ps2

Ramon Crespo

9/12/2018

Problem1 Some concepts that I used in the implementation of this homework are: 1. version control 2. good syntax

Problem2 Part a. In the csv file each value is stored separately as a ASCII, meaning each value takes a bite. Since each entry is 12 characters long, the size of the file should be 12*1e7. If you add the commas and the occasional legative sign, then the file size would increase to a value close to 133,887,710. The Rda file is able to compress the information into a specific file format and use less than a bite per character.

Part b. Because there still needs to be a separator between the rows. In this case the character separating the columns was replaced by a separator of rows.

Part c. First Comparison Scan looks for a specific entry, in this case ",". It is faster as it does not need to look into each of the characters (1e7 * 12), but instead it just looks for the commas and returns the code that matched the parameter.

Second Comparison Because you gave the command "numeric", the program only needs to look into where the separators are. So what the program is doing is just looking into where the commas are and then whatever matches the search its being assigned the character class. You save the program the time it would have needed to search in the ASCII table the values to then be able to asign them the numeric class. The time needed is less than the scan because even when they are looking for the same entrys, ",", the scan still needs to look up what type of entry it is getting.

Third Comparison Rda is just better at this task because it the program does not go character by character looking for the pattern that matches. Instead the program already knows where to look, thus making the calculation almost 100 times faster. Rda is just better at this, thus it should be used when possible.

```
## First comparison
system.time(a0 <- read.csv('/tmp/tmp.csv', header = FALSE))
```

```
##    user  system elapsed
##  48.237   1.520  51.853
```

```
##    user  system elapsed
##  35.348   0.248  35.599
system.time(a1 <- scan('/tmp/tmp.csv', sep = ','))
```

```
##    user  system elapsed
##   3.599   0.222   3.900
```

```
##    user  system elapsed
##   5.236   0.024   5.261

## Second comparison
system.time(a0 <- read.csv('/tmp/tmp.csv',header = FALSE, colClasses = 'numeric'))
```

```
##    user  system elapsed
##   3.671   0.219   3.965
```

```
##    user  system elapsed
##   5.236   0.044   5.281
system.time(a1 <- scan('/tmp/tmp.csv', sep = ','))
```

```
##    user  system elapsed
```

```
##   3.678   0.231   4.024
```

```
##    user  system elapsed
##   5.256   0.032   5.289

## Third comparison
system.time(a1 <- scan('/tmp/tmp.csv', sep = ',')) ## user system elapsed
```

```
##    user  system elapsed
##   3.755   0.101   3.931
```

```
##   5.288   0.020   5.307
system.time(load('/tmp/tmp.Rda'))
```

```
##    user  system elapsed
##   0.292   0.063   0.373
```

```
##    user  system elapsed
##   0.076   0.008   0.085
```

Part d The number of entries and the size of the numbers is the same for both cases. The difference comes in the information that the compressed .Rda file needs to keep to remember the matrix format. This information ends up making the file many times larger than a file that would just keep the values.

