Under the Wood

Getting and Cleaning Data: Quiz 2

2016-03-04

Quiz

1. Register an application with the Github API here https://github.com/settings/applications. Access the API to get information on your instructors repositories (hint: this is the url you want "https://api.github.com/users/jtleek/repos"). Use this data to find the time that the datasharing repo was created. What time was it created?

   This tutorial may be useful (https://github.com/hadley/httr/blob/master/demo/oauth2-github.r). You may also need to run the code in the base R package and not R studio.

   - 2012-06-20T18:39:06Z
   - 2013-08-28T18:18:50Z
   - 2013-11-07T13:25:07Z
   - 2012-06-21T17:28:38Z

2. The sqldf package allows for execution of SQL commands on R data frames. We will use the sqldf package to practice the queries we might send with the dbSendQuery command in RMySQL.

   Download the American Community Survey data and load it into an R object called

   ```
   acs
   ```

   https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Fss06pid.csv

   Which of the following commands will select only the data for the probability weights pwgtp1 with ages less than 50?

   - sqldf("select pwgtp1 from acs where AGEP < 50")
   - sqldf("select * from acs")

- sqldf("select pwgtp1 from acs")
- sqldf("select * from acs where AGEP < 50")

3. Using the same data frame you created in the previous problem, what is the equivalent function to unique(acs$AGEP)

   - sqldf("select distinct pwgtp1 from acs")
   - sqldf("select unique AGEP from acs")
   - sqldf("select AGEP where unique from acs")
   - sqldf("select distinct AGEP from acs")

4. How many characters are in the 10th, 20th, 30th and 100th lines of HTML from this page:

   http://biostat.jhsph.edu/~jleek/contact.html

   (Hint: the nchar() function in R may be helpful)

   - 45 0 2 2
   - 45 31 2 25
   - 45 92 7 2
   - 43 99 7 25
   - 45 31 7 31
   - 45 31 7 25
   - 43 99 8 6

5. Read this data set into R and report the sum of the numbers in the fourth of the nine columns.

   https://d396qusza40orc.cloudfront.net/getdata%2Fwksst8110.for

   Original source of the data: http://www.cpc.ncep.noaa.gov/data/indices/wksst8110.for

   (Hint this is a fixed width file format)

   - 36.5

   - 35824.9

   - 32426.7

- 222243.1
- 101.83
- 28893.3

My Solution

1. Load library and require the package:

```
library(httr)
require(httpuv)
require(jsonlite)
```

setting the OAuth for github:

```
oauth_endpoints("github")
```

after register an application through https://github.com/settings/applications

```
myapp <- oauth_app("quiz2",
                   key = "b87a2841e89ce55fe15e",
                   secret = "b1d4ede3d6f11878f51d0a842e3aa4c219b59c27")
```

get OAuth credentials:

```
github_token <- oauth2.0_token(oauth_endpoints("github"), myapp)
```

use API:

```
req <- GET("https://api.github.com/users/jtleek/repos",
           config(token = github_token))
stop_for_status(req)
output <- content(req)
jsondata <- fromJSON(toJSON(output))
subset(jsondata, name == "datasharing", select = c(created_at))
```

the result is `2013-11-07T13:25:07Z`.

2. After install the package "sqldf", download the file from the Internet, use browser or commands:

```
library(sqldf)
fileUrl <- "https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Fss06pid.csv"
download.file(url=fileUrl,destfile="getdata.csv", method="curl")
acs <- read.csv("getdata.csv")
sqldf("select pwgtp1 from acs where AGEP < 50")
```

so the answer is `sqldf("select pwgtp1 from acs where AGEP < 50")`.

3. The same file, so use command:

```
sqldf("select distinct AGEP from acs")
```

the answer is `sqldf("select distinct AGEP from acs")`.

4. Read the URL:

```
char <- url("http://biostat.jhsph.edu/~jleek/contact.html")
htmlCode <- readLines(char)
close(char)
```

then find out the result:

```
nchar(htmlCode[10])
nchar(htmlCode[20])
nchar(htmlCode[30])
```

the result is `45,31,7`.

5. Try to combine the "download" and "read" function together use the command line:

```
data <- read.fwf(file = "https://d396qusza40orc.cloudfront.net/getdata%2Fwksst8110
                 skip = 4,
                widths = c(12, 7,4, 9,4, 9,4, 9,4))
sum(data[, 4])
```

the result is `32426.7` .

A very useful blog: thoughtfulbloke aka David Hood

The END.

‹ Getting and Cleanning Data: Quiz 1

Getting and Cleaning Data: Quiz 3 ›