20181023a Scraping through lists of hyperlinks

cheungngo

23 October 2018

library(rvest)  
library(tidyverse)

### Drug bank

We are going to search ‘calcium channel blocker’ in the drug bank

home = 'https://www.drugbank.ca/unearth/q?utf8=%E2%9C%93&searcher=targets&query=calcium+channel+blocker'  
home\_html = read\_html(home)

### Accessing to the hyperlinks to individual drug groups

home\_nodes = html\_nodes(home\_html,'.hit-link a')  
home\_text = html\_text(home\_nodes)  
home\_href = html\_attr(home\_nodes,'href')  
home\_href[1:4]

## [1] "/biodb/bio\_entities/BE0004422" "/biodb/bio\_entities/BE0000030"  
## [3] "/biodb/bio\_entities/BE0004906" "/biodb/bio\_entities/BE0003564"

# We can see the href of the links is missing something  
home\_href = str\_c('https://www.drugbank.ca',home\_href)  
home\_href[1:4]

## [1] "https://www.drugbank.ca/biodb/bio\_entities/BE0004422"  
## [2] "https://www.drugbank.ca/biodb/bio\_entities/BE0000030"  
## [3] "https://www.drugbank.ca/biodb/bio\_entities/BE0004906"  
## [4] "https://www.drugbank.ca/biodb/bio\_entities/BE0003564"

### Selecting the useful hyperlinks

ind = grep('calcium channel',home\_text)  
home\_text[ind]

## [1] "Voltage-gated calcium channel beta 1 subunit splice variant CavB1d"  
## [2] "Voltage-dependent N-type calcium channel"   
## [3] "Voltage-dependent calcium channel subunit alpha-2/delta-2"   
## [4] "Voltage-dependent L-type calcium channel subunit beta-1"   
## [5] "Voltage-dependent calcium channel subunit alpha-2/delta-1"   
## [6] "Voltage-dependent calcium channel subunit alpha-2/delta-3"   
## [7] "Voltage-dependent L-type calcium channel subunit beta-4"   
## [8] "Voltage-dependent L-type calcium channel subunit beta-3"   
## [9] "Voltage-dependent L-type calcium channel subunit beta-2"   
## [10] "Voltage-dependent calcium channel gamma-1 subunit"   
## [11] "Voltage-dependent N-type calcium channel subunit alpha-1B"   
## [12] "Voltage-dependent P/Q-type calcium channel"   
## [13] "Voltage dependent L type calcium channel"

use\_href = home\_href[ind]

### Scraping the information from each hyperlink

get\_text = function(url,css) {  
 url %>%  
 read\_html() %>%  
 html\_nodes(css) %>%  
 html\_text()  
}

get\_href = function(url,css) {  
 url %>%  
 read\_html() %>%  
 html\_nodes(css) %>%  
 html\_attr('href')  
}

get\_all = function(url) {  
 html = read\_html(url)  
 name = get\_text(url, 'td:nth-child(2) strong')  
 DBID = get\_text(url, 'td:nth-child(1) a')  
 status = get\_text(url, '#target-relations td:nth-child(3)')  
 href = get\_href(url, 'td:nth-child(1) a')  
 combined = data.frame(Name = name,  
 DBID = DBID,  
 Status = status,  
 Link = href)  
 return(combined)  
}

CCB = use\_href %>%  
 map(get\_all) %>%  
 bind\_rows()

write\_csv(CCB, 'CCB.csv')

### Reading the resulting table

library(readr)  
CCB <- read\_csv("20181022a-web-scraping-(rvest)/CCB.csv")

We can see much duplications here

dim(CCB)

## [1] 84 4

CCB = unique(CCB)  
dim(CCB)

## [1] 39 4

Also to complete the url

CCB[,4] = sapply(CCB[,4],function(i) {  
 str\_c('https://www.drugbank.ca',i)  
})

write\_csv(CCB,'CCB02.csv')

### Accessing into the details of each drug

get\_detail = function(url) {  
 html = read\_html(url)  
 mech = get\_text(url,'.col-sm-8:nth-child(6) p')  
 data.frame(Pharmacodynamics = mech)  
}

drug\_href = as.vector(sapply(CCB[,4], as.character))  
CCB\_details = drug\_href %>%  
 map(get\_detail) %>%  
 bind\_rows()

CCB = bind\_cols(CCB,CCB\_details)

write\_csv(CCB, 'CCB03.csv')