PS2

ShubeiWang

9/7/2018

1

In my code in question(3), I used header information to put metainfo and illustrate what the code does. I also used blank lines to seperate blocks of codes and comments. Moreover, I added some assertions and testing to check the code operates correctly.

2

(a)

The file sizes vary because the data is stored in CSV text file as ASCII text format while in Rda file as binary format. Also there are delimiters and other characters stored in CSV file. In binary format, each number is stored as 8 bytes while in ASCII plain text format each character as one byte. In the CSV text file, there are 133887710 characters since in an ASCII file each character takes up one byte of space.

(b)

Because in this process every comma is actully replaced by a newline character, both of which take up one byte. Thus the file size remains unchanged.

(c)

First: Because read.csv is designed to read data frames which may have columns of very different classes. It uses scan to read the file and then process the results of scan. Unless colClasses is specified, all columns are read as character columns and then converted using type.convert to logical, integer, numeric, complex or factor as appropriate. So it takes much more time for read.csv to read the data than scan.

Second: When colClasses is specified, it saves the time for read.csv to process the data hence the speed between these two situations are very close.

Third: When using scan, if number of items is not specified, the internal mechanism re-allocates memory in powers of two. So it's faster to use load to read binary connections.

(d)

Because save() automatically compress the file and since each element in b is identical, it saves more memory in the process of compressing.

(3)

(a)

```
## Programmatically return a list of the Google Scholar ID and citation page
## of the researcher of interest.
## usage: get_page(x), argument x is the character string of the name of the researcher.

library(xml2)
library(rvest)
library(magrittr)

get_page <- function(name){

# get the user ID
id <- read_html(paste0("https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=", name, "&oq=geof"))%>%
html_nodes(".gs_rt2") %>% sub(".*user=([A-Za-z-]+)&.*","\\1",.)

# get the citation page
page <- read_html(paste0("https://scholar.google.com/citations?user=",id, "&hl=en&oi=ao"))

list <- list(id, page)
names(list) <- c("id", "page")
return (list)
}

#test for Trevor Hastie
get_page("trevorhastie")$id
```

```
## [1] "tQVe-fAAAAAJ"
```

