Web Scraping and Analysis – Web Scraping

Yu-Chen (Amber) Lu

Date: 2018-12-22

Web Scraping

Use the Race Results 1999-2012 link (http://cherryblossom.org/aboutus/results_list.php) to extract the Race Results for Order of Finish – Women for a couple of years.

```
library(XML)

# Given the years I wanted to extract
years = 2001:2012
womenURLs =
  c("results/2001/oof_f.html", "results/2002/ooff.htm",
    "results/2003/CB03-M.HTM","results/2004/women.htm",
    "results/2005/CB05-F.htm","results/2006/women.htm",
    "results/2007/women.htm",
    "results/2008/women.htm", "results/2009/09cucb-F.htm",
    "results/2010/2010cucb10m-f.htm",
    "results/2011/2011cucb10m-f.htm",
    "results/2012/2012cucb10m-f.htm")

extractTable = function(url, year = 1999, file = NULL){

  ## Retrieve data from web site, find preformatted text,
  ## return as a character vector.
  ## From running results for year 2001 to 2012
  ## On http://www.cherryblossom.org/

  if (year == 2001){

    ## Year 2001 is missing header
    ## Add the header and space rows back to the table

    ubase = "http://www.cherryblossom.org/"
    url = paste(ubase, url, sep = "")
    doc = htmlParse(url)
    preNode = getNodeSet(doc, "//pre")

    txt = xmlValue(preNode[[1]])
    els = strsplit(txt, "\\r\\n")[[1]]
    els[[2]] = "PLACE DIV /  NAME                 AG HOMETOWN           TIME    NET "
    els[[3]] = "===== ===== ==================== == ================== ======= ===== "

  }

  else {
    ubase = "http://www.cherryblossom.org/"
  url = paste(ubase, url, sep = "")
  doc = htmlParse(url)
```

```
  preNode = getNodeSet(doc, "//pre")

  txt = xmlValue(preNode[[1]])
  els = strsplit(txt, "\\r\\n")[[1]]
  }


  if (is.null(file)) return(els)

  elss = writeLines(els, con = file)
  return(elss)
}

womenTables = mapply(extractTable, url = womenURLs, year = years)
names(womenTables) = years
save(womenTables, file = "CBWomenTextTables.rda")
```