DataSci Cleaning Data Lecture Notes – XML

XML

* + - 1. Stands for Extensible Markup Language
      2. Used to store structured data
      3. Popular in internet applications
      4. Extraction of XML files is the basis for most web scraping
      5. Components
         1. Markup – labels that give the text structure
         2. Content – actual text of the document

XML Tags, Elements, and Attributes

1. Tags
   1. Correspond to general labels
   2. Start Tag <section>
   3. End Tags </section>
   4. Empty Tags <line-break />
2. Elements
   1. Specific examples of tags
   2. <Greeting> Hello, world! </Greeting>
3. Attributes
   1. Components to the label
   2. <img src=”jeff.jpg” alt=”instructor />
   3. <step number=”3”> Connect A to B. </step>

XML Files look like source code

Text File

Lots of Tags

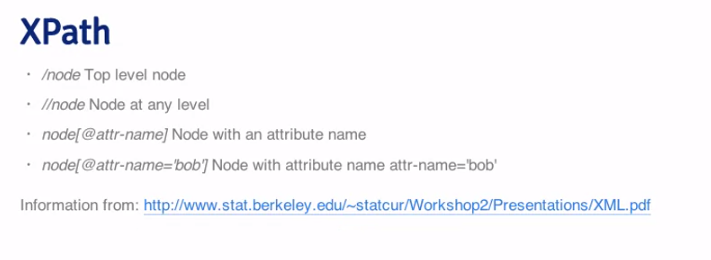
Lots of Content

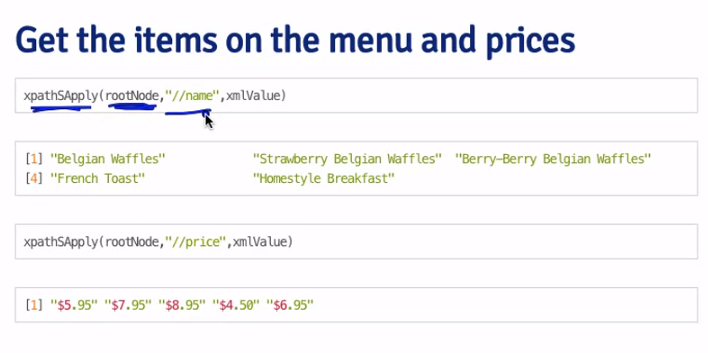
Check [www.w3schools.com/xml](http://www.w3schools.com/xml) for more

Reading XML into R

1. Load the library(XML”
2. xmlTreeParse() loads the document into R
3. Parameters fileURL,
   1. The file URl
4. useInternal
   1. if TRUE, R will use the internal tags for headers
5. rootNode <-xmlRoot(read in file)
   1. the wrapper element for the entire XML file
6. xmlName(rootNode)
   1. gives you the name of the file
7. pull out individual parts of the XML document using double brackets
8. **xmlSApply(rootnode, xmlValue)** returns a readable text of the file

Using XPath to read XML into R





See <http://www.stat.berkeley.edu/~statcur/Workshop2/Presentations/XML.pdf> for further instructions

<http://www.omegahat.org/RSXML/shortIntro.pdf>

http://www.omegahat.org/RSXML/Tour.pdf