#JSON包

library(bitops)

library(RCurl)

library(jsonlite)

#Selenium包

library(stringr) # String Processing

library(dplyr)# 调用%>%管道

library(RSelenium) #爬取动态网页

library(xml2)

library(rvest)

library(rjson)

##Selenium

remote = remoteDriver("localhost", 4444L, browserName="firefox")

remote$open()

#Visit the Page Sources

url = "http://data.eastmoney.com/cjsj/fdi.html"

remote$navigate(url)

#Create a NULL list

data\_list = list(NULL)

for(i in 2:7) #6個變量

{

data\_list[i]=NULL

}

#設置模板(網址在不同頁的型態)

mode = "http://data.eastmoney.com/cjsj/foreigndirectinvestment.aspx?p="

counter = 2

#Crawl the rows of data

for (z in 1:7) #7頁

{

#使每一頁的Page Sources都可以被獲取

tpage = remote$getPageSource()

PageSource = tpage[[1]]

web = read\_html(PageSource)

#在Page Sources的基礎下，找尋並提取所有的row

tbrows = web %>% html\_nodes("#tb>tbody") %>% html\_nodes("tr")

for(i in 2:length(tbrows)) #2:length，而非1，因為不包含第一row中的項目名稱

{

tds = tbrows[i] %>% html\_nodes("td")

for(j in 1:6) #6個變量(column)，故1:6

{

data\_list[[j]] = c(data\_list[[j]], gsub("[ \n]", "", tds[j] %>% html\_text()))

#Preserve what we get into our NULL list, and

#Use gsub to replace all the blank and /n with "", which is nothing

}

}

#Use the loop to crawl data repetedly

if (z == 7)

{

break

}

url2 = paste0(mode, counter)

remote$navigate(url2) #Direct R to navigate what is actually inside url2

counter = counter+1

}

#See what it actually is

print(data\_list)

#Create a data frame to match the data we just crawled

FDI = data.frame(

Month = data\_list[[1]],

Monthly.Volume\_USD = data\_list[[2]],

Year\_on\_year\_Growth = data\_list[[3]],

Month\_on\_Month\_Growth = data\_list[[4]],

Accumulation\_in\_Month = data\_list[[5]],

Year\_on\_year\_Growth\_Accumulation = data\_list[[6]],

stringsAsFactors = FALSE

)

#Write the data frame into the file

write.csv(FDI, "D:\\FDI\_Data\_China.csv")