

Web scraping with **rvest** in R

Jerid Francom

March 1, 2015

In this ExploRation, I will demonstrate how to scrape text data from the web with R. This particular example aims to collect a series of State of the Union (SOTU) speeches [1947-present] from <http://www.presidency.ucsb.edu/> and write the plain-text contents to disc. The bulk of the work will be done with the recently released **rvest** package. The scripting will also employ the **magrittr** package for writing legible code.

To get started first we identify the sub-page [../sou.php](http://www.presidency.ucsb.edu/./sou.php) that contains the links of interest.

Audio/Video Index	reflects each message's placement in the President's term.									
Elections	President George W. Bush delivered his last State of the Union Address on January 28, 2008. Bush had the right to deliver either a written or oral State of the Union in the days immediately before leaving office in 2009. However, like Presidents Reagan, George H.W. Bush, and Clinton, he chose not to do so. Presidents Truman, Eisenhower, Johnson, Ford, and Carter chose to do so.									
Election Index										
Florida 2000										
Links										
Presidential Libraries										
President	years of term	Delivered as a Speech					Delivered as a Written Message			
		Political Time (see essay above)					Political Time (see essay above)			
		1st	2nd	3rd	4th	end 4th	1st	2nd	3rd	4th
Barack Obama	2013-pres. 2009-2013	2013 2009*	2014 2010	2015 2011						
George W. Bush	2005-2009 2001-2005	2005 2001*	2006 2002	2007 2003	2008 2004					
William J. Clinton	1997-2001 1993-1997	1997 1993*	1998 1994	1999 1995	2000 1996					
George Bush	1989-1993	1989*	1990	1991	1992					
Ronald Reagan	1985-1989 1981-1985	1985 1981*	1986 1982	1987 1983	1988 1984					
Jimmy Carter	1977-1981		1978	1979	1980		1978	1979	1980	
Gerald R. Ford	1974-1977			1975	1976	1977				
Richard M. Nixon	1973-1974 1969-1973		1974 1970	1971	1972		1973+	1974		1972
Lyndon B. Johnson	1965-1969 1964-1965	1965	1966	1967	1968 1964	1969				
John F. Kennedy	1961-1963	1961	1962	1963						

This page contains links to pages in which all of the SOTU addresses. To load that page into R, as a parsed html object we use **rvest**'s `html()` function.

```
library("rvest")
# Load the page
main.page <- html(x = "http://www.presidency.ucsb.edu/sou.php")
```

Once we have the page, the next step is to identify how to isolate the links that we are interested in from other links on the page. The documentation for the package refers to **Selectorgadget** a bookmarklet for your browser that allows you to point-and-click your way to identifying either the CSS or XPATH need to get the target html objects.

Activating Selectorgadget, you then click on the html object you want and then see what becomes highlighted. In most cases this will highlight more objects than you want, so then you click again on the object(s) you do not want to isolate. In our case, clicking first the "2013" link in the SOTU listing and then the "Florida 2000" link leaves us with the right objects selected.

Audio/Video Index	reflects each message's placement in the President's term.									
Elections	President George W. Bush delivered his last State of the Union Address on January 28, 2008. Bush had the right to deliver either a written or oral State of the Union in the days immediately before leaving office in 2009. However, like Presidents Reagan, George H.W. Bush, and Clinton, he chose not to do so. Presidents Truman, Eisenhower, Johnson, Ford, and Carter chose to do so.									
Election Index										
1790-2015										
Links										
Presidential Libraries										

President	years of term	Delivered as a Speech					Delivered as a Written Message			
		Political Time (see essay above)					Political Time (see essay above)			
		1st	2nd	3rd	4th	end 4th	1st	2nd	3rd	4th
Barack Obama	2013-pres. 2009-2013	2013	2014	2015						
		2009*	2010	2011	2012					
George W. Bush	2005-2009 2001-2005	2005	2006	2007	2008					
		2001*	2002	2003	2004					
William J. Clinton	1997-2001 1993-1997	1997	1998	1999	2000					
		1993*	1994	1995	1996					
George Bush	1989-1993	1989*	1990	1991	1992					
Ronald Reagan	1985-1989 1981-1985	1985	1986	1987	1988					
		1981*	1982	1983	1984					
Jimmy Carter	1977-1981		1978	1979	1980		1978	1979	1980	
Gerald R. Ford	1974-1977			1975	1976	1977				
Richard M. Nixon	1973-1974 1969-1973		1974				1973†	1974		
			1970	1971	1972					1972
Lyndon B. Johnson	1965-1969 1964-1965	1965	1966	1967	1968	1969				
					1964					
John F. Kennedy	1961-1963	1961	1962	1963						

