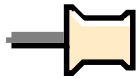
Introduction to Webscraping with R

By Robert Vesco

>> For Access to R Code, Please Open this Presentation in Dedicated PDF Application and Click on Pin <<



Outline

Introduction to Webscraping with R

By Robert Vesco

Why Use R for Webscraping

Why XML, XPATH Approach

The Basics of Webscraping

R Example

RCurl

Practical Advice

References

1 Why Use R for Webscraping

2 Why XML, XPATH Approach

3 The Basics of Webscraping

4 R Example

5 RCurl

6 Practical Advice

7 References

Why Use R for Webscraping

Introduction to Webscraping with R

By Robert Vesco

Why Use R for Webscraping

Why XML, XPATH Approach

The Basics of Webscraping

R Example

RCurl

Practical Advice

References

- No external languages or scripts needed
- Makes workflow more efficient
- Easier to share with othe[R] colleagues
- Can accomplish most webscraping needs quickly and efficiently

Outline

Introduction to Webscraping with R

By Robert Vesco

Why Use R for Webscraping

Why XML, XPATH Approach

The Basics of Webscraping

R Example

RCurl

Practical Advice

References

1 Why Use R for Webscraping

2 Why XML, XPATH Approach

3 The Basics of Webscraping

4 R Example

5 RCurl

6 Practical Advice

7 References

Why XML, XPATH Approach

Introduction to Webscraping with R

By Robert Vesco

Why Use R for Webscraping

Why XML, XPATH Approach

The Basics of Webscraping

R Example

RCurl

Practical Advice

References

- Faster than using regular expressions
- More robust
- Nearly all languages now support XPATH approach
- HTML code in the wild getting better all the time – and hence makes XPATH more reliable
- Can and should be used with regular expressions

Outline

Introduction to Webscraping with R

By Robert Vesco

Why Use R for Webscraping

Why XML, XPATH Approach

The Basics of Webscraping

R Example

RCurl

Practical Advice

References

1 Why Use R for Webscraping

2 Why XML, XPATH Approach

3 The Basics of Webscraping

4 R Example

5 RCurl

6 Practical Advice

7 References

HTML - The Code Behind the Web

Introduction to Webscraping with R

By Robert Vesco

Why Use R for Webscraping

Why XML, XPATH Approach

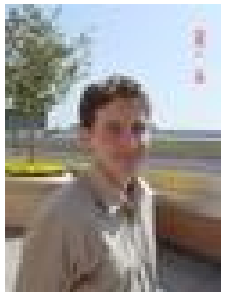
The Basics of Webscraping

R Example

RCurl

Practical Advice

References



Robert V

Joined: April 12, 2010

"1) I'm a love[r] not a hate[r] 2)

```
<li id="member_4403063">
  <div id="image_4403063" class="profphoto">
    <a href="http://www.meetup.com/R-users-DC/members/4403063/">
    <img class="memPhoto" alt=""
    src="http://photos3.meetupstatic.com/photos/member/c/a/c/7/thumb_2031911.jpeg"></a>
  </div>
  <div class="memberItem">
    <div class="memberInfo">
      <div id="memberInfo_4403063" class="D_title ">
        <a class="memName" href="http://www.meetup.com/R-users-DC/members/4403063/">Robert V</a>
        <span id="role_4403063" class="memRole">
        </span>
  <ul class="D_less memStats">
<li>
  <span class="bold">Joined</span>: April 12, 2010
</li>
</ul>
<p class="D_less">
"1) I'm a love[r] not a hate[r]
2)
```

Using XPATH to Select Items

Introduction to Webscraping with R

By Robert Vesco

Why Use R for Webscraping

Why XML, XPATH Approach

The Basics of Webscraping

R Example

RCurl

Practical Advice

References

We want to select just my name

```
.//*[@class='memName']
```

or use fuller path

```
.//*div/div/div/ etc .... /a[@class='memName']
```

Will both pull out "Robert V" from this code:

```
<a class="memName" href="http://www.meetup.com/R-users-DC/members/4403063/">Robert V</a>
```

BUT - more importantly, the above code will also pull out every name if you're looking at all the code on the page!

