Problem Set 2

Yue Hu #3033030912

Sept/15/2017

Problem 1

1 a)

for chars <- sample(letters, 1e6, replace = TRUE) write.table(chars, file = ‘tmp1.csv’, row.names = FALSE, quote = FALSE, col.names = FALSE) each letter takes up 1 byte, and each line-break takes up 1 byte, so 1e6 items in an ASCII file takes 2e6 bytes.

for chars <- paste(chars, collapse = ‘’) write.table(chars, file = ’tmp2.csv’, row.names = FALSE, quote = FALSE, col.names = FALSE) each letter takes up 1 byte and there is no line-break in between. so in this ASCII file 1e6 letters and a line-break in the end takes (1e6+1) bytes.

for nums <- rnorm(1e6) save(nums, file = ‘tmp3.Rda’) each number is treated as double format, taking 8bytes, so 1e6 numbers take approx. 8e6 bytes.

for write.table(nums, file = ‘tmp4.csv’, row.names = FALSE, quote = FALSE, col.names = FALSE, sep = ‘,’)
each number is treated as a series of characters, each taking 1 byte, so the file size is obviously larger.

for write.table(round(nums, 2), file = ‘tmp5.csv’, row.names = FALSE, quote = FALSE, col.names = FALSE, sep = ‘,’) each number is 2 decimal, with 4 characters. so 1e6 numbers plus line-breaks will take approx. 5e6 bytes.

1 b)

for chars <- sample(letters, 1e6, replace = TRUE) chars <- paste(chars, collapse = ‘’) save(chars, file = ’tmp6.Rda’) The save function has deault seting of ascii = FALSE and compress= isTRUE, and will write a binary file applying gzip compresssion. That’s why the file format is much smaller.

For chars <- rep(‘a’, 1e6) chars <- paste(chars, collapse = ‘’) save(chars, file = ’tmp7.Rda’) The gzip is based on DEFLATE algorithm. If a duplicate series of bytes is spotted (a repeated string), then a back-reference is inserted, linking to the previous location of that identical string instead. So 1e6 identical character ‘a’ is even more compressed.

Problem2

2 (a)

Create a function whose input is the character string of the name of the researcher and whose output is the html text corresponding to the researcher citation page as well as the researche’s Google Scholar ID

```
scholar <- function(name){
  library(XML)
  name2 <- gsub("\ ","+", name)
  #from the http request we can see that each blank space (including in the
beginning and multiple space in the space) is replaced by a +.
  baseURL <- "http://scholar.google.com"
  filter1 <- paste0("/citations?view_op=search_authors&mauthors=", name2, "&
hl=en&oi=ao")
  url1 <- paste0(baseURL, filter1)
  #construct url
  html1 <- htmlParse(url1)
  #download html. The object returned by htmlParse() produces nicely formatt
ed text
  nodeh3set <- getNodeSet(html1, "//h3[@class = 'gsc_1usr_name']")
  #find nodeset named h3 and has attribute as gsc_1usr_name, where the ID li
es.
  a <- sapply(nodeh3set, xmlChildren)
  href <- sapply(a,xmlGetAttr, "href")
  #href is the attribute of its child node a
  url2 <- paste0(baseURL,href)
  html2 <- htmlParse(url2)
  #download the new html
  id <- gsub('.*user=(.*?)&.*', '\\1', href)
  #the href, for example "/citations?user=CXJuZ5YAAAAJ&hl=en&oe=ASCII" ,incl
udes the user id.use regex to extract the part after "user="and before "&",
non-greedy match.
  #return a list of id and html
  slist <- list("id"=id, "html"=html2)
  return(slist)
}
d <- scholar(" Geoffrey Hinton")
```

2 (b)

