Reading HTML from a URL

Arin Parsa

7/13/2020

Example for reading HTML from a URL

```
connection <- url("http://biostat.jhsph.edu/~jleek/contact.html")

## readLines gives a character array back for every line from the HTML file
txt <- readLines(connection)
close(connection)
class(txt)
```

```
## [1] "character"
```

```
txt
```

```
##   [1] "<!DOCTYPE html PUBLIC \"-//W3C//DTD XHTML 1.0 Strict//EN\" \"http://www.w3.org/TR/xhtml1/DTD/
##   [2] ""
##   [3] "<html xmlns=\"http://www.w3.org/1999/xhtml\" xml:lang=\"en\" lang=\"en\">"
##   [4] ""
##   [5] "<head>"
##   [6] ""
##   [7] "<meta name=\"Description\" content=\"Welcome to Jeff Leek's Research Group\" />"
##   [8] "<meta name=\"Keywords\" content=\"Johns Hopkins University, Bloomberg School of Public Health
##   [9] "<meta http-equiv=\"Content-Type\" content=\"text/html; charset=iso-8859-1\" />"
##  [10] "<meta name=\"Distribution\" content=\"Global\" />"
##  [11] "<meta name=\"Robots\" content=\"index,follow\" />"
##  [12] ""
##  [13] "<link rel=\"icon\" type=\"image/vnd.microsoft.icon\" href=\"images/favicon.ico\" />"
##  [14] "<link rel=\"stylesheet\" href=\"images/PixelGreen.css\" type=\"text/css\" />"
##  [15] ""
##  [16] ""
##  [17] ""
##  [18] "<title>jeffrey leek contact</title>"
##  [19] ""
##  [20] "<script type=\"text/javascript\">"
##  [21] ""
##  [22] "  var _gaq = _gaq || [];"
##  [23] "  _gaq.push(['_setAccount', 'UA-20898652-1']);"
##  [24] "  _gaq.push(['_trackPageview']);"
##  [25] ""
##  [26] "  (function() {"
##  [27] "    var ga = document.createElement('script'); ga.type = 'text/javascript'; ga.async = true;"
##  [28] "    ga.src = ('https:' == document.location.protocol ? 'https://ssl' : 'http://www') + '.goog
##  [29] "    var s = document.getElementsByTagName('script')[0]; s.parentNode.insertBefore(ga, s);"
##  [30] "  })();"
```

```
##  [31] ""
##  [32] "</script>"
##  [33] "</head>"
##  [34] ""
##  [35] "<body>"
##  [36] "<!-- wrap starts here -->"
##  [37] "<div id=\"wrap\">"
##  [38] ""
##  [39] "\t<div id=\"header\"><div id=\"header-content\">\t"
##  [40] "\t\t"
##  [41] "\t\t<h1 id=\"logo\"><a href=\"research.html\" title=\"\">jeffrey<span class=\"gray\">leek</spa
##  [42] "\t\t<h2 id=\"slogan\">High Dimensional Data and Genomics</h2>\t\t"
##  [43] "\t\t"
##  [44] "\t\t<!-- Menu Tabs -->"
##  [45] "\t\t<ul>"
##  [46] "\t\t\t<li><a href=\"research.html\">Research</a></li>"
##  [47] "\t\t\t<li><a href=\"publications.html\">Publications</a></li>"
##  [48] "\t\t\t<li><a href=\"software.html\">Software and Data</a></li>"
##  [49] "\t\t\t<li><a href=\"about.html\">About</a></li>"
##  [50] "\t\t\t<li><a href=\"contact.html\" id=\"current\">Contact</a></li>\t\t\t"
##  [51] "\t\t</ul>\t\t"
##  [52] "\t\t\t"
##  [53] "\t"
##  [54] "\t</div></div>"
##  [55] "\t"
##  [56] "\t<div class=\"headerphoto-cont\"></div>"
##  [57] "\t\t\t\t"
##  [58] "\t<!