

The Bigger Picture

Matthew J. Denny

mzd5530@psu.edu – www.mjdenny.com – @MatthewJDenny

www.mjdenny.com/ICPSR_Data_Science_2015.html

July 31, 2015



INSTITUTE FOR
SOCIAL SCIENCE
RESEARCH

UNIVERSITY OF MASSACHUSETTS AMHERST



Overview

1. Using C++ with R – Rcpp.
2. Web Scraping.
3. Package Development.

1. C++ Code with R

Rcpp, RcppArmadillo, and BH

1. Rcpp is integrated with RStudio – easy C++ coding
2. RcppArmadillo – gives you access to linear algebra libraries.
3. BH – gives you access to random number generation.
4. Shallow vs. deep data structures.
5. Best for sampling and looping.

Basic RcppArmadillo C++ function

```
#include <RcppArmadillo.h>
#include <string>
//[[Rcpp::depends(RcppArmadillo)]]
using namespace Rcpp;
//[[Rcpp::export]]
List My_Function(
    int some_number,
    List some_vectors,
    arma::vec a_vector,
    arma::mat example_matrix
){
    ...
    List to_return(1);
    to_return[0] = some_data;
    return to_return;
}
```

Looping + Conditionals (ex. word counter)

```
for(int n = 0; n < number_of_bills; ++n){
    report(n);
    int length = Bill_Lengths[n];
    std::vector<std::string> current = Bill_Words[n];
    for(int i = 0; i < length; ++i){
        int already = 0;
        int counter = 0;
        while(already == 0){
            if(unique_words[counter] == current[i]){
                unique_word_counts[counter] += 1;
                already = 1;
            }
            counter +=1;
        }
    }
}
```

Drawing Random Numbers

```
// add to second and third lines of file
#include <random>
#include <math.h>

// set RNG and seed
boost::mt19937_64 generator(seed);

// define a uniform distribution and draw from it
boost::uniform_real_distribution<double> udist(0.0, 1.0);
double rand_num = udist(generator);

// define a normal distribution and draw from it
boost::normal_distribution<double> ndist(mu,sigsq);
my_matrix(k,b) = ndist(generator);
```

Other Useful Stuff

```
# In R define
Report <- function(string){print(string)}

// In C++ we write (inside function definition)
Function report("Report");

// now we can print stuff back up to R
report(n);

// initialize a vector/matrix to zeros
arma::vec myvec = arma::zeros(len);

// some math operators
double d = exp(log(pow(2,4)));
```


Things To Watch Out For

1. Use Armadillo data structures – Rcpp data structures can overflow memory.
2. Cast integers as doubles before dividing.
3. Low latency + faster looping = 50-2,000x speedup.

[www.mjdenny.com/
Rcpp_Intro.html](http://www.mjdenny.com/Rcpp_Intro.html)

2. Web Scraping

I Can Has All Your Webpages?

1. Automate downloading webpages.
2. Automate capturing tweets, posts, files, etc.
3. Scraping vs. Crawling
4. What you ask for is what you get.
5. Legal issues.

HTML and URLs

1. Before you start, need to look at source code.
2. Identify patterns in what you will scrape.
3. Write text processing code to extract useful fields.
4. Get vector of URLs.

Take Me To The Source

The screenshot shows a Google Chrome browser window displaying the page for H.R. 1599 - Safe and Accurate Food Labeling Act of 2015. The address bar shows the URL <https://www.congress.gov/bill/114th-congress/house-bill/1599>. The browser's menu bar is open, showing options like New Tab, New Window, and History. The 'More tools' submenu is also open, with 'View Source' highlighted. The page content includes the title 'H.R.1599 - Safe and Accurate Food Labeling Act of 2015', the sponsor 'Rep. Pompeo, Mike [R-KS-4]', and the committee 'House - Energy and Commerce; Agriculture | Nutrition. and Forestrv'. The 'View Source' option is highlighted with a red box, indicating the next step in the process.

