          images/Dell_vostro.jpg
3  12   Macbook Air 2017 MQD32 18990000            images/MacbooAir.jpg
4   3          Panasonic 40 inch  6590000 images/panansonic_40_6590000.jpg
5   5            Samsung 32 inch  5999000    images/samsung_32_5990000.jpg
6   9  Samsung Galaxy J2 Prime  2690000          images/SamsungPrime.jpg
7   6              Sharp 45 inch  7490000      images/sharp_45_7490000.jpg
8   7               Sony 48 inch 11599000      images/sony_48_11599000.png
9   4               Sony 55 inch 20590000      images/sony_55_20590000.jpg
10  8                TCL 55 inch  9900000        images/tcl_55_9900000.jpg
11  1            Toshiba 32 inch  4590000     images/toshiba32_4590000.jpg
12 10  iPad WiFi 32GB New 2018  8390000     images/iPad_Samsung_Wifi.jpg
```

In [33]:
```
#ghi noi dung nay vao tap tin tivi.json'
Tivis_json <- toJSON(Tivis)
write(Tivis_json, file="Du_lieu/tivis_new.json")
```

In [35]:
```
# Give the input file name to the function.
result <- fromJSON(txt= "Du_lieu/tivis_new.json")
# Print the result.
print("json data read from file:")
data <- data.frame(result)
data
```

```
[1] "json data read from file:"
```

pk	name	fee	image
2	Asanzo 50 inch	9000000	images/asanzo_4k_50_11690000.jpg
11	Dell Vostro V3568 XF6C61	11999000	images/Dell_vostro.jpg
12	Macbook Air 2017 MQD32	18990000	images/MacbooAir.jpg
3	Panasonic 40 inch	6590000	images/panansonic_40_6590000.jpg
5	Samsung 32 inch	5999000	images/samsung_32_5990000.jpg
9	Samsung Galaxy J2 Prime	2690000	images/SamsungPrime.jpg
6	Sharp 45 inch	7490000	images/sharp_45_7490000.jpg
7	Sony 48 inch	11599000	images/sony_48_11599000.png
4	Sony 55 inch	20590000	images/sony_55_20590000.jpg
8	TCL 55 inch	9900000	images/tcl_55_9900000.jpg
1	Toshiba 32 inch	4590000	images/toshiba32_4590000.jpg
10	iPad WiFi 32GB New 2018	8390000	images/iPad_Samsung_Wifi.jpg