(b)

```
## Create a dataframe of article title, authors, journal information, year of publication,
## and number of citations for the reseacher of interest.
## usage: get_article(x), argument x is the character string of the name of the researcher.

get_article <- function(page){

title <- page %>% html_nodes( ".gsc_a_at") %>% html_text()

author <- page %>%  html_nodes(".gs_gray") %>%
  html_text() %>% as.data.frame(stringAsFactors=FALSE) %>%
  .[seq(1, 20, by = 2),]

journal <- page %>%  html_nodes(".gs_gray") %>%
  html_text() %>% as.data.frame(stringAsFactors=FALSE) %>%
  .[seq(0, 20, by = 2),]

year <- suppressWarnings(page %>% html_nodes(".gsc_a_y") %>%
html_text() %>% as.numeric() %>% na.omit())
```

```
num_citation <- page %>% html_nodes(".gsc_a_ac") %>% html_text()

data <- data.frame(
  title = title,
  author = author,
  journal = journal,
  year = year,
  num_citation = num_citation
)

return (data)
}

#test for Trevor Hastie
page1 <- get_page("trevorhastie")$page
get_article(page1)
```

```
##
## 1                                                                                Unsupervis
## 2                                                                              Generalized addi
## 3          Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical
## 4                                                  Regularization and variable selection via the
## 5                                                                                  Least angl
## 6  Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by
## 7                                    Regularization paths for generalized linear models via coordi
## 8                                                          An introduction to statisti
## 9                                      Estimating the number of clusters in a data set via the ga
## 10                                                           The elements of statisti
## 11                                   The Dantzig selector: Statistical estimation when p is much la
## 12                                         Sparse inverse covariance estimation with the grap
## 13                                                                        Statistical
## 14                                                  A statistical explanation of MaxEnt fo
## 15                                 Diagnosis of multiple cancer types by shrunken centroids of gene
## 16                                                 Missing value estimation methods for DNA
## 17                                                         A working guide to boosted regre
## 18                                                                 Sparse principal compone
## 19                                                                         Varying-coeffi
## 20                                                                 Classification by pairwi
##                                                 author
## 1                                T Hastie, R Tibshirani, J Friedman
## 2                                                TJ Hastie
## 3  T Sørlie, CM Perou, R Tibshirani, T Aas, S Geisler, H Johnsen, T Hastie, ...
## 4                                                H Zou, T Hastie
## 5                             B Efron, T Hastie, I Johnstone, R Tibshirani
## 6                                J Friedman, T Hastie, R Tibshirani
## 7                                J Friedman, T Hastie, R Tibshirani
## 8                             G James, D Witten, T Hastie, R Tibshirani
## 9                                 R Tibshirani, G Walther, T Hastie
## 10                               J Friedman, T Hastie, R Tibshirani
## 11                               T Hastie, R Tibshirani, J Friedman
## 12                                               TJ Hastie
## 13 T Sørlie, CM Perou, R Tibshirani, T Aas, S Geisler, H Johnsen, T Hastie, ...
## 14                                               H Zou, T Hastie
## 15                            B Efron, T Hastie, I Johnstone, R Tibshirani
```

```
## 16                                          J Friedman, T Hastie, R Tibshirani
## 17                                          J Friedman, T Hastie, R Tibshirani
## 18                                  G James, D Witten, T Hastie, R Tibshirani
## 19                                           R Tibshirani, G Walther, T Hastie
## 20                                          J Friedman, T Hastie, R Tibshirani
##                                                                          journal
## 1                                 The elements of statistical learning, 485-585, 2009
## 2                                                 Statistical models in S, 249-307, 2017
## 3           Proceedings of the National Academy of Sciences 98 (19), 10869-10874, 2001
## 4  Journal of the Royal Statistical Society: Series B (Statistical Methodology ..., 2005
## 5                                      The Annals of statistics 32 (2), 407-499, 2004
## 6                                      The annals of statistics 28 (2), 337-407, 2000
## 7                                    Journal of statistical software 33 (1), 1, 2010
## 8                                                                  springer, 2013
## 9  Journal of the Royal Statistical Society: Series B (Statistical Methodology ..., 2001
## 10                                          Springer series in statistics 1 (10), 2001
## 11                                The elements of statistical learning, 485-585, 2009
## 12                                                Statistical models in S, 249-307, 2017
## 13          Proceedings of the National Academy of Sciences 98 (19), 10869-10874, 2001
## 14 Journal of the Royal Statistical Society: Series B (Statistical Methodology ..., 2005
## 15                                     The Annals of statistics 32 (2), 407-499, 2004
## 16                                     The annals of statistics 28 (2), 337-407, 2000
## 17                                   Journal of statistical software 33 (1), 1, 2010
## 18                                                                 springer, 2013
## 19 Journal of the Royal Statistical Society: Series B (Statistical Methodology ..., 2001
## 20                                          Springer series in statistics 1 (10), 2001
##    year num_citation
## 1  2009        40041
## 2  2017        15769
## 3  2001        11890
## 4  2005         8007
## 5  2004         7843
## 6  2000         6260
## 7  2010         5405
## 8  2013         3286
## 9  2001         3252
## 10 2001         2895
## 11 2007         2889
## 12 2008         2867
## 13 1992         2822
## 14 2011         2798
## 15 2002         2709
## 16 2001         2703
## 17 2008         2264
## 18 2006         2036
## 19 1993         1850
## 20 1998         1652
```

```r
#test for Geoffrey Hinton
page2 <- get_page("geoffreyhinton")$page
get_article(page2)
```

```
##
## 1                                                 Learning internal representations by error-pr
## 2                                                    Learning representations by back-propagati
```

```
## 3                                                 Imagenet classification with deep convolutional neural
## 4                                                         Learning internal representations by error pr
## 5                                                    Learning representations by back-propagati
## 6                                                                                                 Deep
## 7                                                              A fast learning algorithm for deep be
## 8                                                Dropout: a simple way to prevent neural networks from ov
## 9                                                              The appeal of parallel distributed pr
## 10                                                Reducing the dimensionality of data with neural
## 11                                                                                Visualizing data us
## 12 Deep neural networks for acoustic modeling in speech recognition: The shared views of four researc
## 13                                                 Rectified linear units improve restricted boltzmann
## 14                                                                      Adaptive mixtures of local
## 15                                                          A learning algorithm for Boltzmann
## 16                                                Training products of experts by minimizing contrastive di
## 17                                     Improving neural networks by preventing co-adaptation of feature d
## 18                           A view of the EM algorithm that justifies incremental, sparse, and other
## 19                                                  Phoneme recognition using time-delay neural
## 20                                                 Learning multiple layers of features from ti
##                                                                author
## 1                                    DE Rumelhart, GE Hinton, RJ Williams
## 2                                    DE Rumelhart, GE Hinton, RJ Williams
## 3                                    A Krizhevsky, I Sutskever, GE Hinton
## 4                                    DE Rumelhart, GE Hinton, RJ Williams
## 5                                    DE Rumelhart, GE Hinton, RJ Williams
## 6                                             Y LeCun, Y Bengio, G Hinton
## 7                                          GE Hinton, S Osindero, YW Teh
## 8  N Srivastava, G Hinton, A Krizhevsky, I Sutskever, R Salakhutdinov
## 9                                  JL McClelland, DE Rumelhart, GE Hinton
## 10                                            GE Hinton, RR Salakhutdinov
## 11                                   DE Rumelhart, GE Hinton, RJ Williams
## 12                                   DE Rumelhart, GE Hinton, RJ Williams
## 13                                   A Krizhevsky, I Sutskever, GE Hinton
## 14                                   DE Rumelhart, GE Hinton, RJ Williams
## 15                                   DE Rumelhart, GE Hinton, RJ Williams
## 16                                            Y LeCun, Y Bengio, G Hinton
## 17                                         GE Hinton, S Osindero, YW Teh
## 18 N Srivastava, G Hinton, A Krizhevsky, I Sutskever, R Salakhutdinov
## 19                                 JL McClelland, DE Rumelhart, GE Hinton
## 20                                            GE Hinton, RR Salakhutdinov
##                                                                                journal
## 1  Parallel Distributed Processing: Explorations in the Microstructure of ..., 1986
## 2                                                          Nature 323, 533-536, 1986
## 3                Advances in neural information processing systems, 1097-1105, 2012
## 4                                  CALIFORNIA UNIV SAN DIEGO LA JOLLA INST FOR, 1985
## 5                                                 nature 323 (6088), 533, 1986
## 6                                                 nature 521 (7553), 436, 2015
## 7                                       Neural computation 18 (7), 1527-1554, 2006
## 8                    The Journal of Machine Learning Research 15 (1), 1929-1958, 2014
## 9  Parallel distributed processing: Explorations in the microstructure of ..., 1986
## 10                                          science 313 (5786), 504-507, 2006
## 11 Parallel Distributed Processing: Explorations in the Microstructure of ..., 1986
## 12                                                         Nature 323, 533-536, 1986
## 13               Advances in neural information processing systems, 1097-1105, 2012
## 14                                 CALIFORNIA UNIV SAN DIEGO LA JOLLA INST FOR, 1985
```

```
## 15                                                    nature 323 (6088), 533, 1986
## 16                                                    nature 521 (7553), 436, 2015
## 17                                  Neural computation 18 (7), 1527-1554, 2006
## 18              The Journal of Machine Learning Research 15 (1), 1929-1958, 2014
## 19 Parallel distributed processing: Explorations in the microstructure of ..., 1986
## 20                                                science 313 (5786), 504-507, 2006
##    year num_citation
## 1  1986        44497
## 2  1986        39827
## 3  2012        28366
## 4  1985        25393
## 5  1986        15724
## 6  2015         9610
## 7  2006         8898
## 8  2014         7893
## 9  1986         7751
## 10 2006         7665
## 11 2008         5629
## 12 2012         4608
## 13 2010         3964
## 14 1991         3753
## 15 1985         3643
## 16 2002         3440
## 17 2012         3027
## 18 1998         2646
## 19 1990         2631
## 20 2009         2624
```

