Aquisição de Arquivos e Datasets

Lendo arquivos XML

Delermando Branquinho Filho

XML

- Extensible Markup Language
- Freqüentemente usado para armazenar dados estruturados
- Particularmente utilizado em aplicações na Internet
- Extrair XML é a base para a maioria dos raspagem web
- Componentes
- Marcação - rótulos que dão a estrutura de texto
- Conteúdo - o texto real do documento

Tags, elementos e atributos

- As etiquetas correspondem a etiquetas gerais
- Iniciar tags `<section>`
- End tags `</ section>`
- Etiquetas vazias `<linha-quebra />`
- Os elementos são exemplos específicos de tags
- `<Saudação> Olá, mundo </ Saudação>`
- Os atributos são componentes do rótulo
- `<Img src =" jeff.jpg "alt =" instrutor "/>`
- `<step number="3"> Connect A to B. </step>`

Lendo arquivos no R

```r
library(XML)
fileUrl <- "data/arquivoXML.xml"
doc <- xmlTreeParse(fileUrl,useInternal=TRUE)
rootNode <- xmlRoot(doc)
xmlName(rootNode)
```

```
## [1] "NFe"
```

```r
names(rootNode)
```

```
##      infNFe   Signature
##    "infNFe" "Signature"
```

Acessar diretamente partes do documento XML

```
rootNode[[1]]
```

```
## <infNFe Id="NFe35080599999090910270550010000000015180051273" versao="1.10">
##   <ide>
##     <cUF>35</cUF>
##     <cNF>518005127</cNF>
##     <natOp>Venda a vista</natOp>
##     <indPag>0</indPag>
##     <mod>55</mod>
##     <serie>1</serie>
##     <nNF>1</nNF>
##     <dEmi>2008-05-06</dEmi>
##     <dSaiEnt>2008-05-06</dSaiEnt>
##     <tpNF>0</tpNF>
##     <cMunFG>3550308</cMunFG>
##     <tpImp>1</tpImp>
##     <tpEmis>1</tpEmis>
##     <cDV>3</cDV>
##     <tpAmb>2</tpAmb>
##     <finNFe>1</finNFe>
##     <procEmi>0</procEmi>
##     <verProc>NF-eletronica.com</verProc>
##   </ide>
##   <emit>
##     <CNPJ>99999090910270</CNPJ>
##     <xNome>NF-e Associacao NF-e</xNome>
##     <xFant>NF-e</xFant>
##     <enderEmit>
##       <xLgr>Rua Central</xLgr>
##       <nro>100</nro>
##       <xCpl>Fundos</xCpl>
##       <xBairro>Distrito Industrial</xBairro>
##       <cMun>3502200</cMun>
##       <xMun>Angatuba</xMun>
##       <UF>SP</UF>
##       <CEP>17100171</CEP>
##       <cPais>1058</cPais>
##       <xPais>Brasil</xPais>
##       <fone>1733021717</fone>
##     </enderEmit>
##     <IE>123456789012</IE>
##   </emit>
##   <dest>
##     <CNPJ>00000000000191</CNPJ>
##     <xNome>DISTRIBUIDORA DE AGUAS MINERAIS</xNome>
##     <enderDest>
##       <xLgr>AV DAS FONTES</xLgr>
##       <nro>1777</nro>
##       <xCpl>10 ANDAR</xCpl>
##       <xBairro>PARQUE FONTES</xBairro>
##       <cMun>5030801</cMun>
##       <xMun>Sao Paulo</xMun>