Advanced | Tips

Print Subscribe Share/Save Give Feedback

H.R.1599 - Safe and Accurate Food Labeling Act of 2015

114th Congress (2015-2016) | [Get alerts](#)

BILL Hide Overview X

Sponsor: [Rep. Pompeo, Mike \[R-KS-4\]](#) (Introduced 03/25/2015)

Committees: House - Energy and Commerce; Agriculture | Nutrition. and Forestrv

More tools

View Source

Take Me To The Source

```
1 <!DOCTYPE html><html class="no-js" lang="en">
2 <head>
3   <title>H.R.1599 - 114th Congress (2015-2016): Safe and Accurate Food
  Labeling Act of 2015 | Congress.gov | Library of Congress</title>   <meta
  charset="UTF-8">
4 <meta http-equiv="Content-Type" content="text/html; charset=UTF-8" >
5 <meta name="description" content="Summary of H.R.1599 - 114th Congress (2015-
  2016): Safe and Accurate Food Labeling Act of 2015" >
6 <meta name="dc.creator" content="Pompeo, Mike" >
7 <meta name="dc.title" content="H.R.1599 - 114th Congress (2015-2016): Safe and
  Accurate Food Labeling Act of 2015" >
8 <meta name="dc.identifier" content="https://www.congress.gov/bill/114th-
  congress/house-bill/1599" >
9 <meta name="dc.subject" content="House Bill" >
10 <meta name="dc.coverage" content="2015/2016" >
11 <meta name="dc.coverage" content="2015-03-25" >
12 <meta name="dc.coverage" content="07/24/2015" >
13 <meta name="dc.date" content="07/24/2015" >
14 <meta name="dc.subject" content="Agriculture and Food" >
15 <meta name="dc.language" content="eng" >
16 <meta name="dc.type" content="legislation" >
17 <meta name="dc.type" content="webpage" >
18 <meta name="dc.rights" content="Text is government work" >
19 <meta name="dc.subject" content="Legislative Data" >
20 <meta name="viewport" content="width=device-width,initial-scale=1" >   <link
```

Take Me To The Source

```
503 </div><div id="main" class="wrapper_std" role="main"><div
class="tntFormWrapper" id="summarySelector">
504 <p>There is one summary for this bill. <a href="/help/legislative-
glossary/#glossary_billsummary" target="_blank">Bill summaries</a> are
authored by <a href="/help/legislative-glossary/#glossary_crs"
target="_blank">CRS</a>.</p></div>
505 <h3 class="currentVersion">Shown Here:<br><span>Introduced in House
(03/25/2015)</span></h3>
506 <p><b>Safe and Accurate Food Labeling Act of 2015</b></p> <p>This bill amends
the Federal Food, Drug, and Cosmetic Act to require the developer of a
bioengineered organism intended as food to submit a premarket biotechnology
notification to the Food and Drug Administration (FDA). A "bioengineered
organism" (commonly called a "genetically modified organism"
or "GMO") is a plant or part of a plant that has been modified
through recombinant DNA techniques in a way that could not be obtained using
conventional breeding techniques.</p> <p>The premarket notification must
include the developer's determination that food from, containing, or
consisting of the GMO (GMO food) is as safe as a comparable non-GMO food. For
the GMO to be sold as food, the FDA must not object to the developer's
determination. If the FDA determines that there is a material difference
between a GMO food and a comparable non-GMO food, the FDA can specify labeling
that informs consumers of the difference.</p> <p>A food label can only claim
that a food is non-GMO if the ingredients are subject to certain supply chain
process controls. No food label can suggest that non-GMO foods are safer than
GMO foods. A food can be labeled as non-GMO even if it is produced with a GMO
```


1. Don't scrape too fast – that's a DDOS = Jail Time!
2. Check site policies.
3. Identify yourself.
4. Contact site administrator.

Code

```
# Load the library
library(scrapeR)

# Set options (need to do this)
my_opts <- list(ssl.verifypeer = FALSE)

# Scrape a page
page <- getURL(url, .opts = my_opts)

# Scrape many pages
for(i in 1:length(urls)){
  pages[i] <- getURL(urls[i],
                    .opts = my_opts)
  Sys.sleep(6)
}
```

[www.mjdenny.com/
R_Tutorial.html](http://www.mjdenny.com/R_Tutorial.html)

3. Package Development

[www.mjdenny.com/
R_Package_Pictorial.html](http://www.mjdenny.com/R_Package_Pictorial.html)