Northeastern University

CS6020: Collecting, Storing, and Retrieving Information

Data Collection Through Web Scraping

Data Collection Through Web Scraping

WEB SCRAPING IN R

Web Scraping in R

APPROACH & PACKAGES

Lesson Objectives

- After completing this lesson, you are able to:
  - understand the structure of an HTML document
  - recognize data in an HTML document
  - programmatically extract data from an HTML document

Required Libraries

- The following R libraries need to be loaded:
  - ***RCurl***
  - ***XML***
  - ***scrapeR***
- Be sure to install first the packages if they have not yet been installed.

Web Scraping Packages in R

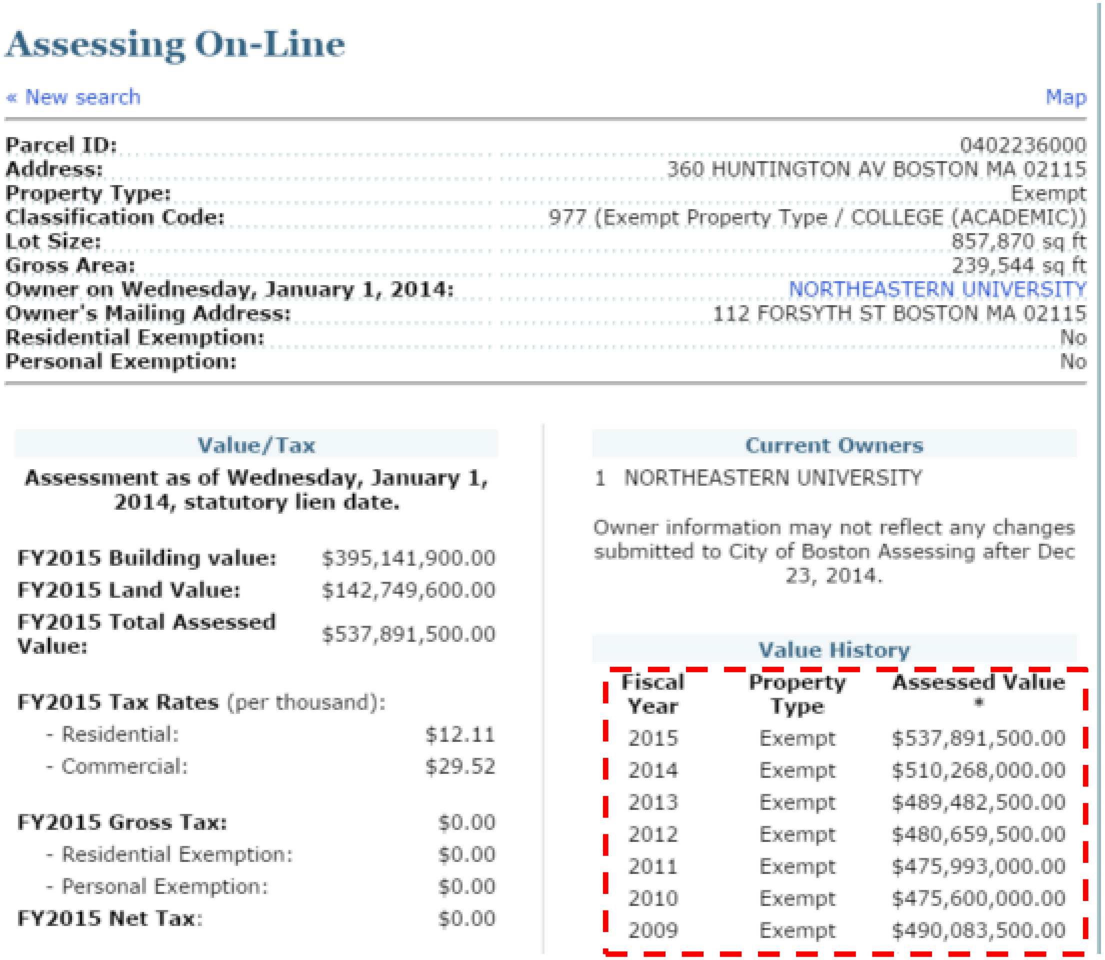
- RCurl
  - The RCurl package is an R-interface to the libcurl library that provides HTTP facilities
  - This allows us to download files from Web servers by GETting forms
  - The primary top-level entry points are : `getURL()`, `getURLContent()`
- XML
  - The XML package is necessary to parse the XML and HTML code
  - This also offers access to an *XPath* "interpreter"
- scrapeR
  - The scrapeR package is necessary to extract the data from the XML and HTML documents
  - Provides a function `scrape()` that assists the user with retrieving HTML and XML files, parsing their contents and diagnosing potential errors that may occur along the way