Problem3 #Parta. Specify the researcher's name and return the HTML for the researcher's citation page.

```
library(curl)
library(rvest)
```

```
## Loading required package: xml2
```

```
library(assertthat)
library(testthat)

researcher_scraper <- function(name){
  #Step0. Revise for correct input format. Check for type of input and length.
  #Input = text
  #Length = 2. Name and LastName
  assert_that(is.character(name))

  name_input <- strsplit(name," ")
  name_input <- unlist(name_input)
  stopifnot(length(name_input)==2)

  #Step1. input researcher's name and obtain the search result information.
  researcher <- name
  researcher_name <- regmatches(researcher, regexpr("^.[[:alpha:]]+", researcher))
  researcher_lastname <- regmatches(researcher, regexpr(" [[:alpha:]]*", researcher))
  researcher_lastname <-gsub(" ", "", researcher_lastname, fixed = TRUE)
  URL <- "https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=<name>+<lastname>&btnG="
  URL <- sub("<name>", researcher_name, URL)
  URL <- sub("<lastname>",researcher_lastname, URL)
  html <- read_html(URL)

  ##Step 2 -> Search for the citation page of the researcher and outputs the html text
  links <- read_html(URL) %>% html_nodes("a") %>% html_attr('href')
  link <- links[41]
  researcher_id <- regmatches(link, regexpr("?=.{12}", link))
```

```
  researcher_id <-gsub("=", "", researcher_id, fixed = TRUE)
  new_url <- c("https://scholar.google.com", links[41])
  new_url <- paste(new_url, collapse="")
  html <- read_html(new_url)
  my_list <- list(researcher_id, html,new_url)
return(my_list)
}



researcher_articles <- function(name){
  #obtain url of citation page for the specified scholar
  information <- researcher_scraper(name)
  URL <- information[[3]]

  #Extract the title of the articles. Its a little verbose, but the different steps makes it clear what
  art_titles <- read_html(URL) %>% html_nodes("a")
  art_titles_clean <- grep("class=\"gsc_a_at\".*",art_titles, value=TRUE)
  art_titles_clean <- regmatches(art_titles_clean, regexpr(">(.*)<", art_titles_clean))
  art_titles_clean <-gsub("<", "", art_titles_clean, fixed = TRUE)
  art_titles_clean <-gsub(">", "", art_titles_clean, fixed = TRUE)
  art_titles_clean

  #Extract the relevant information from the different papers
  art_info <- read_html(URL) %>% html_nodes("div")
  art_info <- grep("^<div class=\"gs_gray\".*",art_info, value=TRUE)

  art_info_authors <- grep("class=\"gs_oph\">",art_info, value=TRUE, invert = TRUE)
  art_info_authors <- regmatches(art_info_authors, regexpr(">(.*)<", art_info_authors))
  art_info_authors <-gsub("<", "", art_info_authors, fixed = TRUE)
  art_info_authors <-gsub(">", "", art_info_authors, fixed = TRUE)
  art_info_authors

  art_info_journal <- grep("class=\"gs_oph\">",art_info, value=TRUE)
  art_info_journal <- regmatches(art_info_journal, regexpr(">(.*)<span", art_info_journal))
  art_info_journal <-gsub("<span", "", art_info_journal, fixed = TRUE)
  art_info_journal <-gsub(">", "", art_info_journal, fixed = TRUE)
  art_info_journal

  art_info_year <- grep("class=\"gs_oph\">",art_info, value=TRUE)
  art_info_year <- regmatches(art_info_year, regexpr("<span class=\"gs_oph\">, [[:digit:]]*", art_info_y
  art_info_year <-gsub("<span class=\"gs_oph\">, ", "", art_info_year, fixed = TRUE)
  art_info_year

  art_info_citations <- read_html(URL) %>% html_nodes("td")
  art_info_citations <- regmatches(art_info_citations, regexpr("class=\"gsc_a_ac gs_ibl\">[[:digit:]]*"
  art_info_citations <-gsub("class=\"gsc_a_ac gs_ibl\">", "", art_info_citations, fixed = TRUE)
  art_info_citations

  #Generate and return dataframe
  d <- data.frame("titles"=art_titles_clean, "authors" = art_info_authors, "journal"=art_info_journal,
  return(d)
}
```

```
TrevorHastie <- researcher_articles("Trevor Hastie")
ScottMoura <- researcher_articles("Scott Moura")

test_that(
  "Check if first article correspond to online information",
  {
    expect_that(as.vector(ScottMoura$titles[1]),equals("A stochastic optimal control approach for power
    expect_equal(as.vector(ScottMoura$authors[1]),"SJ Moura, HK Fathy, DS Callaway, JL Stein")
    expect_equal(as.vector(ScottMoura$journal[1]),"IEEE Transactions on control systems technology 19 (3
    expect_equal(as.vector(ScottMoura$year[1]), "2011")
    expect_equal(as.vector(ScottMoura$citations[1]), "480")
  }
)
TrevorHastie
```

```