-- content-wrap starts here -->"
##  [59] "\t<div id=\"content-wrap\"><div id=\"content\">\t\t"
##  [60] "\t\t"
##  [61] "\t\t<div id=\"sidebar\" >"
##  [62] "\t\t"
##  [63] "\t\t\t<div class=\"sidebox\">"
##  [64] "\t\t\t"
##  [65] "\t\t\t\t<h1>About</h1>"
##  [66] "\t\t\t\t"
##  [67] "\t\t\t\t<p>Jeff is an assistant professor in the Biostatistics Department of the Johns Hopkins
##  [68] "\t\t\t  methods for high-dimensional data and genomics. For more info check out his <a href=\
##  [69] "\t\t\t\t\t\t"
##  [70] "\t\t\t</div>\t\t\t"
##  [71] ""
##  [72] "\t\t\t<!--"
##  [73] "\t\t\t<div class=\"sidebox\">\t"
##  [74] "\t\t\t"
##  [75] "\t\t\t\t<h1 class=\"clear\">Recent Projects</h1>"
##  [76] "\t\t\t\t<ul class=\"sidemenu\">"
##  [77] "\t\t\t           <li><a href=\"./papers/sets.pdf\" class=\"top\"> Interpretable Gene Sets </a
##  [78] "\t\t\t\t   <li><a href=\"./papers/acsvd.pdf\" class=\"top\"> Asymptotic SVD </a></li>"
##  [79] "\t\t\t\t   <li><a href=\"./papers/tspair.pdf\" class=\"top\"> Top-Scoring Pairs </a></li>"
##  [80] "\t\t\t\t   <li><a href=\"./papers/sva.pdf\" class=\"top\" > Expression heterogeneity </a></li
##  [81] "\t\t\t\t</ul>\t"
##  [82] "\t\t\t\t"
##  [83] "\t\t\t</div>\t"
##  [84] "\t\t\t-->"
```

```
##  [85] "\t\t\t"
##  [86] "\t\t\t\t<div class=\"sidebox\">"
##  [87] "\t\t\t\t<h1>Media</h1>"
##  [88] "\t\t\t\t<ul class=\"sidemenu\">"
##  [89] "\t\t\t\t\t<li><a href=\"http://simplystatistics.tumblr.com/\">Simply Statistics (Blog)</a></li
##  [90] "\t\t\t\t\t<li><a href=\"http://www.twitter.com/leekgroup\">@leekgroup</a></li>\t\t\t\t\t\t"
##  [91] "\t\t\t\t</ul>"
##  [92] "\t\t\t\t"
##  [93] "\t\t\t      </div>"
##  [94] "\t\t\t"
##  [95] "\t\t\t  "
##  [96] "\t\t\t<div class=\"sidebox\">"
##  [97] "\t\t"
##  [98] "\t\t\t"
##  [99] "\t\t\t\t<h1>Affiliations</h1>"
## [100] "\t\t\t\t<ul class=\"sidemenu\">"
## [101] "\t\t\t\t\t<li><a href=\"http://www.jhsph.edu/\" class=\"top\">JHSPH</a></li>"
## [102] "\t\t\t\t\t<li><a href=\"http://www.biostat.jhsph.edu/\">Biostatistics Department</a></li>"
## [103] "\t\t\t\t\t<li><a href=\"http://www.biostat.jhsph.edu/genomics/\">Genomics @ Biostat</a></li>"
## [104] "\t\t\t\t\t<li><a href=\"http://genomics.jhu.edu/\">Computational Genomics</a></li>\t\t\t\t\t\t
## [105] "\t\t\t\t</ul>"
## [106] "\t\t\t\t"
## [107] "\t\t\t</div>"
## [108] "\t\t\t"
## [109] "\t\t\t\t\t"
## [110] "\t\t</div>\t"
## [111] "\t"
## [112] "\t\t<div id=\"main\">\t\t"
## [113] "\t\t"
## [114] "\t\t\t<div class=\"post\">"
## [115] "\t\t\t"
## [116] "\t\t\t\t<a name=\"TemplateInfo\"></a>\t"
## [117] "\t\t\t\t<h1>Contact