SNS COLLEGE OF ENGINEERING

Kurumbapalayam (Po), Coimbatore – 641 107

An Autonomous Institution

Accredited by NAAC – UGC with 'A' Grade
Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

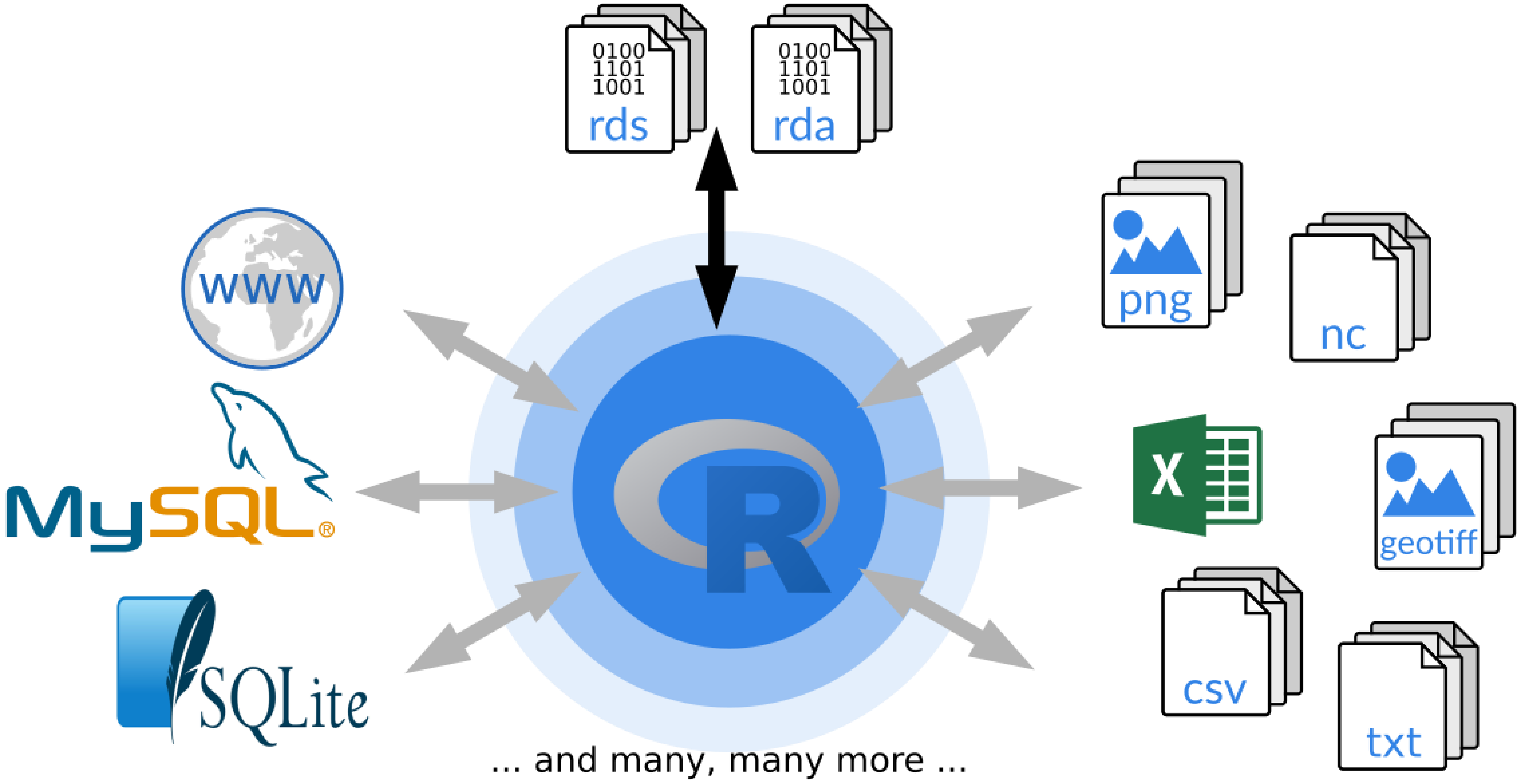
DEPARTMENT OF COMPUTER SCIENCE AND TECHNOLOGY

COURSE NAME :19CS407 DATA ANALYTICS WITH R

II YEAR /IV SEMESTER

Unit 5- DATA VISUALIZATION USING R

Topic : Web Data



... and many, many more ...

R - Web Data

- ✓ Many websites provide data for consumption by its users. For example the World Health Organization(WHO) provides reports on health and medical information in the form of CSV, txt and XML files.

- ✓ Using R programs, we can programmatically extract specific data from such websites. Some packages in R which are used to scrap data form the web are − "RCurl",XML", and "stringr".

- ✓ They are used to connect to the URL’s, identify required links for the files and download them to the local environment.

Install R Packages

The following packages are required for processing the URL’s and links to the files. If they are not available in your R Environment, you can install them using following commands.

```
install.packages("RCurl")
install.packages("XML")
install.packages("stringr")
install.packages("plyr")
```

Input Data

We will visit the URL weather data and download the CSV files using R for the year 2015.

https://www.geos.ed.ac.uk/~weather/jcmb_ws/

Example

We will use the function getHTMLLinks() to gather the URLs of the files. Then we will use the function download.file() to save the files to the local system.

As we will be applying the same code again and again for multiple files, we will create a function to be called multiple times.

The filenames are passed as parameters in form of a R list object to this function.

Example

```
# Read the URL.
url <- "http://www.geos.ed.ac.uk/~weather/jcmb_ws/"

# Gather the html links present in the webpage.
links <- getHTMLLinks(url)

# Identify only the links which point to the JCMB 2015 files.
filenames <- links[str_detect(links, "JCMB_2015")]

# Store the file names as a list.
filenames_list <- as.list(filenames)

# Create a function to download the files by passing the URL and filename list.
downloadcsv <- function (mainurl,filename) {
  filedetails <- str_c(mainurl,filename)
  download.file(filedetails,filename)
}

# Now apply the l_ply function and save the files into the current R working directory.
l_ply(filenames,downloadcsv,mainurl = "http://www.geos.ed.ac.uk/~weather/jcmb_ws/")
.
```

Verify the File Download

After running the above code, you can locate the following files in the current R working directory.

"JCMB_2015.csv" "JCMB_2015_Apr.csv" "JCMB_2015_Feb.csv" "JCMB_2015_Jan.csv"

"JCMB_2015_Mar.csv"

Assessment 1

References

1. João Moreira, Andre Carvalho, Tomás Horvath – “A General Introduction to Data Analytics” – Wiley -2018

2. https://www.tutorialspoint.com/r/r_web_data.htm

Thank You