XPATH Tutorials

Introduction to Webscraping with R

By Robert Vesco

Why Use R for Webscraping

Why XML, XPATH Approach

The Basics of Webscraping

R Example

RCurl

Practical Advice

References

- http://www.zvon.org/xxl/XPathTutorial/General/examples.html
- http://www.w3schools.com/xpath/

Important XML Functions in R

Introduction to Webscraping with R
By Robert Vesco
Why Use R for Webscraping
Why XML, XPATH Approach
The Basics of Webscraping
R Example
RCurl
Practical Advice
References

htmlTreeParse() parses file,cleans malformed HTML, and make it available for querying

```
web <- htmlTreeParse(chrFileUrl,error=function(...){}, useInternalNodes =↩
    TRUE,trim=TRUE)
```

getNodeset() takes xml documents and queries it for particular nodes (html tags)

```
nArticles <- getNodeSet(web,"//*/div[@class='D_title ']")
```

xmlValue() takes nodes from getNodeSet and extracts text value

```
xmlValue(nArticles[[1]])
```

xmlGetAttr() get attributes from html tags such as href links or class names

```
xmlGetAttr(nArticles[[1]],"href")
```

Lastly, there exists convenience functions such as xpathSApply()

Some Tools - Firebug + FireXPATH

Introduction to Webscraping with R

By Robert Vesco

Why Use R for Webscraping

Why XML, XPATH Approach

The Basics of Webscraping

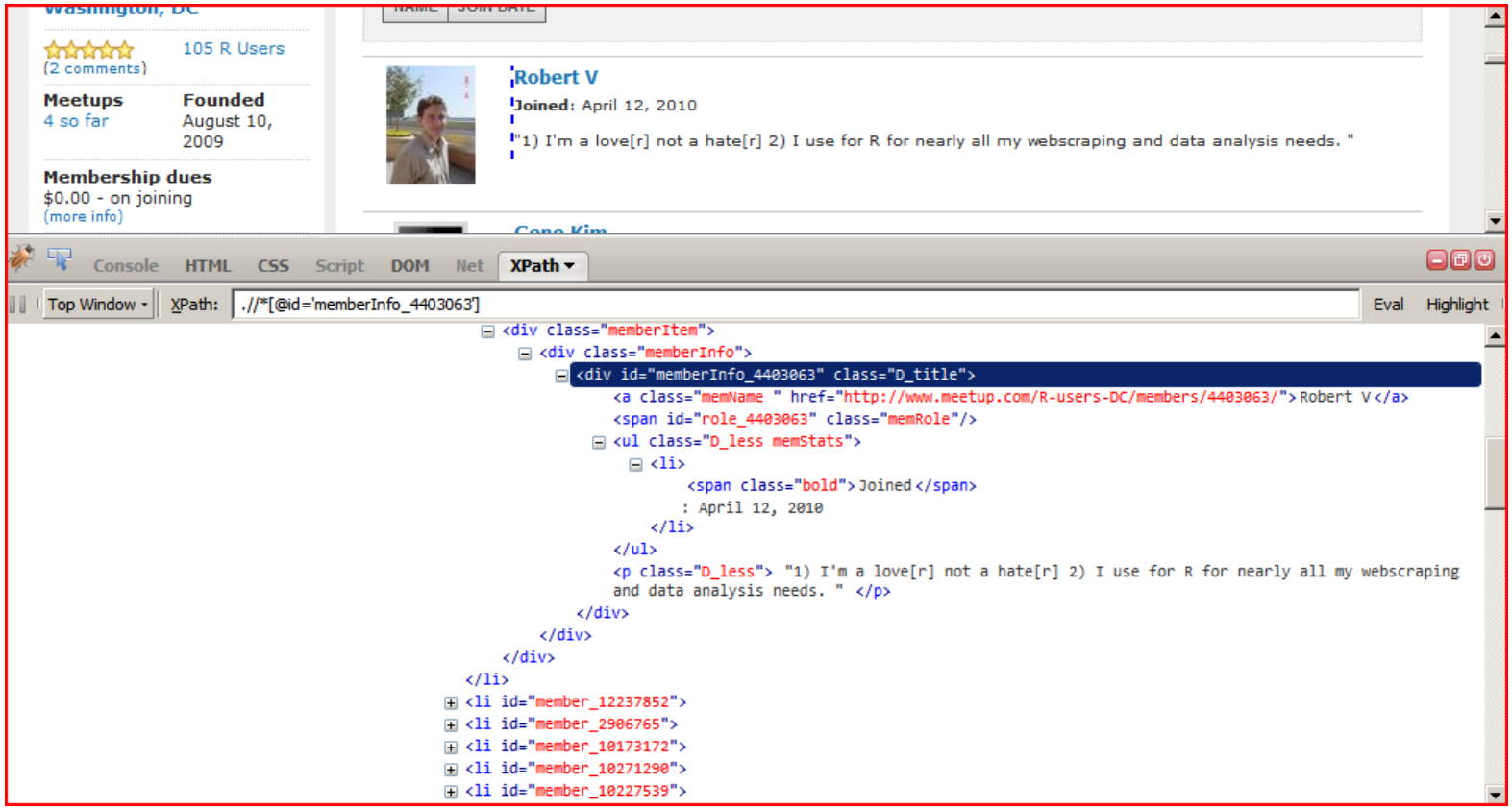
R Example

RCurl

Practical Advice

References

- Allows you to select items on a webpage and inspect their underlying tags. Also allows you to query your XPATH to see what it will select!