Case Study: Property Tax History

- Let’s say that we need historical tax information for properties in Boston.
- This data is available through the web at www.cityofboston.gov, although the city does not provide the data for download.
- We will build an R script that “scrapes” the needed data from the relevant web page on the web site for the desired property.

The Web Page

http://www.cityofboston.gov/assessing/search/?pid=0402236000



Assessing On-Line

« New search Map

Parcel ID:	0402236000
Address:	360 HUNTINGTON AV BOSTON MA 02115
Property Type:	Exempt
Classification Code:	977 (Exempt Property Type / COLLEGE (ACADEMIC))
Lot Size:	857,870 sq ft
Gross Area:	239,544 sq ft
Owner on Wednesday, January 1, 2014:	NORTHEASTERN UNIVERSITY
Owner's Mailing Address:	112 FORSYTH ST BOSTON MA 02115
Residential Exemption:	No
Personal Exemption:	No

Value/Tax

Assessment as of Wednesday, January 1, 2014, statutory lien date.

FY2015 Building value:	$395,141,900.00
FY2015 Land Value:	$142,749,600.00
FY2015 Total Assessed Value:	$537,891,500.00
FY2015 Tax Rates (per thousand):	
- Residential:	$12.11
- Commercial:	$29.52
FY2015 Gross Tax:	$0.00
- Residential Exemption:	$0.00
- Personal Exemption:	$0.00
FY2015 Net Tax:	$0.00

Current Owners

1 NORTHEASTERN UNIVERSITY

Owner information may not reflect any changes submitted to City of Boston Assessing after Dec 23, 2014.

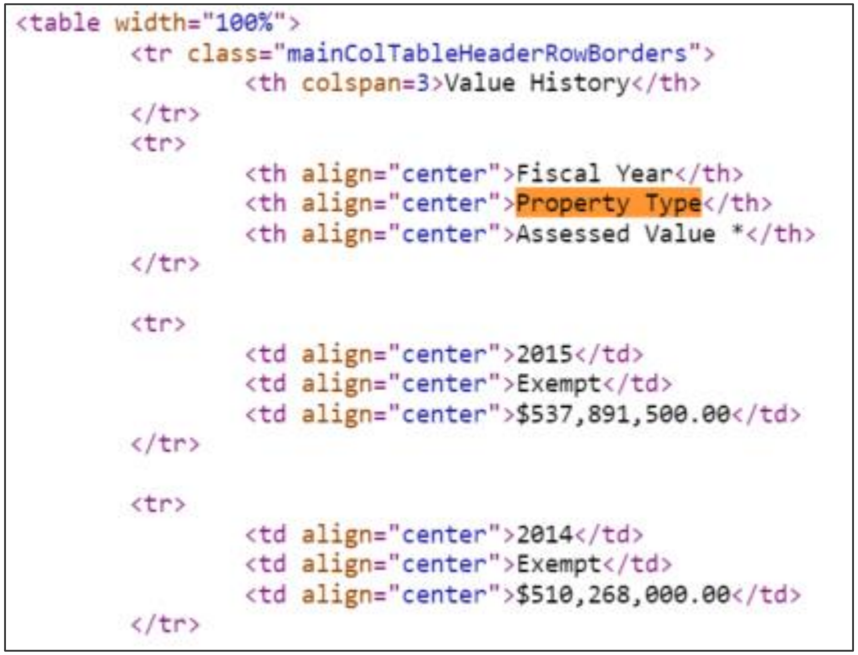
Value History

Fiscal Year	Property Type	Assessed Value *
2015	Exempt	$537,891,500.00
2014	Exempt	$510,268,000.00
2013	Exempt	$489,482,500.00
2012	Exempt	$480,659,500.00
2011	Exempt	$475,993,000.00
2010	Exempt	$475,600,000.00
2009	Exempt	$490,083,500.00

This is the data we would like to extract. Note the text labels around the data. We need to look for that "pattern" in the HTML document that defines this page.