##       <UF>SP</UF>
##       <CEP>13950000</CEP>
##       <cPais>1058</cPais>
##       <xPais>BRASIL</xPais>
##       <fone>1932011234</fone>
##     </enderDest>
##     <IE> </IE>
##   </dest>
##   <retirada>
##     <CNPJ>99171171000194</CNPJ>
##     <xLgr>AV PAULISTA</xLgr>
##     <nro>12345</nro>
##     <xCpl>TERREO</xCpl>
##     <xBairro>CERQUEIRA CESAR</xBairro>
##     <cMun>3550308</cMun>
##     <xMun>SAO PAULO</xMun>
##     <UF>SP</UF>
##   </retirada>
##   <entrega>
##     <CNPJ>99299299000194</CNPJ>
##     <xLgr>AV FARIA LIMA</xLgr>
##     <nro>1500</nro>
##     <xCpl>15 ANDAR</xCpl>
##     <xBairro>PINHEIROS</xBairro>
##     <cMun>3550308</cMun>
##     <xMun>SAO PAULO</xMun>
##     <UF>SP</UF>
##   </entrega>
##   <det nItem="1">
##     <prod>
##       <cProd>00001</cProd>
##       <cEAN/>
##       <xProd>Agua Mineral</xProd>
##       <CFOP>5101</CFOP>
##       <uCom>dz</uCom>
##       <qCom>1000000.0000</qCom>
##       <vUnCom>1</vUnCom>
##       <vProd>10000000.00</vProd>
##       <cEANTrib/>
##       <uTrib>und</uTrib>
##       <qTrib>12000000.0000</qTrib>
##       <vUnTrib>1</vUnTrib>
##     </prod>
##     <imposto>
##       <ICMS>
##         <ICMS00>
##           <orig>0</orig>
##           <CST>00</CST>
##           <modBC>0</modBC>
##           <vBC>10000000.00</vBC>
##           <pICMS>18.00</pICMS>
##           <vICMS>1800000.00</vICMS>
##         </ICMS00>
##       </ICMS>
##       <PIS>
##         <PISAliq>
##           <CST>01</CST>
##           <vBC>10000000.00</vBC>
##           <pPIS>0.65</pPIS>
##           <vPIS>65000</vPIS>
##         </PISAliq>
##       </PIS>
##       <COFINS>
##         <COFINSAliq>
##           <CST>01</CST>
##           <vBC>10000000.00</vBC>
##           <pCOFINS>2.00</pCOFINS>
##           <vCOFINS>200000.00</vCOFINS>
##         </COFINSAliq>
##       </COFINS>
##     </imposto>
##   </det>
##   <det nItem="2">
##     <prod>
##       <cProd>00002</cProd>
##       <cEAN/>
##       <xProd>Agua Mineral</xProd>
##       <CFOP>5101</CFOP>
##       <uCom>pack</uCom>
##       <qCom>5000000.0000</qCom>
##       <vUnCom>2</vUnCom>
##       <vProd>10000000.00</vProd>
##       <cEANTrib/>
##       <uTrib>und</uTrib>
##       <qTrib>3000000.0000</qTrib>
##       <vUnTrib>0.3333</vUnTrib>
##     </prod>
##     <imposto>
##       <ICMS>
##         <ICMS00>
##           <orig>0</orig>
##           <CST>00</CST>
##           <modBC>0</modBC>
##           <vBC>10000000.00</vBC>
##           <pICMS>18.00</pICMS>
##           <vICMS>1800000.00</vICMS>
##         </ICMS00>
##       </ICMS>