Some Tools - Selector Gadget

Introduction to Webscraping with R

By Robert Vesco

Why Use R for Webscraping

Why XML, XPATH Approach

The Basics of Webscraping

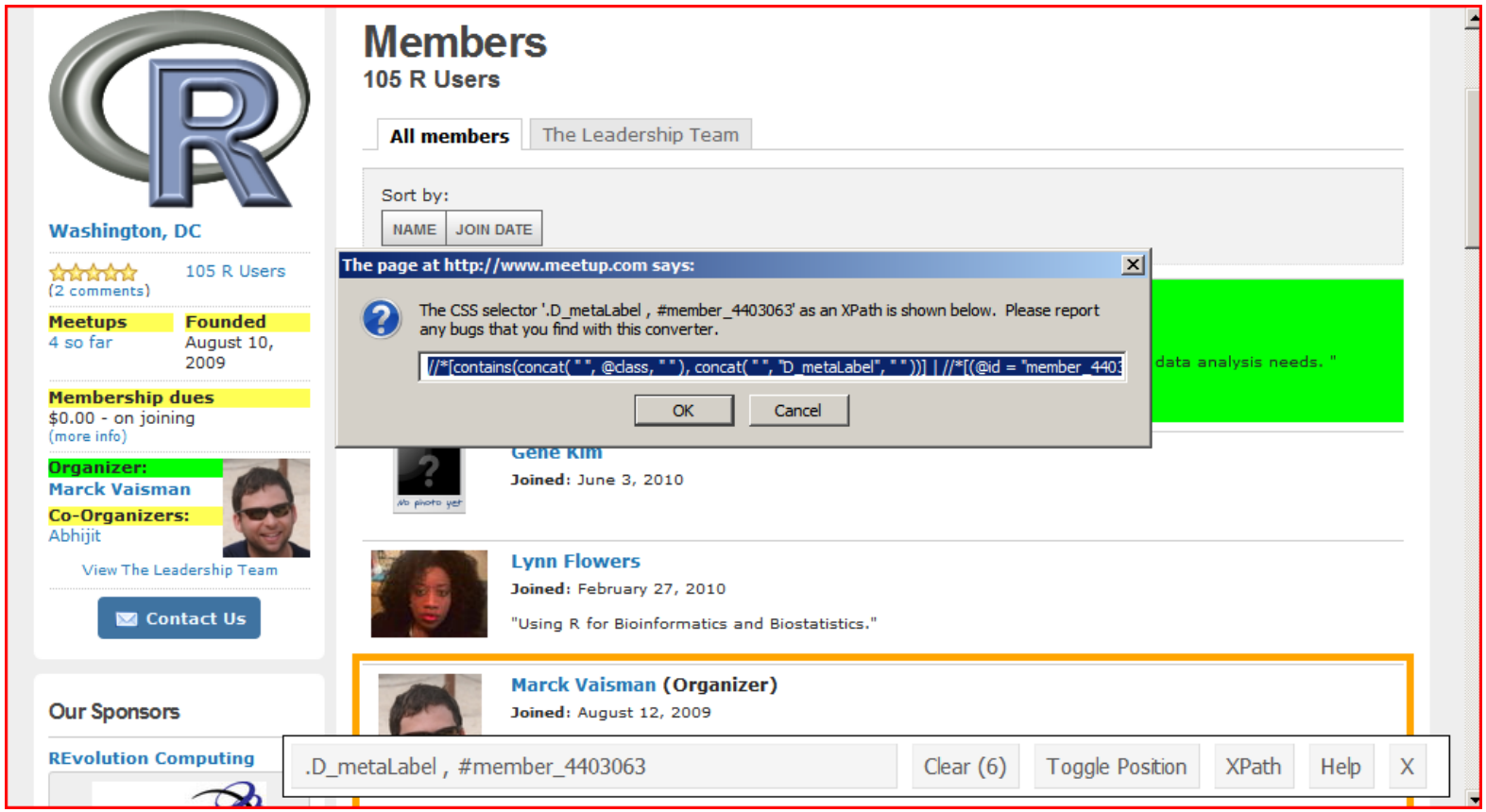
R Example

RCurl

Practical Advice

References

- Allows you to select multiple elements on a screen
- Useful for very complicated layouts



Outline

Why Use R for Webscraping

Why XML, XPATH Approach

The Basics of Webscraping

R Example

RCurl

Practical Advice

References

1 Why Use R for Webscraping

2 Why XML, XPATH Approach

3 The Basics of Webscraping

4 R Example

5 RCurl

6 Practical Advice

7 References

General Steps

Introduction to Webscraping with R

By Robert Vesco

Why Use R for Webscraping

Why XML, XPATH Approach

The Basics of Webscraping

R Example

RCurl

Practical Advice

References

- Select and download the pages you want
- Query the document
- Select which items you want and save to dataframe
- Repeat

Find Which Pages Have the Info You Need + Download

Introduction to Webscraping with R

By Robert Vesco

Why Use R for Webscraping

Why XML, XPATH Approach

The Basics of Webscraping

R Example

RCurl

Practical Advice

References


```r
#Seq Variables
cStartSeq <- 0