The HTML Code

- This is the relevant section of the HTML code[1].



```
<table width="100%">
        <tr class="mainColTableHeaderRowBorders">
                <th colspan=3>Value History</th>
        </tr>
        <tr>
                <th align="center">Fiscal Year</th>
                <th align="center">Property Type</th>
                <th align="center">Assessed Value *</th>
        </tr>

        <tr>
                <td align="center">2015</td>
                <td align="center">Exempt</td>
                <td align="center">$537,891,500.00</td>
        </tr>

        <tr>
                <td align="center">2014</td>
                <td align="center">Exempt</td>
                <td align="center">$510,268,000.00</td>
        </tr>
```

You can locate the relevant section of HTML by using the Search function in your browser or text editor (generally CTRL-F.)

[1] You can get the HTML code for any web page by right-clicking on the page in your browser and selecting "View Page Source" or a similar choice depending on the browser.

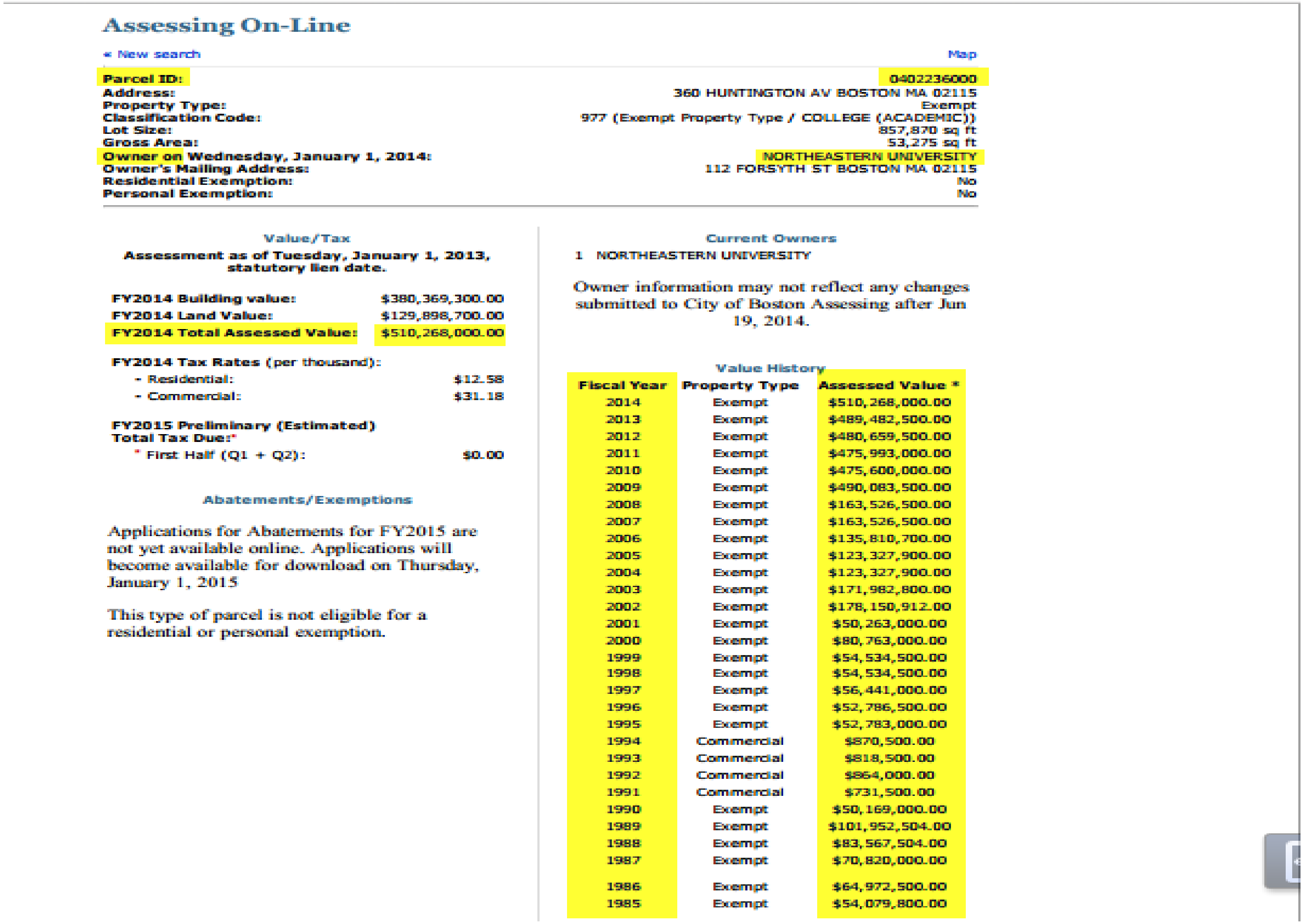
Web Scraping in R

A WORKED EXAMPLE

Web Scraping Example

- In order to explain web scraping we will consider the example of scraping the following website to extract some useful data on Northeastern University.
- Website to be scraped : http://www.cityofboston.gov/assessing/search/?pid=0402236000
- As the entire data available on the page is not important to us we will scrape only the highlighted portions of this page.

Web Scraping Example



Assessing On-Line

« New search

Map

Parcel ID:	0402236000
Address:	360 HUNTINGTON AV BOSTON MA 02115
Property Type:	Exempt
Classification Code:	977 (Exempt Property Type / COLLEGE (ACADEMIC))
Lot Size:	857,870 sq ft
Gross Area:	53,275 sq ft