Create a second function to process the resulting HTML to create an R data frame that contains the article title, authors, journal information, year of publication, and number of citations as five columns of information.

```
df <- function(name){
  html2 <- scholar(name)[["html"]]
  #use function in 2a to get the html
  #All information lies in one table . Since article name, author and journa
l is in the same table cell, separated by <div>, directly using readHTMLTabl
e will paste them together. So I use XPath to substrct the table, read each
part as a vector, and merge them to a dataframe.
  #All information lies in one table element whose class is 'gsc_a_t',use //
td[@class = 'gsc_a_t'] to locate them
  narticle <- xpathSApply(html2, "//td[@class = 'gsc_a_t']/a")
  article <- sapply(narticle, xmlValue)
    #get the child element, whose value is the article name
  nauthor <- xpathSApply(html2, "//td[@class = 'gsc_a_t']/div[1]")
  author <- sapply(nauthor, xmlValue)
  #get the first "div" child element, whose value is author
  njn <- xpathSApply(html2, "//td[@class = 'gsc_a_t']/div[2]")
  journal <- sapply(njn, xmlValue)
  #get the  second "div" child element, whose value is journal
  ncited <- xpathSApply(html2, "//td[@class = 'gsc_a_c']")
  nyear <- xpathSApply(html2, "//td[@class = 'gsc_a_y']")
  cited <- sapply(ncited, xmlValue)
  year <- sapply(nyear, xmlValue)
  year <- sapply(year, as.numeric)
  cited <- sapply(cited, as.numeric)
  #get table cell elements whose classes are 'gsc_a_c' and 'gsc_a_y', whose
value is cited number and year. change them to numeric
  df <- cbind.data.frame(article, author, journal, cited, year, stringsAsFac
tors=FALSE)
  #merge them to form a dataframe
  return(df)
}

#Try the function on a second researcher to provide more confidence that th
e function is working properly.

df("Steven Glaser")
```

```
#
#
                                                                    article
## 1                                                                    H
ealth monitoring of civil infrastructures using wireless sensor networks
##
2
               Some real-world applications of wireless sensor nodes
## 3                                                          OpenWSN: a st
andardsa<80><90>based lowa<80><90>power wireless development environment
## 4  Sensor technology innovation for the advancement of structural health
monitoring: a strategic program of US-China research for the next decade
##
5
Acoustic emission sensor calibration for absolute source measurements
## 6                                                                 Wavele
t denoising techniques with applications to experimental geophysical data
## 7                                                    Beamforming array
techniques for acoustic emission monitoring of large concrete structures
##
8
             Physical activity monitoring for assisted living at home
##
9
 Influence of rock mass strength on the erosion rate of alpine cliffs
## 10                                                                    M
ulti-purpose wireless accelerometers for civil infrastructure monitoring
## 11                                                              Hertzian
impact: Experimental study of the force pulse and resulting stress waves
## 1
2
                    Mobile transit trip planning with real-time data
## 13                                                     Mobile phones
as seismologic sensors: Automating data extraction for the iShake system
## 1
4                                                                    Sy
stem identification estimation of soil properties at the Lotung site
## 1
5
                           Frontiers in sensors and sensing systems
## 16                                                                 Sen
se of sensing: from data to informed decisions for the built environment
## 17                            Design and performance of a wireless senso
r network for catchmenta<80><90>scale snow and soil moisture measurements
## 18                                          Fault healing promotes hi
gh-frequency earthquakes in laboratory experiments and on natural faults
## 19                                                                 Mic
romechanics of asperity rupture during laboratory stick slip experiments
## 20                                                  Body waves record
ed inside an elastic half-space by an embedded, wideband velocity sensor
##                                                                  author
## 1     S Kim, S Pakzad, D Culler, J Demmel, G Fenves, S Glaser, M Turon
```

```
## 2                                                                SD Glaser
## 3  T Watteyne, X Vilajosana, B Kerkez, F Chraim, K Weekly, Q Wang, ...
## 4                                SD Glaser, H Li, ML Wang, J Ou, J Lynch
## 5                                                    GC McLaskey, SD Glaser
## 6                                                AC To, JR Moore, SD Glaser
## 7                                    GC McLaskey, SD Glaser, CU Grosse
## 8                           R Jafari, W Li, R Bajcsy, S Glaser, S Sastry
## 9                           JR Moore, JW Sanders, WE Dietrich, SD Glaser
## 10        SN Pakzad, S Kim, GL Fenves, SD Glaser, DE Culler, JW Demmel
## 11                                                   GC McLaskey, SD Glaser
## 12    J Jariyasunant, DB Work, B Kerkez, R Sengupta, S Glaser, A Bayen
## 13         J Reilly, S Dashti, M Ervasti, JD Bray, SD Glaser, AM Bayen
## 14                                                     SD Glaser, LG Baise
## 15                                  SD Glaser, RA Shoureshi, D Pescovitz
## 16                                                   SD Glaser, A Tolman
## 17                              B Kerkez, SD Glaser, RC Bales, MW Meadows
## 18                           GC McLaskey, AM Thomas, SD Glaser, RM Nadeau
## 19                                                   GC McLaskey, SD Glaser
## 20                                      SD Glaser, GG Weiss, LR Johnson
#
#
        journal
## 1  Proceedings of the 6th international conference on Information process
ing in ..., 2007
## 2                                                          Proc. of SPIE Vo
l 5391, 345, 2004
## 3          Transactions on Emerging Telecommunications Technologies 23
(5), 480-493, 2012
## 4                                        Smart Structures and Systems 3
(2), 221-244, 2007
## 5                                Journal of Nondestructive Evaluation 31
(2), 157-168, 2012
## 6                                                 Signal Processing 89
(2), 144-160, 2009
## 7                                  Journal of Sound and Vibration 329 (1
2), 2384-2394, 2010
## 8   4th International Workshop on Wearable and Implantable Body Sensor Ne
tworks ..., 2007
## 9                           Earth Surface Processes and Landforms 34 (1
0), 1339-1352, 2009
## 10           Proceedings of the 5th International Workshop on Structural
Health ..., 2005
## 11              The Journal of the Acoustical Society of America 128
(3), 1087-1096, 2010
## 1
2

## 13          IEEE Transactions on Automation Science and Engineering 10
(2), 242-251, 2013
## 14                          Soil Dynamics and Earthquake Engineering 19
(7), 521-531, 2000
## 15                                       Smart Structures and Systems 1
```

```
(1), 103-120, 2005
## 16                                    Journal of infrastructure systems 1
4 (1), 4-14, 2008
## 17                                                 Water Resources Resea
rch 48 (9), 2012
## 18                                                              Nature 491
(7422), 101, 2012
## 19                                              Geophysical Research Lette
rs 38 (12), 2011
## 20              The Journal of the Acoustical Society of America 104
(3), 1404-1412, 1998
##    cited year
## 1   1071 2007
## 2    177 2004
## 3    175 2012
## 4     99 2007
## 5     85 2012
## 6     82 2009
## 7     80 2010
## 8     76 2007
## 9     68 2009
## 10    64 2005
## 11    52 2010
## 12    51 2011
## 13    49 2013
## 14    49 2000
## 15    48 2005
## 16    46 2008
## 17    44 2012
## 18    42 2012
## 19    41 2011
## 20    38 1998
```