cEndSeq <- 100
cStepSeq <- 20
#link variables
chrURLPrefix <- "http://www.meetup.com/R-users-DC/members/?offset="
chrURLSuffix <- "&desc=1&sort=chapter_member.atime"
#Files will be read and save to this path. Make sure it exists on your computer
#or change path to wherever you save this script to!!
chrSetDir = "~/R/R_Meetup/"
#setwd(chrSetDir) #commented out for sweave
chrDir <- paste(chrSetDir,"RawData/",sep="")
#Check to see if folder exists, else create it.
if("RawData" %in% dir(chrSetDir)==FALSE){dir.create(chrDir)}
for(w in seq(cStartSeq,cEndSeq,cStepSeq))
{
 #Create URL that will download page
 url <- paste(chrURLPrefix,w,chrURLSuffix,sep="")
 #Create name for URL. Important because sometimes URL names have illegal ↩
      characters
 #or lengths for files systems.
 urlName <- paste(chrDir,w,".html",sep="")
 #Without error catching script will crash. Websites frequently time out!
 err <- try(download.file(url,destfile = urlName,quiet = TRUE), silent = TRUE)
 if(class(err)=="try-error")
 {
  #you may be hitting the server too hard, so backoff and try again later.
  Sys.sleep(5) #in seconds, adjust as necessary
  try(download.file(url,destfile = urlName,quiet = TRUE), silent = TRUE)
 }
}
```

