DataCollection\_WebScrapping.R

Admin

Thu Aug 01 18:56:07 2019

library(rvest)

## Loading required package: xml2

library(xml2)  
library(openxlsx)  
library(rio)  
library(tidyr)  
  
  
setwd("F:/Vivek")  
  
###############################################################  
# Collecting Data for population, density, sex ratio, litracy #  
###############################################################  
  
links = c("https://www.census2011.co.in/census/state/districtlist/west+bengal.html",  
 "https://www.census2011.co.in/census/state/districtlist/bihar.html",  
 "https://www.census2011.co.in/census/state/districtlist/jharkhand.html",  
 "https://www.census2011.co.in/census/state/districtlist/orissa.html",  
 "https://www.census2011.co.in/census/state/districtlist/assam.html")  
  
for (i in 1:5)  
{  
 url <- links[i]  
 webpage <- read\_html(url)  
   
 #converting html to dataframe  
 state\_html <- html\_nodes(webpage,xpath='/html/body/div[1]/div/div[1]/table')  
 state\_data <-html\_table(state\_html, fill = TRUE)  
 typeof(state\_data)  
 state\_data <- as.data.frame(state\_data);state\_data  
   
 #Cleaning the dataframe  
 statename = c("westBengal","bihar","jharkhand","orissa","Assam")  
 state\_data <- state\_data[,-c(1,3,9:99)]  
 colnames(state\_data) <- tolower(make.names(colnames(state\_data)))   
 state\_data <- state\_data[-c(8,16),]  
 state\_data$state <- statename[i]  
 assign(paste("statedata\_",statename[i], sep = ""), state\_data)  
 rm(state\_data)  
}   
  
collateddata <- rbind(statedata\_westBengal,statedata\_jharkhand,statedata\_orissa,statedata\_bihar,statedata\_Assam)  
#write.xlsx(masterdata, "sampel\_scarped\_data.xlsx")  
  
#############################################  
  
#State wise Unemployment rate  
  
url <- "https://unemploymentinindia.cmie.com/"  
webpage <- read\_html(url)  
  
# Converting html to datafram  
tbls <- html\_nodes(webpage, "table")  
unemp\_html <- html\_nodes(tbls, xpath = '/html/body/table/tbody/tr/td/table/tbody/tr[2]/td/table/tbody/tr/td[2]/table[2]')  
unempdata <- html\_table(tbls,fill = TRUE)  
#unempdata <- as.data.frame(unempdata)  
  
unempdata <- unempdata[[16]]  
  
######################################  
  
#Growth under MFI and SHGs (2015)  
  
mfgrowth <- read.csv("rs\_session\_239\_AU1837\_1.1\_1.csv")  
#mfgrowth <- subset(mfgrowth, mfgrowth$REGION == "EASTERN REGION" )  
mfgrowth <- mfgrowth[-c(1,6),-c(1,2)]  
  
######################################  
  
#GSDP per capita Data  
  
url <- "https://statisticstimes.com/economy/gdp-capita-of-indian-states.php"  
webpage <- read\_html(url)  
tbls <- html\_nodes(webpage, "table")  
  
gsdp\_html <- html\_nodes(tbls, xpath ='//\*[@id="table\_id"]')  
gsdp\_data <- html\_table(gsdp\_html,fill = TRUE)  
gsdp\_data <- as.data.frame(gsdp\_data)  
  
#Cleaning  
  
gsdp\_data <- gsdp\_data[,-c(1,6,8:10)]  
  
#####################################  
  
#Writing the excel file  
  
# Create a blank workbook  
masterdata <- createWorkbook()  
  
# Add some sheets to the workbook  
addWorksheet(masterdata, "Demographic")  
addWorksheet(masterdata, "UnempData")  
addWorksheet(masterdata, "MFIGrowth")  
addWorksheet(masterdata, "GSDPPCData")  
  
# Write the data to the sheets  
writeData(masterdata, sheet = "Demographic", x = collateddata)  
writeData(masterdata, sheet = "UnempData", x = unempdata)  
writeData(masterdata, sheet = "MFIGrowth", x = mfgrowth)  
writeData(masterdata, sheet = "GSDPPCData", x = gsdp\_data)  
  
# Reorder worksheets  
worksheetOrder(masterdata) <- c(1,2,3,4)  
  
# Export the file  
#saveWorkbook(masterdata, "Masterdata.xlsx")