동적 크롤링

Day08

install.packages("RSelenium")

library(RSelenium)

remDr <- remoteDriver(remoteServerAddr = "localhost" , port = 4445, browserName = "chrome")

remDr$open()

remDr$navigate("http://www.google.com/")

webElem <- remDr$findElement(using = "css", "[name = 'q']")

webElem$sendKeysToElement(list("PYTHON", key = "enter"))

remDr$navigate("http://www.naver.com/")

str(remDr)

webElem <- remDr$findElement(using = "css", "#query")

webElem$sendKeysToElement(list("PYTHON", key = "enter"))

# [ 네이버 웹툰 댓글 읽기 ]

# 베댓과 전댓 url이 안바뀌기 때문에 정적 크롤링 불가

# 페이지 소스에 댓글이 보이지 않음 // 개발자도구에는 보임(렌더링이 끝난 곳에는 보임)

# 서버로 보내진 소스에는 없음 --> Ajax로 개발한 것

url<-'http://comic.naver.com/comment/comment.nhn?titleId=570503&no=135'

remDr$navigate(url)

#단수형으로 노드 추출

more<-remDr$findElement(using='css','#cbox\_module > div > div.u\_cbox\_sort > div.u\_cbox\_sort\_option > div > ul > li:nth-child(2) > a')

more$getElementTagName() # 태그 // 리스트로 리턴

more$getElementText() # 태그 텍스트 내용 // 리스트로 리턴

more$clickElement() # 클릭 이벤트 발생 // 클릭하는 효과

#cbox\_module > div > div.u\_cbox\_sort > div.u\_cbox\_sort\_option > div >

ul > li.u\_cbox\_sort\_option\_wrap.u\_cbox\_sort\_option\_on > a > span.u\_cbox\_sort\_label

# 2페이지부터 10페이지까지 링크 클릭하여 페이지 이동하기

# json으로 구현되었기에 클릭 이벤트를 하나씩 발생시켜야만 함

for (i in 4:12) {

nextCss <- paste0("#cbox\_module>div>div.u\_cbox\_paginate>div> a:nth-child(",i,") > span")

nextPage<-remDr$findElement(using='css',nextCss) # findelement임

nextPage$clickElement() # 클릭이벤트 발생

Sys.sleep(2) # 2초마다 수행

}

#복수형으로 노드 추출

url<-'http://comic.naver.com/comment/comment.nhn?titleId=570503&no=135'

remDr$navigate(url)

#베스트 댓글 내용 읽어오기 --> 베댓만의 공통된 css 선택자를 설정 // findelements임

bestReviewNodes<-remDr$findElements(using ="css selector","ul.u\_cbox\_list span.u\_cbox\_contents")

sapply(bestReviewNodes,function(x){x$getElementText()}) # apply와 같은 함수 // 매트릭스의 가로 세로 기준으로 주어진 함수를 수행해라 ~ // 첫번째로 주어진 객체를 가지고, 두번째 함수를 시행하라

#전체 댓글 링크 클릭후에 첫 페이지 내용 읽어오기

totalReview <- remDr$findElement(using='css','#cbox\_module > div > div.u\_cbox\_sort > div.u\_cbox\_sort\_option > div > ul > li:nth-child(2) > a')

totalReview$clickElement()

totalReviewNodes<-remDr$findElements(using ="css selector","ul.u\_cbox\_list span.u\_cbox\_contents")

sapply(totalReviewNodes,function(x){x$getElementText()})

# 링크 클릭으로 AJAX 로 처리되는 네이버 웹툰 댓글 읽어 오기

repl\_v = NULL;

remDr <- remoteDriver(remoteServerAddr = "localhost" , port = 4445, browserName = "chrome")

remDr$open()

url<-'http://comic.naver.com/comment/comment.nhn?titleId=570503&no=135'

remDr$navigate(url)

#베스트 댓글 내용 읽어오기

doms1<-remDr$findElements(using ="css","ul.u\_cbox\_list span.u\_cbox\_contents")

vest\_repl <- sapply(doms1,function(x){x$getElementText()})

repl\_v <- c(repl\_v, unlist(vest\_repl))

toralReview <- remDr$findElement(using='css','span.u\_cbox\_in\_view\_comment')

toralReview$clickElement()

#전체 댓글의 첫 페이지 내용 읽어오기

doms2<-remDr$findElements(using ="css selector","ul.u\_cbox\_list span.u\_cbox\_contents")

repl <-sapply(doms2,function(x){x$getElementText()})

repl\_v <- c(repl\_v, unlist(repl))

repeat {

for (i in 4:12) {

nextCss <- paste("#cbox\_module>div>div.u\_cbox\_paginate>div> a:nth-child(",i,") > span", sep="")

try(nextListLink<-remDr$findElement(using='css',nextCss))

if(length(nextListLink) == 0) break;

nextListLink$clickElement()