Process Pages and Extract Data

Introduction to Webscraping with R

By Robert Vesco

Why Use R for Webscraping

Why XML, XPATH Approach

The Basics of Webscraping

R Example

RCurl

Practical Advice

References


```
require(XML)
require(xtable)
vFiles ← list.files(chrDir)# put files in rawdata folder into vector and get length
iLenFilesList ← length(vFiles) #create list to store dataframes
ls ← list()
for(i in 1:iLenFilesList)
{
 #each i will pull a different URL
 url ← vFiles[i]
    chrFileUrl ← paste(chrDir, url,sep="")
 # this function works on dirty html, adding closing tags and such.
 web ← htmlTreeParse(chrFileUrl,error=function(...){}, useInternalNodes = TRUE,encoding = "UTF-8",trim=TRUE)
 #Use vectorized function to get names
 vNames ← xpathSApply(web,'//*[@class="memName"]',xmlValue)
 #Same as above, but use regex to clean up a bit
 vDates ← gsub("Joined: |\r\n","",xpathSApply(web,'//*/ul[@class="D_less memStats"]/li',xmlValue))
 #Since not every person has a quote, we break the problem into part getting chunks of code
 vQuote2 ← getNodeSet(web,"//*/div[@class='D_title']")
 #now we look for quotes -  notice ".//" this means subquery -- IMPORTANT!
 vQuote3 ← sapply(vQuote2,function(x) xpathSApply(x,".//p[@class='D_less']",xmlValue))
 # we get list() for node with no quotes. Replace list() with NULL
 vQuote4  ← gsub('\r|\n|["]',"",sapply(vQuote3,function(x) ifelse(is.list(x),NA,x)))
 #add df to list. This is ok for small scrapes, but for larges ones, you need to write to file or db
 ls[[i]] ← data.frame(Name=vNames,Date=vDates,Quote=vQuote4,stringsAsFactors=FALSE)
}
```

Combine Data from Each Loop

Introduction to Webscraping with R

By Robert Vesco

Why Use R for Webscraping

Why XML, XPATH Approach

The Basics of Webscraping

R Example

RCurl

Practical Advice

References

```
#combine df
df <- do.call(rbind,ls)
#sample output for latex
library(Hmisc)
df2 <- df[1:3,]
latex(df2, file='', col.just=c("l","l","p{2in}"))
```

df2	Name	Date	Quote
1	JOEL ROBERTS	June 9, 2010	My name is Joel Roberts. I am a friend of Bryan Stroube who told me about the DC useR Group. I work with several systems that use XML for information exchange between dissimilar computer / software systems.
2	Arun	May 1, 2010	NA
3	Travis M	April 16, 2010	NA

Outline

Introduction to Webscraping with R

By Robert Vesco

Why Use R for Webscraping

Why XML, XPATH Approach

The Basics of Webscraping

R Example

RCurl

Practical Advice

References

1 Why Use R for Webscraping

2 Why XML, XPATH Approach

3 The Basics of Webscraping

4 R Example

5 RCurl

6 Practical Advice

7 References

RCurl Package

Introduction to Webscraping with R

By Robert Vesco

Why Use R for Webscraping

Why XML, XPATH Approach

The Basics of Webscraping

R Example

RCurl

Practical Advice

References

- More flexible, allows one to modify headers, referers, etc...
- Beyond http: https, ftp, sftp, scp, ldap, etc.....
- Come from c library libcurl so fast, extensive, and actively developed
- Can use persistent connections, cookies, and process requests as they come in rather than sequentially