(c)

```
## Include checks in the code in (a) and carry out some tests.

library(testthat)
```

```
##
## Attaching package: 'testthat'
## The following objects are masked from 'package:magrittr':
##
##     equals, is_less_than, not
```

```
library(assertthat)

get_page <- function(name){

# check if the input is valid
is_valid <- function(x) {
  is.character(x)
}
on_failure(is_valid) <- function(call, env) {
  "invalid input!"
}
assert_that(is_valid(name))
```

```
# get the user ID
id <- read_html(paste0("https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=", name, "&oq=geof"))%>%
html_nodes(".gs_rt2") %>% sub(".*user=([A-Za-z-]+)&.*","\\1",.)

# check if Google Scholar returns a result
have_result <- function(x) {
  length(x) != 0
}
on_failure(have_result) <- function(call, env) {
  print("can't find a result!")
}
assert_that(have_result(id))

# get the citation page
page <- read_html(paste0("https://scholar.google.com/citations?user=",id, "&hl=en&oi=ao"))

list <- list(id, page)
names(list) <- c("id", "page")
return (list)
}

# carry out some tests
test_that("get_page can detect invalid input and check if there is no result",{
  name1 <- "geoffreyhinton"
  name2 <- "trevorhastie"
  name3 <- "shubeiwang"

  expect_type(get_page(name1),'list')
  expect_type(get_page(name2),'list')
  expect_error(get_page(name3))
})
```

```
## [1] "can't find a result!"
```

