ps2

Baoyue Liang

9/10/2018

Problem 1

I mind the following tips in my code:

1. I break down tasks into core units.
2. I write functions that take data as an argument and not lines of code that operate on specific data objects.
3. I build fuctions with single task, meaningful name and comment on its purpose.
4. I use variables instead of hard code numbers

Problem 2

(a)

For csv, each random number has 11 digits, 1 dot, 50% possibility to have a minus sign, and followed by a comma/line break, which means each random number would take up 13.5 bytes on average. However, if a random number end with 0, 00, 000 etc., the zeros will not be saved. This could make each random number 0.11111 bite shorter. (13.5-0.111)*10ˆ7 = 13.389*10ˆ7, which is very close to 133887710.

For rda, there are 10ˆ7 numbers, each saved as double float and takes up 8 bytes. Therefore, the file size is 8*10ˆ7 bite.

(b)

Each random is either followed by a comma or a line break (/n). Even though commas are no longer saved, line break (/n) would replace comma and still take up 1 byte.

(c)

Comparation 1 As for read.csv, unless colClasses is specified, all columns are read as character columns and then converted using type.convert to logical, integer, numeric, complex or (depending on as.is) factor as appropriate. However, scan treats the random numbers in the csv files as numeric by default. Therefore, scan is faster than read.csv.

Comparation 2 Specifying the colClasses argument explicitly make it for faster for read.csv to read files. Here, numric are explicitly assigned. Therefore the two method takes appximately the same time.

Comparation 3 On one hand, the rda file is smaller. On the other hand, it saved the time for string processing since the random number is directly saved as double float in .rda. THat's why the load command is way faster.

(d)

The save fuction has a compress argument, compress = isTRUE(!ascii), which will compress the file it it is not ASCII. In b, all of the numbers are the same. Therefore, the file can be compressed in to smaller size.

Problem 3

(a)

```
library(xml2)
library(rvest)

library(assertthat)
library(testthat)

get_http_ID = function(authorname){

  nametrans = gsub(" ","+",authorname)

  #test required in question (c)
  assert_that(is.character(nametrans))

  URL1 = paste("https://scholar.google.com/citations?view_op=search_authors&mauthors=",nametrans,"&hl=en

  links <- read_html(URL1) %>% html_nodes("a") %>% html_attr('href')

  authorID = grep("\\?user=",links,value = TRUE)

  # fetch the line of url with author id
  authorID = substring(authorID,17,(nchar(authorID)-15))
  expect_length(authorID, 1)

  URL2 = paste("https://scholar.google.com/citations?user=",authorID,"&hl=en",sep="")
  html1 = read_html(URL2)

  print(authorID)
  return(html1)

}

get_http_ID("Trevor Hastie")
```

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

```
## {xml_document}
## <html>
## [1] <head>\n<title>Trevor Hastie - Google Scholar Citations</title>\n<me ...
## [2] <body><div id="gs_top" onclick="">\n<style>#gs_md_s,.gs_md_wnw{z-ind ...
```

(b)

```
#
get_citation = function(html){
  tbls = html_table(html_nodes(html, "table"))
  tbls[[2]]
}

tbles = get_citation(http_ID1)
head(tbles)
```

```
##
## 1
## 2 National, regional, and global trends in body-mass index since 1980: systematic analysis of health
## 3     National, regional, and global trends in fasting plasma glucose and diabetes prevalence since
## 4            on behalf of the Global Burden of Metabolic Risk Factors of Chronic Diseases Collaborat
## 5       National, regional, and global trends in systolic blood pressure since 1980: systematic analy
## 6                                                                                              Na
##
## 1 Cited by Year
## 2     3722 2011
## 3     3304 2011
## 4    1077* 2011
## 5      884 2011
## 6      618 2012
```

(d)

I put sebsection (d) first since I would like to write the test function for the get_all_result function.

```
get_all_result = function(authorname){

  nametrans = gsub(" ","+",authorname)

  #test required in question (c)
  assert_that(is.character(nametrans))

  URL1 = paste("https://scholar.google.com/citations?view_op=search_authors&mauthors=",nametrans,"&hl=e

  links <- read_html(URL1) %>% html_nodes("a") %>% html_attr('href')

  authorID = grep("\\?user=",links,value = TRUE)

  # fetch the line of url with author id
  authorID = substring(authorID,17,(nchar(authorID)-15))
  expect_length(authorID, 1)

  tbls = data.frame()
  tbls = tbls[-1,]

  i = 1

  while (TRUE) {
    URL2 = paste("https://scholar.google.com/citations?user=",authorID,"&hl=en&cstart=",i,"&pagesize=100
    i = i + 100

    html1 = read_html(URL2)
    tbls1 = html_table(html_nodes(html1, "table"))
    tbls1 = tbls1[[2]]
    tbls = rbind(tbls,tbls1[-1,])
    nrow = nrow(tbls1)
    if (nrow < 101) { break }
  }

  return (tbls)
}

output = get_all_result("Christopher Paciorek")
head(output)
```

```
##
## 2          National, regional, and global trends in fasting plasma glucose and diabetes prevalence si
## 3                 on behalf of the Global Burden of Metabolic Risk Factors of Chronic Diseases Collab
## 4            National, regional, and global trends in systolic blood pressure since 1980: systematic
## 5
## 6 Global, regional, and national trends in haemoglobin concentration and prevalence of total and seve
## 7                National, regional, and global trends in serum total cholesterol since 1980: systema
##
## 2  3304 2011
## 3 1077* 2011
## 4   884 2011
## 5   618 2012
## 6   529 2013
## 7   456 2011
```

(c)

```
test_that("WRFP", {
    expect_error(get_all_result("ASDFGSJKD"),"")
})

test_that("Trevor Hastie", {

    tbles = get_all_result("Trevor Hastie")
    expect_is(tbles,"data.frame")
})
```

Problem 4

As is shown in the “https://scholar.google.com/robots.txt”, it is disallowed to perform “/search”, “/index.html”, “/scholar”, “/citations?”, “/citations?*cstart=“,”/citations?user*=%40”,"/citations?user=*@".But it is allowed to search “/citations?user=”. What we search is “https://scholar.google.com/citations?user=xUXVgn8AAAAJ&hl=en”. As far as I am concerned, i believe that what we are doin is legal.