Now we can return to R, and use the CSS selector `'ver12 a'` to get our links. The `html_nodes()` function gets the elements we want, but they come with `html-warts` and all. For the URLs we use the `html_attr()` function and specify that we want the part contained under `href` (ex. `1790`). The same basic process is applied to get the link text, but instead we use the `html_text()` function to get the '1790' part of the previous URL example. Then we combine the results into a data frame `sotu`.

```
# Get link URLs
urls <- main.page %>% # feed `main.page` to the next step
  html_nodes(".ver12 a") %>% # get the CSS nodes
  html_attr("href") # extract the URLs

# Get link text
links <- main.page %>% # feed `main.page` to the next step
  html_nodes(".ver12 a") %>% # get the CSS nodes
  html_text() # extract the link text

# Combine `links` and `urls` into a data.frame
sotu <- data.frame(links = links, urls = urls, stringsAsFactors = FALSE)
head(sotu)
```

```
##   links                                     urls
## 1  2013 http://www.presidency.ucsb.edu/ws/index.php?pid=102826
## 2  2014 http://www.presidency.ucsb.edu/ws/index.php?pid=104596
## 3  2015 http://www.presidency.ucsb.edu/ws/index.php?pid=108031
## 4  2009 http://www.presidency.ucsb.edu/ws/index.php?pid=85753
## 5  2010 http://www.presidency.ucsb.edu/ws/index.php?pid=87433
## 6  2011 http://www.presidency.ucsb.edu/ws/index.php?pid=88928
```

The results look great. We still need to extract only those addresses we are interested in, dates between 1947-2015. To do this we simply use the `%in%` operator to filter our `sotu$links` column by the vector `1947:2015`.

```
sotu <- subset(x = sotu, links %in% 1947:2015) # Truman to Obama
```

The next step is to follow each of these links, extract the text, and write the text to disc. To keep our files organized, we are going to dynamically generate the file names marking them as either `republican` or

democrat by using the dates that Republicans held the presidency and then append the date. This will result in files with the format: republican-2001.txt.

First the filter: dates which Republicans were in office.

```
# Vector to mark SOTU address political party
republicans <- c(1954:1960, 1970:1974, 1974:1977, 1981:1988, 1989:1992, 2001:2008)
```

Now the aim is to loop through each of the links in our `sotu` data.frame (i.e. the number of rows `nrow(sotu)`), grabbing the parsed html (`html()`) and isolating (`".displaytext"`) and extracting the relevant text (`html_text()`). After the text has been scraped then we decide if the text should be marked Republican or Democrat using the previous filter and an `ifelse()` statement, compile the file name, and write that file to disc.

```
# Loop over each row in `sotu`
for(i in seq(nrow(sotu))) {
  text <- html(sotu$urls[i]) %>% # load the page
    html_nodes(".displaytext") %>% # isolate the text
    html_text() # get the text
  # Find the political party of this link
  party <- ifelse(test = sotu$links[i] %in% republicans,
                 yes = "republican", no = "democrat")
  # Create the file name
  filename <- paste0("texts/", party, "-", sotu$links[i], ".txt")
  sink(file = filename) %>% # open file to write
  cat(text) # write the file
  sink() # close the file
}
```