Owner on Wednesday, January 1, 2014:	NORTHEASTERN UNIVERSITY
Owner's Mailing Address:	112 FORSYTH ST BOSTON MA 02115
Residential Exemption:	No
Personal Exemption:	No

Value/Tax

Assessment as of Tuesday, January 1, 2013, statutory lien date.

FY2014 Building value:	$380,369,300.00
FY2014 Land Value:	$129,898,700.00
FY2014 Total Assessed Value:	$510,268,000.00

FY2014 Tax Rates (per thousand):

- Residential:	$12.58
- Commercial:	$31.18

FY2015 Preliminary (Estimated) Total Tax Due:*

* First Half (Q1 + Q2):	$0.00

Abatements/Exemptions

Applications for Abatements for FY2015 are not yet available online. Applications will become available for download on Thursday, January 1, 2015

This type of parcel is not eligible for a residential or personal exemption.

Current Owners

1 NORTHEASTERN UNIVERSITY

Owner information may not reflect any changes submitted to City of Boston Assessing after Jun 19, 2014.

Value History

Fiscal Year	Property Type	Assessed Value *
2014	Exempt	$510,268,000.00
2013	Exempt	$489,482,500.00
2012	Exempt	$480,659,500.00
2011	Exempt	$475,993,000.00
2010	Exempt	$475,600,000.00
2009	Exempt	$490,083,500.00
2008	Exempt	$163,526,500.00
2007	Exempt	$163,526,500.00
2006	Exempt	$135,810,700.00
2005	Exempt	$123,327,900.00
2004	Exempt	$123,327,900.00
2003	Exempt	$171,982,800.00
2002	Exempt	$178,150,912.00
2001	Exempt	$50,263,000.00
2000	Exempt	$80,763,000.00
1999	Exempt	$54,534,500.00
1998	Exempt	$54,534,500.00
1997	Exempt	$56,441,000.00
1996	Exempt	$52,786,500.00
1995	Exempt	$52,783,000.00
1994	Commercial	$870,500.00
1993	Commercial	$818,500.00
1992	Commercial	$864,000.00
1991	Commercial	$731,500.00
1990	Exempt	$50,169,000.00
1989	Exempt	$101,952,504.00
1988	Exempt	$83,567,504.00
1987	Exempt	$70,820,000.00
1986	Exempt	$64,972,500.00
1985	Exempt	$54,079,800.00