##
## 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 for
## 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
##                                                                 authors
## 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                                                            E Candes, T Tao
## 12                                          J Friedman, T Hastie, R Tibshirani
## 13                                                      JM Chambers, TJ Hastie
## 14                  J Elith, SJ Phillips, T Hastie, M Dudík, YE Chee, CJ Yates
## 15                                   R Tibshirani, T Hastie, B Narasimhan, G Chu
## 16   O Troyanskaya, M Cantor, G Sherlock, P Brown, T Hastie, R Tibshirani, ...
```

```
## 17                                                J Elith, JR Leathwick, T Hastie
## 18                                                  H Zou, T Hastie, R Tibshirani
## 19                                                         T Hastie, R Tibshirani
## 20                                                         T Hastie, R Tibshirani
##                                                                            journal
## 1                                    The elements of statistical learning, 485-585
## 2                                                 Statistical models in S, 249-307
## 3           Proceedings of the National Academy of Sciences 98 (19), 10869-10874
## 4  Journal of the Royal Statistical Society: Series B (Statistical Methodology ...
## 5                                      The Annals of statistics 32 (2), 407-499
## 6                                      The annals of statistics 28 (2), 337-407
## 7                                    Journal of statistical software 33 (1), 1
## 8                                                                       springer
## 9  Journal of the Royal Statistical Society: Series B (Statistical Methodology ...
## 10                                        Springer series in statistics 1 (10)
## 11                                   The Annals of Statistics 35 (6), 2313-2351
## 12                                                Biostatistics 9 (3), 432-441
## 13                     Wadsworth &amp;amp; Brooks/Cole Advanced Books &amp;amp; Software
## 14                                   Diversity and distributions 17 (1), 43-57
## 15            Proceedings of the National Academy of Sciences 99 (10), 6567-6572
## 16                                                Bioinformatics 17 (6), 520-525
## 17                                   Journal of Animal Ecology 77 (4), 802-813
## 18              Journal of computational and graphical statistics 15 (2), 265-286
## 19  Journal of the Royal Statistical Society. Series B (Methodological), 757-796
## 20                     Advances in neural information processing systems, 507-513
##    year citations
## 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
```

```
ScottMoura
```

```
##
## 1                                        A stochastic optimal control approach for power manageme
## 2                                      Plug-in hybrid electric vehicle charge pattern optimization
## 3      Adaptive Partial Differential Equation Observer for Battery State-of-Charge/State-of-Health Es
## 4        Battery-health conscious power management in plug-in hybrid electric vehicles via electroch
## 5  Genetic identification and fisher identifiability analysis of the Doyle-Fuller-Newman model from
```

```
## 6                    Tradeoffs between battery energy capacity and stochastic optimal power manageme
## 7                                                  Velocity Predictors for Predictive Energy Ma
## 8                                                PDE estimation techniques for advanced battery mana
## 9                           Integrated optimization of battery sizing, charging, and power manageme
## 10         Stochastic control of smart home energy management with plug-in electric vehicle battery
## 11                                          Dynamic traffic feedback data enabled energy manageme
## 12                                  Quantifying EV battery end-of-life through analysis of travel
## 13                                                          Air flow control in fuel cell
## 14                                                Optimal control of film growth in lithiu
## 15                                                                  Adaptive PDE ob
## 16                                  Impact of battery sizing on stochastic optimal power manageme
## 17                                                  Lyapunov-based switched extremum seeki
## 18                              On the aggregate grid load imposed by battery health-conscious chargi
## 19                   Asymptotic convergence through Lyapunov-based switching in extremum seeking wi
## 