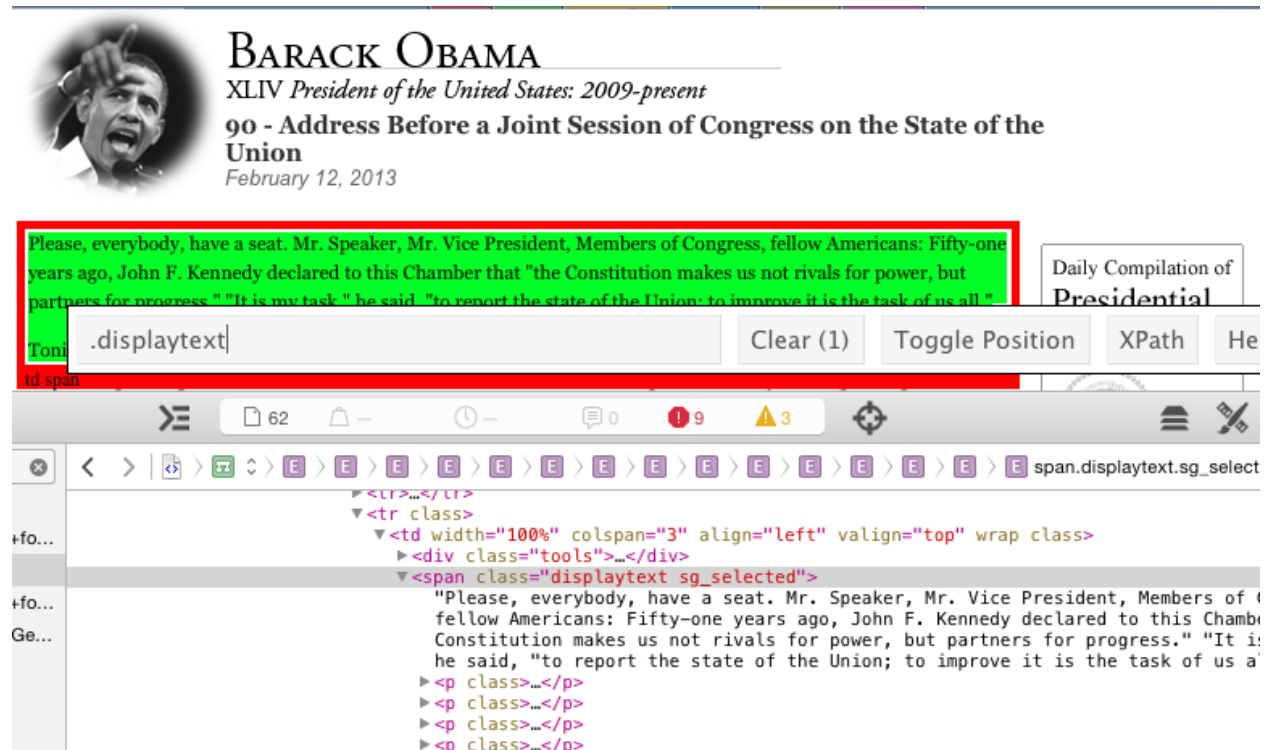
And that should do it. Looking at our directory we see that the files are now there and in order.

```
# View the `texts/` directory
dir(path = "texts", full.names = TRUE)
```

```
## [1] "texts/democrat-1947.txt" "texts/democrat-1948.txt"
## [3] "texts/democrat-1949.txt" "texts/democrat-1950.txt"
## [5] "texts/democrat-1951.txt" "texts/democrat-1952.txt"
## [7] "texts/democrat-1953.txt" "texts/democrat-1961.txt"
## [9] "texts/democrat-1962.txt" "texts/democrat-1963.txt"
## [11] "texts/democrat-1964.txt" "texts/democrat-1965.txt"
## [13] "texts/democrat-1966.txt" "texts/democrat-1967.txt"
## [15] "texts/democrat-1968.txt" "texts/democrat-1969.txt"
## [17] "texts/democrat-1978.txt" "texts/democrat-1979.txt"
## [19] "texts/democrat-1980.txt" "texts/democrat-1993.txt"
## [21] "texts/democrat-1994.txt" "texts/democrat-1995.txt"
## [23] "texts/democrat-1996.txt" "texts/democrat-1997.txt"
## [25] "texts/democrat-1998.txt" "texts/democrat-1999.txt"
## [27] "texts/democrat-2000.txt" "texts/democrat-2009.txt"
## [29] "texts/democrat-2010.txt" "texts/democrat-2011.txt"
## [31] "texts/democrat-2012.txt" "texts/democrat-2013.txt"
## [33] "texts/democrat-2014.txt" "texts/democrat-2015.txt"
## [35] "texts/republican-1954.txt" "texts/republican-1955.txt"
## [37] "texts/republican-1956.txt" "texts/republican-1957.txt"
```

```
## [39] "texts/republican-1958.txt" "texts/republican-1959.txt"
## [41] "texts/republican-1960.txt" "texts/republican-1970.txt"
## [43] "texts/republican-1971.txt" "texts/republican-1972.txt"
## [45] "texts/republican-1974.txt" "texts/republican-1975.txt"
## [47] "texts/republican-1976.txt" "texts/republican-1977.txt"
## [49] "texts/republican-1981.txt" "texts/republican-1982.txt"
## [51] "texts/republican-1983.txt" "texts/republican-1984.txt"
## [53] "texts/republican-1985.txt" "texts/republican-1986.txt"
## [55] "texts/republican-1987.txt" "texts/republican-1988.txt"
## [57] "texts/republican-1989.txt" "texts/republican-1990.txt"
## [59] "texts/republican-1991.txt" "texts/republican-1992.txt"
## [61] "texts/republican-2001.txt" "texts/republican-2002.txt"
## [63] "texts/republican-2003.txt" "texts/republican-2004.txt"
## [65] "texts/republican-2005.txt" "texts/republican-2006.txt"
## [67] "texts/republican-2007.txt" "texts/republican-2008.txt"
```