Step-by-Step Web Scraping in R

- Step 1: *Get the web page via its URL*
- Step 2: *Parse the HTML that defines the page*
- Step 3: *Extract leaf items which is the data*
- Step 4: *Clean the extracted data*

Step 1: Get the web page

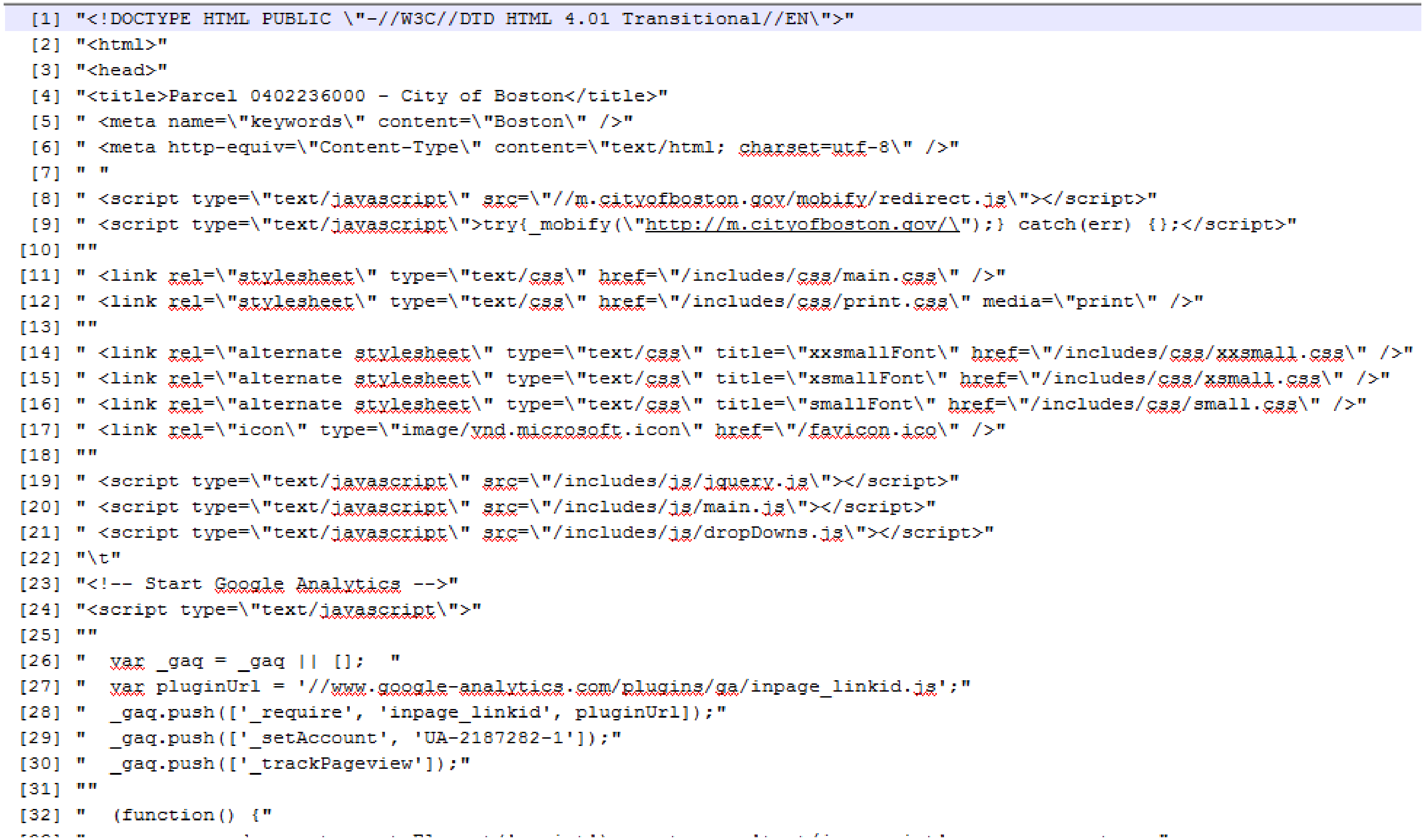
- Download the raw HTML content of the webpage using the these two functions :

```
> webpage <- getURL(URLPath)
> webpage <- readLines(tc <- textConnection(webpage));
```

- These functions fetch the entire HTML page into a parsable object.