2(c)

include checks in your code so that it fails gracefully if the user provides invalid input or Google Scholar doesn't return a result. Also write some test code that uses the testthat package to carry out a small number of tests of your function.

```
scholar <- function(name=NULL){
  if (is.null(name))
        stop("Need to specify a name.")
  #check if a value is entered
  if (is.character(name)==FALSE) stop("'name' must be string")
  #check if it is a string
  name2 <- gsub("\ ","+", name)
  baseURL <- "http://scholar.google.com"
  filter1 <- paste0("/citations?view_op=search_authors&mauthors=", name2, "&
hl=en&oi=ao")
  url1 <- paste0(baseURL, filter1)
  html1 <- htmlParse(url1)
  nodeh3set <- getNodeSet(html1, "//h3[@class = 'gsc_1usr_name']")
  a <- sapply(nodeh3set, xmlChildren)
  href <- sapply(a,xmlGetAttr, "href")
  url2 <- paste0(baseURL,href)
  html2 <- htmlParse(url2)
  id <- gsub('.*user=(.*?)&.*', '\\1', href)
  #check if the scholar exists
  if (length(id)==0) stop("didn't match any user profiles")
  slist <- list("id"=id, "html"=html2)
  return(slist)
}

#use testthat for tests

library(testthat)
context("Test df")

test_that("returning dataframe has dimension of (20,5)", {
  expect_equal(dim(df("Steven Glaser")),c(20,5) )
  expect_equal(dim(df("Geoffrey Hinton")),c(20,5) )
 })

test_that("returning cited times is numeric", {
  expect_equal(class(df("Steven Glaser")[,"cited"]),"numeric")
  expect_equal(class(df("Geoffrey Hinton")[,"cited"]),"numeric")

})
```