##       <PIS>
##         <PISAliq>
##           <CST>01</CST>
##           <vBC>10000000.00</vBC>
##           <pPIS>0.65</pPIS>
##           <vPIS>65000</vPIS>
##         </PISAliq>
##       </PIS>
##       <COFINS>
##         <COFINSAliq>
##             <CST>01</CST>
##             <vBC>10000000.00</vBC>
##             <pCOFINS>2.00</pCOFINS>
##             <vCOFINS>200000.00</vCOFINS>
##           </COFINSAliq>
##         </COFINS>
##       </imposto>
##     </det>
##     <total>
##       <ICMSTot>
##         <vBC>20000000.00</vBC>
##         <vICMS>18.00</vICMS>
##         <vBCST>0</vBCST>
##         <vST>0</vST>
##         <vProd>20000000.00</vProd>
##         <vFrete>0</vFrete>
##         <vSeg>0</vSeg>
##         <vDesc>0</vDesc>
##         <vII>0</vII>
##         <vIPI>0</vIPI>
##         <vPIS>130000.00</vPIS>
##         <vCOFINS>400000.00</vCOFINS>
##         <vOutro>0</vOutro>
##         <vNF>20000000.00</vNF>
##       </ICMSTot>
##     </total>
##     <transp>
##       <modFrete>0</modFrete>
##       <transporta>
##         <CNPJ>99171171000191</CNPJ>
##         <xNome>Distribuidora de Bebidas Fazenda de SP Ltda.</xNome>
##         <IE>171999999119</IE>
##         <xEnder>Rua Central 100 - Fundos - Distrito Industrial</xEnder>
##         <xMun>SAO PAULO</xMun>
##         <UF>SP</UF>
##       </transporta>
##       <veicTransp>
##         <placa>BXI1717</placa>
##         <UF>SP</UF>
##         <RNTC>123456789</RNTC>
##       </veicTransp>
##       <reboque>
##         <placa>BXI1818</placa>
##         <UF>SP</UF>
##         <RNTC>123456789</RNTC>
##       </reboque>
##       <vol>
##         <qVol>10000</qVol>
##         <esp>CAIXA</esp>
##         <marca>LINDOYA</marca>
##         <nVol>500</nVol>
##         <pesoL>1000000000.000</pesoL>
##         <pesoB>1200000000.000</pesoB>
##         <lacres>
##         <nLacre>XYZ10231486</nLacre>
##       </lacres>
##     </vol>
##   </transp>
##   <infAdic>
##     <infAdFisco>Nota Fiscal de exemplo NF-eletronica.com</infAdFisco>
##   </infAdic>
## </infNFe>
```

```
rootNode[[1]][[1]]
```

```
## <ide>
##   <cUF>35</cUF>
##   <cNF>518005127</cNF>
##   <natOp>Venda a vista</natOp>
##   <indPag>0</indPag>
##   <mod>55</mod>
##   <serie>1</serie>
##   <nNF>1</nNF>
##   <dEmi>2008-05-06</dEmi>
##   <dSaiEnt>2008-05-06</dSaiEnt>
##   <tpNF>0</tpNF>
##   <cMunFG>3550308</cMunFG>
##   <tpImp>1</tpImp>
##   <tpEmis>1</tpEmis>
##   <cDV>3</cDV>
##   <tpAmb>2</tpAmb>
##   <finNFe>1</finNFe>
##   <procEmi>0</procEmi>
##   <verProc>NF-eletronica.com</verProc>
## </ide>
```