Steps of Web Scraping in R

- Output of fetching a web page using `readLines()` in R



```
 [1] "<!DOCTYPE HTML PUBLIC \"-//W3C//DTD HTML 4.01 Transitional//EN\">"
 [2] "<html>"
 [3] "<head>"
 [4] "<title>Parcel 0402236000 - City of Boston</title>"
 [5] " <meta name=\"keywords\" content=\"Boston\" />"
 [6] " <meta http-equiv=\"Content-Type\" content=\"text/html; charset=utf-8\" />"
 [7] " "
 [8] " <script type=\"text/javascript\" src=\"//m.cityofboston.gov/mobify/redirect.js\"></script>"
 [9] " <script type=\"text/javascript\">try{_mobify(\"http://m.cityofboston.gov/\");} catch(err) {};</script>"
[10] ""
[11] " <link rel=\"stylesheet\" type=\"text/css\" href=\"/includes/css/main.css\" />"
[12] " <link rel=\"stylesheet\" type=\"text/css\" href=\"/includes/css/print.css\" media=\"print\" />"
[13] ""
[14] " <link rel=\"alternate stylesheet\" type=\"text/css\" title=\"xxsmallFont\" href=\"/includes/css/xxsmall.css\" />"
[15] " <link rel=\"alternate stylesheet\" type=\"text/css\" title=\"xsmallFont\" href=\"/includes/css/xsmall.css\" />"
[16] " <link rel=\"alternate stylesheet\" type=\"text/css\" title=\"smallFont\" href=\"/includes/css/small.css\" />"
[17] " <link rel=\"icon\" type=\"image/vnd.microsoft.icon\" href=\"/favicon.ico\" />"
[18] ""
[19] " <script type=\"text/javascript\" src=\"/includes/js/jquery.js\"></script>"
[20] " <script type=\"text/javascript\" src=\"/includes/js/main.js\"></script>"
[21] " <script type=\"text/javascript\" src=\"/includes/js/dropDowns.js\"></script>"
[22] "\t"
[23] "<!-- Start Google Analytics -->"
[24] "<script type=\"text/javascript\">"
[25] ""
[26] "  var _gaq = _gaq || [];  "
[27] "  var pluginUrl = '//www.google-analytics.com/plugins/ga/inpage_linkid.js';"
[28] "  _gaq.push(['_require', 'inpage_linkid', pluginUrl]);"
[29] "  _gaq.push(['_setAccount', 'UA-2187282-1']);"
[30] "  _gaq.push(['_trackPageview']);"
[31] ""
[32] "  (function() {"
```

Step 2: Parse the Data

- Transform raw HTML into a more convenient format to work with using `htmlTreeParse()`.

```
pagetree <- htmlTreeParse(webpage,
                          useInternalNodes = TRUE)
```

- Setting `useInternalNodes=TRUE` allows one to access the parent and ancestor nodes.

Step 3: Extract Leaf Items

- Use `xpathApply()` to extract the leaf items in the HTML document:

```
x <- unlist(xpathApply(pagetree,
                "//*/table[@width='100%']/tr[2]/
                 th[@align='center']", xmlValue))
```

- To eliminate undesired matches, the query restricts the high level table attribute to `width=100%` and table heading attribute aligned to center.
- `xmlValue` is convenient for extracting the text value of the node.