Some RCurl Examples

Introduction to Webscraping with R

By Robert Vesco

Why Use R for Webscraping

Why XML, XPATH Approach

The Basics of Webscraping

R Example

RCurl

Practical Advice

References

getURL() allows the use of https whereas built-in R functions do not (recent R versions may be different, internet2 method must have valid cert)

```
1 txt = getURL("https://www.twitter.com",ssl.verifyhost=0, ssl.verifypeer=0)
```

getCurlHandle() handles allow persistent connections and settings to be used across repeated call to same server which is similar to pass list of arguments, but potentially with better network connectivity.

```
1 curl = getCurlHandle(cookie=cookie, useragent= "Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.8.1.6) Gecko/20070725 Firefox/2.0.0.6")
2 txt = getURL("http://www.meetup.com/R-users-DC/members/",curl=curl,.opts = list(verbose = TRUE))
3 txt2 = getURL("http://www.meetup.com/R-users-DC/members/",curl=curl,.opts = list(verbose = TRUE))
```

Outline

Introduction to Webscraping with R

By Robert Vesco

Why Use R for Webscraping

Why XML, XPATH Approach

The Basics of Webscraping

R Example

RCurl

Practical Advice

References

1 Why Use R for Webscraping

2 Why XML, XPATH Approach

3 The Basics of Webscraping

4 R Example

5 RCurl

6 Practical Advice

7 References

Warning and Practical Advice

Introduction to Webscraping with R

By Robert Vesco

Why Use R for Webscraping

Why XML, XPATH Approach

The Basics of Webscraping

R Example

RCurl

Practical Advice

References

- Be Nice! Go easy on the servers – especially if you're using rCURL
- Check to make sure website does not prohibit scrapping. See if they already have an API. Else, send an email to web owners.
- If login required, then need to use rCurl package.
- Always use a proxy - if for nothing else you don't want your home address to accidently get blocked (ie google automated query blocking).

Warning and Practical Advice

Introduction to Webscraping with R

By Robert Vesco

Why Use R for Webscraping

Why XML, XPATH Approach

The Basics of Webscraping

R Example

RCurl

Practical Advice

References

- Consider setting up your old computer or laptop to run scraping. Frees your main computer up. Set it to email you if it crashes.....
- Get familiar with error catching and debugging
- Even if you try to make your code robust, large jobs will require several retweakings of code, because ex ante, you don't know all the possible permutations. Your code will become more generalized with more exception handling.
- XML packages have slightly different interpretations. Hence going from one language to the next may require 'slightly' different queries

Outline

Introduction to Webscraping with R

By Robert Vesco

Why Use R for Webscraping

Why XML, XPATH Approach

The Basics of Webscraping

R Example

RCurl

Practical Advice

References

1 Why Use R for Webscraping

2 Why XML, XPATH Approach

3 The Basics of Webscraping

4 R Example

5 RCurl

6 Practical Advice

7 References

References

Introduction to Webscraping with R

By Robert Vesco

Why Use R for Webscraping

Why XML, XPATH Approach

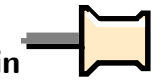
The Basics of Webscraping

R Example

RCurl

Practical Advice

References

- **R code used in this presentation, click on pin** 
- **Tutorials,**
  `http://www.zvon.org/xxl/XPathTutorial/General/examples.html`
  `http://www.w3schools.com/xpath/`
- **Firebug**
  `http://getfirebug.com/`
  `https://addons.mozilla.org/en-US/firefox/addon/11900/`
- **SelectorGadget**
  `http://www.selectorgadget.com/`
- **XML package docs**
  `http://www.omegahat.org/RSXML/`
- **rCurl Package**
  `http://www.omegahat.org/RCurl/`