Programaticamente extrair partes do arquivo

```
xmlSApply(rootNode,xmlValue)
```

```
##
##
##
## "xhTSDMH61e9uqe04lnoHT4ZzLSY=Iz5Z3PLQbzZt9jnBtr6xsmHZMOu/3plXG9xxfFjRCQYGnD1rjlhzBGrqt026Ca2VHHM/bHN
```

```
xpathSApply(rootNode, "//*",xmlValue)
```

```
##   [1] "35518005127Venda a vista055112008-05-062008-05-0603550308113210NF-eletronica.com99999090910270
##   [2] "35518005127Venda a vista055112008-05-062008-05-0603550308113210NF-eletronica.com99999090910270
##   [3] "35518005127Venda a vista055112008-05-062008-05-0603550308113210NF-eletronica.com"
##   [4] "35"
##   [5] "518005127"
##   [6] "Venda a vista"
##   [7] "0"
##   [8] "55"
##   [9] "1"
##  [10] "1"
##  [11] "2008-05-06"
##  [12] "2008-05-06"
##  [13] "0"
##  [14] "3550308"
##  [15] "1"
##  [16] "1"
##  [17] "3"
##  [18] "2"
##  [19] "1"
##  [20] "0"
##  [21] "NF-eletronica.com"
##  [22] "99999090910270NF-e Associacao NF-eNF-eRua Central100FundosDistrito Industrial3502200AngatubaSP
##  [23] "99999090910270"
##  [24] "NF-e Associacao NF-e"
##  [25] "NF-e"
##  [26] "Rua Central100FundosDistrito Industrial3502200AngatubaSP171001711058Brasil1733021717"
##  [27] "Rua Central"
##  [28] "100"
##  [29] "Fundos"
##  [30] "Distrito Industrial"
##  [31] "3502200"
##  [32] "Angatuba"
##  [33] "SP"
##  [34] "17100171"
##  [35] "1058"
##  [36] "Brasil"
##  [37] "1733021717"
##  [38] "123456789012"
##  [39] "00000000000191DISTRIBUIDORA DE AGUAS MINERAISAV DAS FONTES177710 ANDARPARQUE FONTES5030801Sao
##  [40] "00000000000191"
##  [41] "DISTRIBUIDORA DE AGUAS MINERAIS"
##  [42] "AV DAS FONTES177710 ANDARPARQUE FONTES5030801Sao PauloSP139500001058BRASIL1932011234"
##  [43] "AV DAS FONTES"
##  [44] "1777"
##  [45] "10 ANDAR"
##  [46] "PARQUE FONTES"
##  [47] "5030801"
##  [48] "Sao Paulo"
##  [49] "SP"
##  [50] "13950000"
##  [51] "1058"
##  [52] "BRASIL"
##  [53] "1932011234"
##  [54] " "
##  [55] "99171171000194AV PAULISTA12345TERREOCERQUEIRA CESAR3550308SAO PAULOSP"
##  [56] "99171171000194"
##  [57] "AV PAULISTA"
##  [58] "12345"
##  [59] "TERREO"
##  [60] "CERQUEIRA CESAR"
##  [61] "3550308"
##  [62] "SAO PAULO"
##  [63] "SP"
##  [64] "99299299000194AV FARIA LIMA150015 ANDARPINHEIROS3550308SAO PAULOSP"
##  [65] "99299299000194"
##  [66] "AV FARIA LIMA"
##  [67] "1500"
##  [68] "15 ANDAR"
##  [69] "PINHEIROS"
##  [70] "3550308"
##  [71] "SAO PAULO"
##  [72] "SP"
##  [73] "00001Agua Mineral5101dz1000000.0000110000000.00und12000000.00001000010000000.0018.001800000.00
##  [74] "00001Agua Mineral5101dz1000000.0000110000000.00und12000000.00001"
##  [75] "00001"
##  [76] ""
##  [77] "Agua Mineral"
##  [78] "5101"
##  [79] "dz"
##  [80] "1000000.0000"
##  [81] "1"
##  [82] "10000000.00"
##  [83] ""
##  [84] "und"
##  [85] "12000000.0000"
##  [86] "1"
##  [87] "000010000000.0018.001800000.000110000000.000.65650000110000000.002.00200000.00"
##  [88] "000010000000.0018.001800000.00"
##  [89] "000010000000.0018.001800000.00"
##  [90] "0"
##  [91] "00"
##  [92] "0"
##  [93] "10000000.00"
##  [94] "18.00"
##  [95] "1800000.00"
##  [96] "0110000000.000.6565000"
##  [97] "0110000000.000.6565000"
##  [98] "01"
##  [99] "10000000.00"
## [100] "0.65"
## [101] "65000"