Step 4: Clean the Data

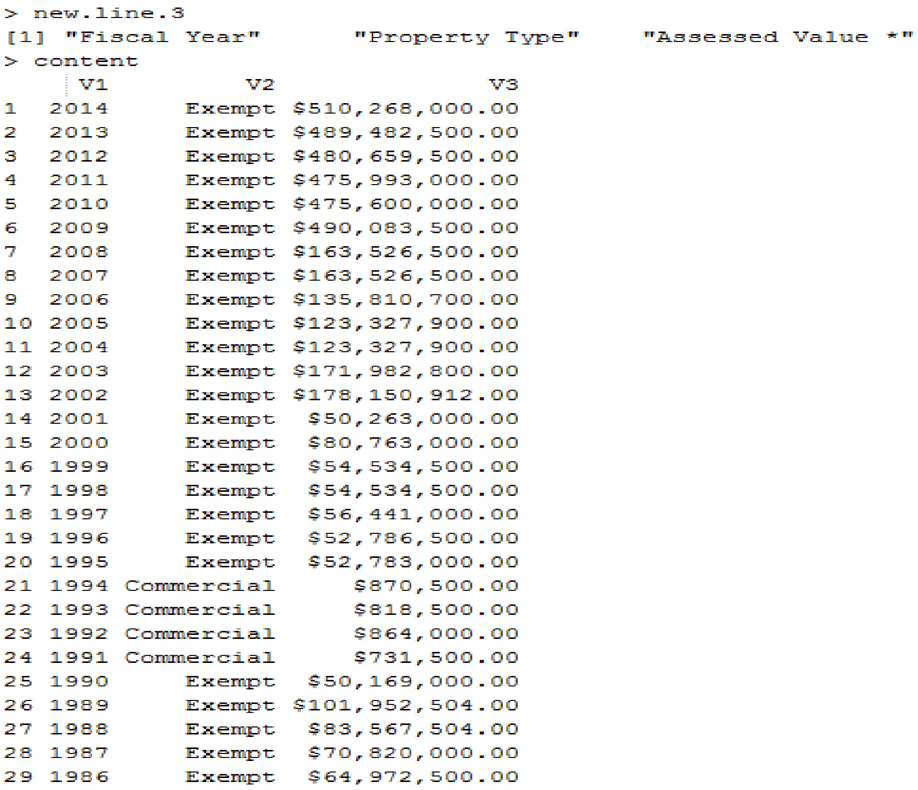
- Example:

```
Content <- gsub(pattern = "([\t\n])",
       replacement = " ", x = x, ignore.case = TRUE)
```

- The R global substitution function `gsub()` changes the “\t\n” combination to an empty string(“”)

Web Scraping in R

- Cleaned up data



```
> new.line.3
[1] "Fiscal Year"      "Property Type"    "Assessed Value *"
> content
     V1         V2              V3
1  2014     Exempt $510,268,000.00
2  2013     Exempt $489,482,500.00
3  2012     Exempt $480,659,500.00
4  2011     Exempt $475,993,000.00
5  2010     Exempt $475,600,000.00
6  2009     Exempt $490,083,500.00
7  2008     Exempt $163,526,500.00
8  2007     Exempt $163,526,500.00
9  2006     Exempt $135,810,700.00
10 2005     Exempt $123,327,900.00
11 2004     Exempt $123,327,900.00
12 2003     Exempt $171,982,800.00
13 2002     Exempt $178,150,912.00
14 2001     Exempt  $50,263,000.00
15 2000     Exempt  $80,763,000.00
16 1999     Exempt  $54,534,500.00
17 1998     Exempt  $54,534,500.00
18 1997     Exempt  $56,441,000.00
19 1996     Exempt  $52,786,500.00
20 1995     Exempt  $52,783,000.00
21 1994 Commercial     $870,500.00
22 1993 Commercial     $818,500.00
23 1992 Commercial     $864,000.00
24 1991 Commercial     $731,500.00
25 1990     Exempt  $50,169,000.00
26 1989     Exempt $101,952,504.00
27 1988     Exempt  $83,567,504.00
28 1987     Exempt  $70,820,000.00
29 1986     Exempt  $64,972,500.00
```

Summary

- In this lesson, you learned that:
  - web scraping can be done in R through parsing a retrieved HTML document
  - markers need to be used to identify the relevant sections of the HTML document
  - the programming fails if the HTML code changes and no longer meets the search pattern

Summary, Review, & Questions…