A note is in order on isolating the text on each SOTU page. Selectorgadget is really handy, but in my experience it isn't fool proof. If you cannot get the highlighting to work, you will need to open up the html page source and do some sleuthing. In Safari on OSX, you will need to enable "Show Develop in menu bar" and then you can choose "Show Web Inspector". Perusing the html structure you need to use some trial and error to find the CSS selector(s) that work. After some poking around, `.displaytext` turns out to do the trick.



The screenshot shows a web browser displaying the SOTU page for Barack Obama. The page title is "BARACK OBAMA" and the subtitle is "XLIV President of the United States: 2009-present". The main heading is "90 - Address Before a Joint Session of Congress on the State of the Union" dated "February 12, 2013". The text of the address is visible, starting with "Please, everybody, have a seat. Mr. Speaker, Mr. Vice President, Members of Congress, fellow Americans: Fifty-one years ago, John F. Kennedy declared to this Chamber that 'the Constitution makes us not rivals for power, but partners for progress.' 'It is my task,' he said, 'to report the state of the Union; to improve it is the task of us all.'" The Selectorgadget tool is overlaid on the page, showing the CSS selector `.displaytext` and the corresponding HTML structure. The HTML structure shows a `tr` element with a `td` element containing a `div` element with the class `displaytext`.

```
sessionInfo()
```

```
## R version 3.1.3 (2015-03-09)
## Platform: x86_64-apple-darwin13.4.0 (64-bit)
## Running under: OS X 10.10.2 (Yosemite)
##
## locale:
```

```
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] graphics grDevices utils      datasets stats      methods  base
##
## other attached packages:
## [1] rvest_0.2.0  EBIImage_4.8.2 kfigr_1.1.0  Rdym_0.2.0  ggplot2_1.0.1
##
## loaded via a namespace (and not attached):
## [1] abind_1.4-0      BiocGenerics_0.12.1 bitops_1.0-6
## [4] colorspace_1.2-5 digest_0.6.8      evaluate_0.5.5
## [7] formatR_1.1      grid_3.1.3       gtable_0.1.2
## [10] htmltools_0.2.6  httr_0.6.1       jpeg_0.1-8
## [13] knitr_1.9        lattice_0.20-30  locfit_1.5-9.1
## [16] magrittr_1.5     MASS_7.3-39      munsell_0.4.2
## [19] parallel_3.1.3   plyr_1.8.1       png_0.1-7
## [22] proto_0.3-10     Rcpp_0.11.5      RCurl_1.95-4.5
## [25] reshape2_1.4.1   rmarkdown_0.5.1  scales_0.2.4
## [28] selectr_0.2-3    stringr_0.6.2    tcltk_3.1.3
## [31] tiff_0.1-5       tools_3.1.3      XML_3.98-1.1
## [34] yaml_2.1.13
```