Sys.sleep(1) # 없으면 에러난다. 쉬어가면서 하는 게 더 나음

#전체 댓글의 해당 페이지 내용 읽어오기

doms3<-remDr$findElements(using ="css selector","ul.u\_cbox\_list span.u\_cbox\_contents")

repl <-sapply(doms3,function(x){x$getElementText()})

repl\_v <- c(repl\_v, unlist(repl))

}

try(nextPage<-remDr$findElement(using='css',

"#cbox\_module > div > div.u\_cbox\_paginate > div > a:nth-child(13) > span.u\_cbox\_cnt\_page")) # try함수 --> 에러가 나더라도 넘어감

if(length(nextPage) == 0) break; # 다음 페이지가 없을 때 종료

nextPage$clickElement()

Sys.sleep(1)

doms2<-remDr$findElements(using ="css selector","ul.u\_cbox\_list span.u\_cbox\_contents")

repl <-sapply(doms2,function(x){x$getElementText()})

repl\_v <- c(repl\_v, unlist(repl))

}

print(repl\_v)

write(repl\_v, "webtoon2.txt")

#===============================================================================

# [ 아고다 페이지에 올려진 신라스테이 호텔에 대한 전체 페이지 댓글 읽기 ]

remDr <- remoteDriver(remoteServerAddr = "localhost" , port = 4445, browserName = "chrome")

remDr$open()

url<-'https://www.agoda.com/ko-kr/shilla-stay-yeoksam/hotel/seoul-kr.html?asq=z91SVm7Yvc0eRE%2FTBXmZWCYGcVeTALbG%2FvMXOYFqqcm2JknkW25Du%2BVdjH%2FesXg8ORIaVs1PaEgwePlsVWfwf3sX%2BVNABRcMMOWSvzQ9BxqOPOsvzl8390%2BEhEylPvEiBp0eoREr2xLYHgqmk0Io4J1HYEzEOqyvdox%2BwS6yxHeonB9lh7mJsBIjSBPoMzBLFW01k%2BU8s2bGO6PcSdsu3T30HwabyNzwNYKiv%2BRDxfs%3D&hotel=699258&tick=637215342272&languageId=9&userId=bcb7ecc6-7719-465f-bf29-951e39733c66&sessionId=uouhnqjisace4freagmzbxxc&pageTypeId=7&origin=KR&locale=ko-KR&cid=-1&aid=130243&currencyCode=KRW&htmlLanguage=ko-kr&cultureInfoName=ko-KR&ckuid=bcb7ecc6-7719-465f-bf29-951e39733c66&prid=0&checkIn=2020-05-30&checkOut=2020-05-31&rooms=1&adults=1&childs=0&priceCur=KRW&los=1&textToSearch=%EC%8B%A0%EB%9D%BC%EC%8A%A4%ED%85%8C%EC%9D%B4%20%EC%97%AD%EC%82%BC%20(Shilla%20Stay%20Yeoksam)&productType=-1&travellerType=0&familyMode=off'

remDr$navigate(url)

Sys.sleep(3) # 천천히

pageLink <- NULL

reple <- NULL

curr\_PageOldNum <- 0

repeat{

doms <- remDr$findElements(using = "css selector", ".Review-comment-bodyText")

Sys.sleep(1)

reple\_v <- sapply(doms, function (x) {x$getElementText()})

print(reple\_v) # 페이지 단위 출력

reple <- append(reple, unlist(reple\_v))

cat(length(reple), "\n")

pageLink <- remDr$findElements(using='css',"#reviewSection > div:nth-child(6) > div > span:nth-child(3) > i ")

remDr$executeScript("arguments[0].click();",pageLink) # 자바로 클릭 발생 // 에러가 더 적음 // 두번째에 클릭 이벤트를 발생시킬 웹 엘먼트를 쓴다.

Sys.sleep(2)

curr\_PageElem <- remDr$findElement(using='css','#reviewSection > div:nth-child(6) > div > span.Review-paginator-numbers > span.Review-paginator-number.Review-paginator-number--current')

curr\_PageNewNum <- as.numeric(curr\_PageElem$getElementText())

cat(paste(curr\_PageOldNum, ':', curr\_PageNewNum,'\n')) # 페이지 넘버

if(curr\_PageNewNum == curr\_PageOldNum) { # 다음 페이지가 현재 페이지와 같으면

cat("종료\n")

break; # 다음 페이지 없으면 브레이크

}

curr\_PageOldNum <- curr\_PageNewNum;

}

cat(length(reple), "개의 댓글 추출\n")

write(reple,"hotel.txt")

# [ YES24의 명견만리 댓글 읽어오기 ]

library(RSelenium)

remDr <- remoteDriver(remoteServerAddr = "localhost", port = 4445, browserName = "chrome")

remDr$open()

remDr$navigate("http://www.yes24.com/24/goods/40936880")

webElem <- remDr$findElement("css", "body")

remDr$executeScript("scrollTo(0, 0)", args = list(webElem))