## [102] "0110000000.002.00200000.00"
## [103] "0110000000.002.00200000.00"
## [104] "01"
## [105] "10000000.00"
## [106] "2.00"
## [107] "200000.00"
## [108] "00002Agua Mineral5101pack5000000.0000210000000.00und3000000.00000.3333000010000000.0018.001800
## [109] "00002Agua Mineral5101pack5000000.0000210000000.00und3000000.00000.3333"
## [110] "00002"
## [111] ""
## [112] "Agua Mineral"
## [113] "5101"
## [114] "pack"
## [115] "5000000.0000"
## [116] "2"
## [117] "10000000.00"
## [118] ""
## [119] "und"
## [120] "3000000.0000"
## [121] "0.3333"
## [122] "000010000000.0018.001800000.000110000000.000.65650000110000000.002.00200000.00"
## [123] "000010000000.0018.001800000.00"
## [124] "000010000000.0018.001800000.00"
## [125] "0"
## [126] "00"
## [127] "0"
## [128] "10000000.00"
## [129] "18.00"
## [130] "1800000.00"
## [131] "0110000000.000.6565000"
## [132] "0110000000.000.6565000"
## [133] "01"
## [134] "10000000.00"
## [135] "0.65"
## [136] "65000"
## [137] "0110000000.002.00200000.00"
## [138] "0110000000.002.00200000.00"
## [139] "01"
## [140] "10000000.00"
## [141] "2.00"
## [142] "200000.00"
## [143] "20000000.0018.000020000000.0000000130000.00400000.00020000000.00"
## [144] "20000000.0018.000020000000.0000000130000.00400000.00020000000.00"
## [145] "20000000.00"
## [146] "18.00"
## [147] "0"
## [148] "0"
## [149] "20000000.00"
## [150] "0"
## [151] "0"
## [152] "0"
## [153] "0"
## [154] "0"
## [155] "130000.00"
## [156] "400000.00"
## [157] "0"
## [158] "20000000.00"
## [159] "099171171000191Distribuidora de Bebidas Fazenda de SP Ltda.171999999119Rua Central 100 - Fundo
## [160] "0"
## [161] "99171171000191Distribuidora de Bebidas Fazenda de SP Ltda.171999999119Rua Central 100 - Fundos
## [162] "99171171000191"
## [163] "Distribuidora de Bebidas Fazenda de SP Ltda."
## [164] "171999999119"
## [165] "Rua Central 100 - Fundos - Distrito Industrial"
## [166] "SAO PAULO"
## [167] "SP"
## [168] "BXI1717SP123456789"
## [169] "BXI1717"
## [170] "SP"
## [171] "123456789"
## [172] "BXI1818SP123456789"
## [173] "BXI1818"
## [174] "SP"
## [175] "123456789"
## [176] "10000CAIXALINDOYA5001000000000.0001200000000.000XYZ10231486"
## [177] "10000"
## [178] "CAIXA"
## [179] "LINDOYA"
## [180] "500"
## [181] "1000000000.000"
## [182] "1200000000.000"
## [183] "XYZ10231486"
## [184] "XYZ10231486"
## [185] "Nota Fiscal de exemplo NF-eletronica.com"
## [186] "Nota Fiscal de exemplo NF-eletronica.com"
## [187] "xhTSDMH61e9uqe04lnoHT4ZzLSY=Iz5Z3PLQbzZt9jnBtr6xsmHZMOu/3plXG9xxfFjRCQYGnD1rjlhzBGrqt026Ca2VH
## [188] "xhTSDMH61e9uqe04lnoHT4ZzLSY="
## [189] ""
## [190] ""
## [191] "xhTSDMH61e9uqe04lnoHT4ZzLSY="
## [192] ""
## [193] ""
## [194] ""
## [195] ""
## [196] "xhTSDMH61e9uqe04lnoHT4ZzLSY="
## [197] "Iz5Z3PLQbzZt9jnBtr6xsmHZMOu/3plXG9xxfFjRCQYGnD1rjlhzBGrqt026Ca2VHHM/bHNepi6FuFkAi595GScKVuHRE
## [198] "MIIEuzCCA6OgAwIBAgIDMTMxMA0GCSqGSIb3DQEBBQUAMIGSMQswCQYDVQQGEwJCUjELMAkGA1UECBMCUlMxFTATBgNVB
## [199] "MIIEuzCCA6OgAwIBAgIDMTMxMA0GCSqGSIb3DQEBBQUAMIGSMQswCQYDVQQGEwJCUjELMAkGA1UECBMCUlMxFTATBgNVB
## [200] "MIIEuzCCA6OgAwIBAgIDMTMxMA0GCSqGSIb3DQEBBQUAMIGSMQswCQYDVQQGEwJCUjELMAkGA1UECBMCUlMxFTATBgNVB
```