Information</h1>"
## [118] ""
## [119] "\t\t\t  <h3>Address </h3>"
## [120] "\t\t\t  <p>"
## [121] "\t\t\t  Johns Hopkins University <br/>"
## [122] "\t\t\t  Bloomberg School of Public Health <br/>"
## [123] "\t\t\t  615 North Wolfe Street <br/>"
## [124] "\t\t\t  Baltimore, MD 21205-2179 <br/>"
## [125] "\t\t\t  </p>"
## [126] "\t\t\t  <h3>Phone</h3>"
## [127] "\t\t\t  <p>410-955-1166 (I am <b>much</b> easier to reach by email)</p>"
## [128] "\t\t\t  <h3>Fax</h3>"
## [129] "\t\t\t  <p>410-955-0958</p>"
## [130] "\t\t\t  <h3>Email</h3>"
## [131] "\t\t\t  <p> jleek || jhsph dot edu </p>"
## [132] "\t\t\t  <h3>Twitter</h3>"
## [133] "\t\t\t  <p> <a href=\"http://www.twitter.com/leekgroup\">@leekgroup</a></p>"
## [134] "\t\t\t  <h3>Blog</h3>"
## [135] "\t\t\t  <p> <a href=\"http://simplystatistics.tumblr.com/\">Simply Statistics</a></p>"
## [136] "\t\t\t  "
## [137] "\t\t\t\t"
## [138] "\t\t\t</div>"
```

```
## [139] "\t\t\t"
## [140] "\t\t\t\t\t"
## [141] "\t\t\t\t\t\t\t\t\t\t"
## [142] "\t\t</div>\t\t\t\t\t"
## [143] "\t\t"
## [144] "\t<!-- content-wrap ends here -->\t\t"
## [145] "\t</div></div>"
## [146] ""
## [147] "<!-- footer starts here -->\t"
## [148] "<div id=\"footer\"><div id=\"footer-content\">"
## [149] "\t"
## [150] "\t\t"
## [151] "\t\t<div class=\"col float-left\">"
## [152] "\t\t\t<h1>Links</h1>"
## [153] "\t\t\t<ul>\t\t\t\t"
## [154] "\t\t\t\t<li><a href=\"http://groups.google.com/group/edge-software?pli=1\">EDGE Discussion Gr
## [155] "\t\t\t\t<li><a href=\"http://www.biostat.jhsph.edu/newsEvent/event/seminar/seminars.shtml\">B
## [156] "\t\t\t</ul>\t\t\t"
## [157] "\t\t</div>"
## [158] ""
## [159] "\t"
## [160] "\t\t<div class=\"col2 float-right\">"
## [161] "\t\t<p>"
## [162] "\t\t&copy; copyright 2009 <strong>Jeffrey Leek</strong><br />"
## [163] "\t\tArtwork &copy copyright 2009 <a href=\"mailto:codyandanna@gmail.com\">Cody Lail</a> <br/>
## [164] "\t\tPhoto <a href = \"attributions.html\">Attributions</a><br/>"
## [165] "\t\tDesign by: <a href=\"http://www.styleshout.com/\"><strong>styleshout</strong></a>   &
## [166] "\t\tValid <a href=\"http://jigsaw.w3.org/css-validator/check/referer\"><strong>CSS</strong></
## [167] "\t\t      <a href=\"http://validator.w3.org/check/referer\"><strong>XHTML</strong></a>"
## [168] "\t\t</p>"
## [169] "\t\t"
## [170] "\t"
## [171] "\t\t</div>"
## [172] "\t"
## [173] "</div></div>"
## [174] "<!-- footer ends here -->"
## [175] "\t"
## [176] "<!-- wrap ends here -->"
## [177] "</div>"
## [178] ""
## [179] "</body>"
## [180] "</html>"
```

```
## Finding number of characters on a line
nchar(txt[10])
```

```
## [1] 45
```

```
nchar(txt[20])
```

```
## [1] 31
```

```
nchar(txt[30])
```

```
## [1] 7
```

```r
nchar(txt[100])
```

```
## [1] 25
```