RCurl

The RCurl package is an R-interface to the libcurl library that provides HTTP facilities. This allows us to download files from Web servers, post forms, use HTTPS (the secure HTTP), use persistent connections, upload files, use binary content, handle redirects, password authentication, etc.

Author: Duncan Temple Lang

Victoria Mansfield (Speaker) , Yue Bai,

Data Science Challenge

Web Scraping

- Getting a page's HTML code and parsing through to get data that we want.
- Posting forms and request to the Internet.

R-language interface

Given the increasing role of HTTP and Web connectivity, and the desire to use R in ways that can access data and services in other domains via HTTP, a more general, flexible and complete R-language interface to client-side HTTP is desirable. The RCurl package provides such an interface for R.

Basic Functionality

There are three high-level functions in RCurl: getURL(), getForm(), and postForm().

- getURL() and getURI(): These functions download one or more URIs (a.k.a. URLs). Input the URL and return the HTML of the Web page.
- getForm() and postForm(): These functions provide facilities for submitting an HTML form. Used for automating retrieval of data sets that might otherwise require a form submission for each data set.

Example

Example (getURL().)

```
# Three tables (total area, land area, and water area) in this Web pages.
# We want to know the land area of the 50 United States.
x<-getURL("https://simple.wikipedia.org/wiki/List_of_U.S._states_by_area")
```

```
## [1] "<!DOCTYPE html>\n<html class=\"client-nojs\" lang=\"en\" dir=\"ltr\">\n<head>\n<meta charset=\"UTF-8\"/>\
n<title>List of U.S. states by area - Simple English Wikipedia, the free encyclopedia</title>\n<script>document.d
ocumentElement.className = document.documentElement.className.replace( /(^|\\s)client-nojs(\\s|$)/, \"$1client-js
$2\" );</script>\n<script>(window.RLQ=window.RLQ||[]).push(function(){mw.config.set({\"wgCanonicalNamespace\":\"\
\",\"wgCanonicalSpecialPageName\":false,\"wgNamespaceNumber\":0,\"wgPageName\":\"List_of_U.S._states_by_area\",\"w
```

(area<-readHTMLTable(x))

```
## $`NULL`
##      Rank          State       km²     miles²
## 1       1         Alaska 1,717,854    663,267
## 2       2          Texas   696,621    268,581
## 3       3     California   423,970    163,696
## 4       4        Montana   380,838    147,042
## 5       5     New Mexico   314,915    121,589
## 6       6        Arizona   295,254    113,998
## 7       7         Nevada   286,351    110,561
## 8       8       Colorado   269,601    104,094
## 9       9         Oregon   254,805     98,381
## 10     10        Wyoming   253,348     97,818
```

(land_area<-area[[2]])

```
##      Rank          State       km²  sq.miles
## 1       1         Alaska 1,481,347   567,400
## 2       2          Texas   678,051   261,797
## 3       3     California   403,933   155,959
## 4       4        Montana   376,979   145,552
## 5       5     New Mexico   314,309   121,356
## 6       6        Arizona   294,312   113,635
## 7       7         Nevada   284,448   109,826
## 8       8       Colorado   268,627   103,718
## 9       9         Oregon   254,818    98,386
## 10     10        Wyoming   251,489    97,105
## 11     11          Idaho   214,314    82,747
```