2(d)

(Extra credit) Fix your function so that you get all of the results for a researcher and not just the irst 20.

```
scholar2 <- function(name){
  library(XML)
  name2 <- gsub("\ ","+", name)
  baseURL <- "http://scholar.google.com"
  filter1 <- paste0("/citations?view_op=search_authors&mauthors=", name2, "&
hl=en&oi=ao")
  url1 <- paste0(baseURL, filter1)
  html1 <- htmlParse(url1)
  nodeh3set <- getNodeSet(html1, "//h3[@class = 'gsc_1usr_name']")
  a <- sapply(nodeh3set, xmlChildren)
  href <- sapply(a,xmlGetAttr, "href")

  #Check the http request and found that another aurgument named pages is ap
plied. rewrite the url accordingly.
  url2 <- paste0(baseURL,href,"&pagesize=80")
  html2 <- htmlParse(url2)
  id <- gsub('.*user=(.*?)&.*', '\\1', href)
  slist <- list("id"=id, "html"=html2)
  return(slist)
}


df2 <- function(name){
  html2 <- scholar2(name)[["html"]]
  node <- getNodeSet(html2,"//td[@class = 'gsc_a_t']")
  narticle <- xpathSApply(html2, "//td[@class = 'gsc_a_t']/a")
  article <- sapply(narticle, xmlValue)
  nauthor <- xpathSApply(html2, "//td[@class = 'gsc_a_t']/div[1]")
  author <- sapply(nauthor, xmlValue)
  njn <- xpathSApply(html2, "//td[@class = 'gsc_a_t']/div[2]")
  journal <- sapply(njn, xmlValue)
  ncited <- xpathSApply(html2, "//td[@class = 'gsc_a_c']")
  nyear <- xpathSApply(html2, "//td[@class = 'gsc_a_y']")
  cited <- sapply(ncited, xmlValue)
  year <- sapply(nyear, xmlValue)
  year <- sapply(year, as.numeric)
  cited <- sapply(cited, as.numeric)
  df <- cbind.data.frame(article, author, journal, cited, year, stringsAsFac
tors=FALSE)


  return(df)
}

head(df2("steven glaser"),)
```

```
## Warning in lapply(X = X, FUN = FUN, ...): 强制改变过程中产生了NA
```

```
#
#
                                                                    article
## 1                                                                    He
alth monitoring of civil infrastructures using wireless sensor networks
##
2
               Some real-world applications of wireless sensor nodes
## 3                                                          OpenWSN: a sta
ndardsa<80><90>based lowa<80><90>power wireless development environment
## 4 Sensor technology innovation for the advancement of structural health m
onitoring: a strategic program of US-China research for the next decade
##
5                                                                     A
coustic emission sensor calibration for absolute source measurements
## 6                                                                  Wavelet
denoising techniques with applications to experimental geophysical data
##                                                                 author
## 1    S Kim, S Pakzad, D Culler, J Demmel, G Fenves, S Glaser, M Turon
## 2                                                             SD Glaser
## 3 T Watteyne, X Vilajosana, B Kerkez, F Chraim, K Weekly, Q Wang, ...
## 4                              SD Glaser, H Li, ML Wang, J Ou, J Lynch
## 5                                        GC McLaskey, SD Glaser
## 6                                    AC To, JR Moore, SD Glaser
#
#
       journal
## 1 Proceedings of the 6th international conference on Information processi
ng in ..., 2007
## 2                                                  Proc. of SPIE Vol
5391, 345, 2004
## 3         Transactions on Emerging Telecommunications Technologies 23
(5), 480-493, 2012
## 4                                  Smart Structures and Systems 3
(2), 221-244, 2007
## 5                              Journal of Nondestructive Evaluation 31
(2), 157-168, 2012
## 6                                          Signal Processing 89
(2), 144-160, 2009
##   cited year
## 1  1071 2007
## 2   177 2004
## 3   175 2012
## 4    99 2007
## 5    85 2012
## 6    82 2009
```

Notes

Person I worked with : Hangyu Huang