20             Charge trajectory optimization of plug-in hybrid electric vehicles for energy cost re
##                                          authors
## 1           SJ Moura, HK Fathy, DS Callaway, JL Stein
## 2            S Bashash, SJ Moura, JC Forman, HK Fathy
## 3                  SJ Moura, NA Chaturvedi, M Krstić
## 4                      SJ Moura, JL Stein, HK Fathy
## 5             JC Forman, SJ Moura, JL Stein, HK Fathy
## 6           SJ Moura, DS Callaway, HK Fathy, JL Stein
## 7                      C Sun, X Hu, SJ Moura, F Sun
## 8                  SJ Moura, NA Chaturvedi, M Krstic
## 9        X Hu, SJ Moura, N Murgovski, B Egardt, D Cao
## 10              X Wu, X Hu, S Moura, X Yin, V Pickert
## 11           C Sun, SJ Moura, X Hu, JK Hedrick, F Sun
## 12         S Saxena, C Le Floch, J MacDonald, S Moura
## 13                               YA Chang, SJ Moura
## 14 SJ Moura, JC Forman, S Bashash, JL Stein, HK Fathy
## 15                  SJ Moura, M Krstic, NA Chaturvedi
## 16          SJ Moura, DS Callaway, HK Fathy, JL Stein
## 17                               SJ Moura, YA Chang
## 18                    S Bashash, SJ Moura, HK Fathy
## 19                               SJ Moura, YA Chang
## 20                    S Bashash, SJ Moura, HK Fathy
##                                                                                journal
## 1                IEEE Transactions on control systems technology 19 (3), 545-555
## 2                                    Journal of power sources 196 (1), 541-549
## 3           Journal of Dynamic Systems, Measurement, and Control 136 (1), 011015
## 4                IEEE Transactions on Control Systems Technology 21 (3), 679-694
## 5                                        Journal of Power Sources 210, 263-275
## 6                                   Journal of Power Sources 195 (9), 2979-2988
## 7                                IEEE Trans. Contr. Sys. Techn. 23 (3), 1197-1204
## 8                                American Control Conference (ACC), 2012, 559-565
## 9              IEEE Transactions on Control Systems Technology 24 (3), 1036-1043
## 10                                       Journal of Power Sources 333, 203-212
## 11             IEEE Transactions on Control Systems Technology 23 (3), 1075-1086
## 12                                       Journal of Power Sources 282, 265-276
## 13                     American Control Conference, 2009. ACC'09., 1052-1059
## 14                 IEEE Transactions on Industrial Electronics 58 (8), 3555-3566
## 15 ASME 2012 5th Annual Dynamic Systems and Control Conference joint with the ...
## 16     IEEE International Conference on Vehicular Electronics and Safety, 22-24
## 17                                  Control Engineering Practice 21 (7), 971-980
```

```
## 18                                  Journal of Power Sources 196 (20), 8747-8754
## 19                              American Control Conference (ACC), 2010, 3542-3548
## 20                              American Control Conference (ACC), 2010, 5824-5831
##    year citations
## 1  2011       480
## 2  2011       277
## 3  2014       130
## 4  2013       130
## 5  2012       124
## 6  2010       122
## 7  2015       120
## 8  2012       102
## 9  2016        99
## 10 2016        94
## 11 2015        91
## 12 2015        85
## 13 2009        53
## 14 2011        52
## 15 2012        44
## 16 2008        39
## 17 2013        35
## 18 2011        35
## 19 2010        35
## 20 2010        34
```

Problem 4 Problem 3 is not unethical and it follows good web scraping ethics. The objective of problem 3 is to summarize the information from certain a certain scholar and present this information in an organized manner to the user. Obtaining this information through a webscraper is not unethical because: i) only requires a small amount of information, ii) by being able to perform the task, google is authorizing third parties to do it and iii) the information we obtain from the process is publicly available thus we are not stealing information. An unethical practice would violate one of the above rules. The process we performed could be summarized as rapidly organizing the information that would have taken 10 minutes in less than 30 seconds.

The robot.txt file is an efficient manner of communicating with the robot (scraper computer) and setting friendly terms on the information sharing process. It is a common language for webscrapers. This files contains information regarding specific parts of the webpage that should not be accessed, delays in scrapping etc. Large companies, like google, will follow the information in this .txt file.