Adquisición y preparación de nuestros datos en R

Carlos Andrés Saldaña Amézquita

1. LECTURA DE FICHEROS CSV

auto <- read.csv("1_PreparacionDeData/data/auto-mpg.csv", header = TRUE, sep = ",")

names(auto)

Lectura de ficheros cuyo separador es un espacio

read_table == read.csv("filename", sep = " ")

Lectura de ficheros cuyo separador es ";" y la data emplea coma decimal

read_csv2 == read.csv("filename") | por defecto: sep = ";", dec = ","

Imprimimos el nombre de las cabeceras -> names(data_name)

Dataframe sin cabecera

auto_no_header <- read.csv("1_PreparacionDeData/data/auto-mpg-noheader.csv", header = FALSE, sep = ",")

Imprimimos las 5 primeras líneas:

head(auto_no_header, 5)

Dataframe con cabecera personalizada

auto_custom_header <- read.csv("1_PreparacionDeData/data/auto-mpg-noheader.csv", header = FALSE, col.names = c("Número", "Millas_por_galeón", "Cilindrada","Desplazamiento", "Caballos_de_potencia","Peso","Aceleración","Año_del_modelo", "Modelo_del_coche"))

head(auto_custom_header, 2)

Uso de los NA (los valores " " pasan a ser identificados como NA por R)

- Los strings pueden ser tratados como caracteres o factores (por defecto)
- stringAsFactors = TRUE (por defecto R convierte los strings a factores)

auto_NA <- read.csv("1_PreparacionDeData/data/auto-mpg.csv", header = TRUE, sep = ",", strings.na = " ", stringsAsFactors = False)

Carga de datos online

who_from_internet <- read.csv("https://frogames.es/course-contents/r/intro/tema1/WHO.csv", header = TRUE)

- NA: no available
- na.strings = " "
- as.character() -> de variable categórica (factor) a caracter

2. LECTURA DE FICHEROS XML

1. Instalación de paquetes install.packages("XML")
2. Cargamos la librería library(XML)

Guardamos el fichero a emplear

url <- "1_PreparacionDeData/data/cd_catalog.xml"

Puntero que localiza el documento

DataXML <- xmlParse(url)

Obtenemos la estructura de datos del nodo raíz

rootnode <- xmlRoot(DataXML)

Consultamos el primer elemento

rootnode[1]

Recorremos el nodo raíz y aplicamos una función

- x: cada elemento del nodo raíz
- function(x) xmlSApply(x, xmlValue): Extraemos únicamente los valores

cds_data <- xmlSApply(rootnode, function(x) xmlSApply(x, xmlValue))

Ordenamos nuestra tabla

- row.names = NULL: Las filas no tienen nombre

cds.catalog <- data.frame(t(cds_data), row.names = NULL)

Visualización de la data

head(cds.catalog, 3)

Algunas funciones útiles

- xpathSApply()
- getNodeSet()

3. LECTURA DE UN FICHERO HTML

population_url <- "1_PreparacionDeData/data/WorldPopulation-wiki.htm"

tables <- readHTMLTable(population_url) #Extracción de todas tablas

most_populated <- tables[[6]]

head(most_populated, 3)

Extraemos una columna

custom_table <- readHTMLTable(population_url, which = 6)

4. LECTURA DE FICHEROS JSON

Instalación y carga de librerías

install.packages("jsonlite")

install.packages("curl")

library(jsonlite)

library(curl)

Lectura de ficheros locales

data_1 <- fromJSON("1_PreparacionDeData/data/students.json")

data_2 <- fromJSON("1_PreparacionDeData/data/student-courses.json")

Lectura de ficheros online

url <- "http://www.floatrates.com/daily/usd.json"

currencies <- fromJSON(url)

Símbolo $ (acceso a nuestra data)

currencies_data <- currencies$eur | Accedemos a euros

head(data_1, 3)

data_1$Email | Accedemos a Email

5. LECTURA DE FICHEROS DE ANCHO FIJO

Cargamos nuestro fichero de ancho fijo

widths: Tamaño del elemento más grande de la columna

students_data <- read.fwf("1_PreparacionDeData/data/student-fwf.txt", widths = c(4, 15, 20, 15, 4), col.names = c("Id", "Nombre", "Email", "Carrera", "Año"))

Cargamos un fichero de ancho fijo con cabeceras

- sep: Separador empleado por las cabeceras
- skip: Saltamos las 2 primeras líneas del fichero

students_data_header <- read.fwf("1_PreparacionDeData/data/student-fwf-header.txt", widths = c(4, 15, 20, 15, 4), header = TRUE, sep = "|", skip = 2)

Cargamos toda la data, a excepción del email

students_data_no_email <- read.fwf("1_PreparacionDeData/data/student-fwf.txt", widths = c(4, 15, -20, 15, 4), col.names = c("Id", "Nombre", "Carrera", "Año"))

6. LECTURA DE FICHEROS DE R

Creamos algunos objetos de R

clientes <- c("Carlos", "Manuel", "Luis")

pago <- c(250, 88.52, 174.99)

pedidos <- data.frame(clientes, fechas, pago)

Objeto tipo fecha (otación anglosajona)

fechas <- as.Date(c("2018-11-25", "2018-11-5", "2018-4-2"))

Guardamos uno o más objetos de R (.Rdata)

save(pedidos, file = "1_PreparacionDeData/data/pedidos.Rdata")

Guardamos únicamente un objeto de R (.rds)

saveRDS(pedidos, file = "1_PreparacionDeData/data/pedidos.rds")

Limpieza de variables del workspace

remove(pedidos)

Carga de ficheros Rdata y rds

load("1_PreparacionDeData/data/pedidos.Rdata")

orders_rds <- readRDS("1_PreparacionDeData/data/pedidos.rds")

Algunos datasets que vienen por defecto en R

data()

data(iris)

data(cars)

Guardamos todos los objetos de la sesión

save.image(file = "1_PreparacionDeData/data/alldata.Rdata")

Guardamos objetos de manera selectiva

primes <- c(2, 3, 5, 7, 11, 13)

pow2 <- c(2, 4, 8, 16, 32, 64, 128)

save(list = c("primes", "pow2"), file = "1_PreparacionDeData/data/primes_and_pow2.Rdata")

Cargamos objetos (recibimos una notificación si el objeto ya existe)

attach("1_PreparacionDeData/data/primes_and_pow2.Rdata")