Web Scraping password protected website using R and "rvest" package

Mouli

2022-06-29

In this tutorial we are going to web scrape https://the-internet.herokuapp.com/login website. So lets gets started.

About rvest package

rvest is part of `tidyverse package`. It helps you scrape (or harvest) data from web pages. It is designed to work with `magrittr` to make it easy to express common web scraping tasks, inspired by libraries like `beautiful soup` and `RoboBrowser`.

know more about rvest by visiting below links https://rvest.tidyverse.org/
https://github.com/tidyverse/rvest/

Installing and loading required packages

install the rvest package by using `install.packages("rvest")` command.

load the rvest library

```r
library("rvest");
```

Setting URL's and login details

set the URL of the page you want to login

```r
login_url <- "https://the-internet.herokuapp.com/login"
```

set the URL of the page you want to access after login

```r
secure_url <- "https://the-internet.herokuapp.com/secure"
```

set the login details

```r
username_text <- "tomsmith"
password_text <- "SuperSecretPassword!"
```

Creating session and pulling the login form and submitting the form

Create a session with login URL

```
pgsession <- session(login_url)
```

Get the login form from the login page

```
pgform <- html_form(pgsession)[[1]]
```

fill the login details in the form

```
filled_form <- html_form_set(pgform,
                            username = username_text,
                            password = password_text,)
```

Submit the filled login form

```
session_submit(pgsession,filled_form, style = "POST")
```

```
## <session> https://the-internet.herokuapp.com/secure
##   Status: 200
##   Type:   text/html;charset=utf-8
##   Size:   1974
```

Scraping the target website

Get the required text from the secure page by moving the session to the secure page

use selector gadget chrome extension to get the html_nodes

```
text1 <- session_jump_to(pgsession, secure_url) %>%
  read_html() %>%
  html_nodes(".subheader") %>%
  html_text()
```

Display the text from the secure page

```
text1
```

```
## [1] "Welcome to the Secure Area. When you are done click logout below."
```

we got our required text … We can get other text,lines,etc in the same way…