Sys.sleep(1)

remDr$executeScript("scrollBy(0, 3200)", args = list(webElem))

Sys.sleep(1)

remDr$executeScript("scrollBy(0, 3200)", args = list(webElem))

Sys.sleep(1)

remDr$executeScript("scrollBy(0, 3200)", args = list(webElem))

Sys.sleep(3)

repl\_v = NULL

endFlag <- FALSE

page <- 3

repeat {

for(index in 3:7) {

fullContentLinkCSS <- paste("#infoset\_reviewContentList > div:nth-child(",index,") > div.reviewInfoBot.crop > a", sep='')

fullContentLink<-remDr$findElements(using='css selector', fullContentLinkCSS)

if (length(fullContentLink) == 0) {

cat("종료\n")

endFlag <- TRUE

break

}

remDr$executeScript("arguments[0].click();",fullContentLink);

Sys.sleep(1)

fullContentCSS <- paste("#infoset\_reviewContentList > div:nth-child(",index,") > div.reviewInfoBot.origin > div.review\_cont > p", sep='')

fullContent<-remDr$findElements(using='css selector', fullContentCSS)

print(fullContent)

repl <-sapply(fullContent,function(x){x$getElementText()})

print(repl)

repl\_v <- c(repl\_v, unlist(repl))

}

if(endFlag)

break;

if(page == 10){

page <- 3

nextPageCSS <- "#infoset\_reviewContentList > div.review\_sort.sortTop > div.review\_sortLft > div > a.bgYUI.next"

}

else{

page <- page+1;

nextPageCSS <- paste("#infoset\_reviewContentList > div.review\_sort.sortBot > div.review\_sortLft > div > a:nth-child(",page,")",sep="")

}

remDr$executeScript("scrollTo(0, 0)", args = list(webElem))

nextPageLink<-remDr$findElements(using='css selector',nextPageCSS)

remDr$executeScript("arguments[0].click();",nextPageLink);

#sapply(nextPageLink,function(x){x$clickElement()})

Sys.sleep(3)

print(page)

}

write(repl\_v, "yes24.txt")

library(RSelenium)

remDr <- remoteDriver(remoteServerAddr = "localhost", port = 4445, browserName = "chrome")

remDr$open()

remDr$navigate("https://www.istarbucks.co.kr/store/store\_map.do?disp=locale")

selectSeoul <- "#container > div > form > fieldset > div > section > article.find\_store\_cont > article > article:nth-child(4) > div.loca\_step1 > div.loca\_step1\_cont > ul > li:nth-child(1) > a"

nextPageLink<-remDr$findElements(using='css selector',selectSeoul)

sapply(nextPageLink,function(x){x$clickElement()})

selectAll <- paste("#mCSB\_2\_container > ul > li:nth-child(1) > a")

nextPageLink<-remDr$findElements(using='css selector',selectAll)

sapply(nextPageLink,function(x){x$clickElement()})

sizeCss <- "#container > div > form > fieldset > div > section > article.find\_store\_cont > article > article:nth-child(4) > div.loca\_step3 > div.result\_num\_wrap > span"

size <- remDr$findElements(using='css selector',sizeCss)

limit <-sapply(size,function(x){x$getElementText()})

nameList <- NULL

addressList <- NULL

phoneList <- NULL

latList <- NULL

longList <- NULL

for(i in 1 : as.numeric(limit)){

nameCss <- paste("#mCSB\_3\_container > ul > li:nth-child(",i,") > strong", sep = '')

nameLink <- remDr$findElements(using='css selector',nameCss)

a <- sapply(nameLink,function(x){x$getElementText()})

nameList <- c(nameList, unlist(a))

infoCss <- paste("#mCSB\_3\_container > ul > li:nth-child(",i,") > p", sep = '')

infoLink <- remDr$findElements(using='css selector',infoCss)

a <- sapply(infoLink,function(x){x$getElementText()})

words = strsplit(unlist(a), split="\n")

addressList <- c(addressList, words[[1]][1])

phoneList <- c(phoneList, words[[1]][2])

pointCss <- paste("#mCSB\_3\_container > ul > li:nth-child(",i,")", sep = '')

pointLink <- remDr$findElements(using='css selector',pointCss)

latList <- c(latList, sapply(pointLink, function(x){x$getElementAttribute('data-lat')}))

longList <- c(longList, sapply(pointLink, function(x){x$getElementAttribute('data-long')}))

if(i %% 3 == 0 && i != as.numeric(limit)){

remDr$executeScript("var su = arguments[0]; var dom = document.querySelectorAll('#mCSB\_3\_container > ul > li')[su]; dom.scrollIntoView();", list(i))

}

}