(d)

```
## Fix the function in (b) so that it gets all of the results for a researcher.
## usage: get_all_article(x), argument x is the character string of the name of the researcher.

get_all_article <- function(name){

# get user ID
id <- get_page(name)$id

# create an empty data frame for future use
data <- data.frame(
  title = c(NA),
  author = c(NA),
  journal = c(NA),
  year = c(NA),
  num_citation = c(NA) )
```

```r
# use a loop to get all the articles
for(i in 0:100)
  {

# set pagesize = 100
site <- read_html(paste0("https://scholar.google.com/citations?user=", id,"&hl=en&cstart=",as.character

# break if there's no result in that page
message <- site %>% html_nodes(".gsc_a_e") %>%
html_text()
if(length(message)!=0) break

title <- site %>% html_nodes( ".gsc_a_at") %>% html_text()
len <- length(title)

author <- site %>%  html_nodes(".gs_gray") %>% html_text() %>%
as.data.frame(stringAsFactors=FALSE) %>%
.[seq(1, 2*len, by = 2),]

journal <- site %>%  html_nodes(".gs_gray") %>% html_text() %>%
as.data.frame(stringAsFactors=FALSE) %>% .[seq(0, 2*len, by = 2),]

year <- suppressWarnings(site %>% html_nodes(".gsc_a_y") %>%
html_text() %>% as.numeric())
year <- year[-1][-1]

num_citation <- site %>% html_nodes(".gsc_a_ac") %>%
html_text() %>% as.numeric(.) %>% replace(is.na(.),0)

data_app <- data.frame(
  title = title,
  author = author,
  journal = journal,
  year = year,
  num_citation = num_citation
)

# combine the data from each page
data <- rbind(data, data_app)
}
data <- data[-1,]
return (data)
}

# store all data of Trevor Hastie in alldata
alldata <- get_all_article("trevorhastie")
```

(4)

When webscraping data from Google Scholar, we should comply to the rules set by it according to the robot.txt file. It shows that the website allow partial access for crawling. We should avoid crawling the blocked areas such as “https://scholar.google.com/citations?”, etc. Also we should not make queries too

frequently.