My first Web_Scraping with R

SeongJin Kim

2019-04-24 수요일

```
##  ------------------------------------------------------------------------------
##  2019-04-24   KOSPI
##  ------------------------------------------------------------------------------
```

```
#          .
library(tidyverse)
library(httr)
library(rvest)
library(readr)
```

```
#      KOSPI           .
# https://finance.naver.com/sise/sise_index.nhn?code=KOSPI

# HTTP       .
res <- GET(url = 'https://finance.naver.com/sise/sise_index.nhn?code=KOSPI')
```

```
# html
status_code(x = res) ##[1] 200 :
```

```
## [1] 200
```

```
readr::guess_encoding(file="C:/Users/iihsk/Desktop/ds_web scraping/Sisa web scraping/   _     .html")
```

```
## # A tibble: 5 x 2
##   encoding   confidence
##   <chr>           <dbl>
## 1 EUC-KR          1
## 2 GB18030         0.62
## 3 EUC-JP          0.42
## 4 Big5            0.35
## 5 ISO-8859-1      0.290
```

```
#cat(content(x = res, as = 'text', encoding = 'EUC-KR'))

#         .
#print(x = res)
```

```
#               locale
Sys.setlocale(category = 'LC_ALL', locale = 'C')
```

```
## [1] "C"
```

```
#           .
tableContents <- res %>%
  read_html(encoding='EUC-KR') %>%
  html_nodes(css = '#contentarea_left > div.box_top_sub > div > div.subtop_sise_detail > table') %>%
  html_table(trim = TRUE)

#
# print(x = tableContents)
# glimpse(x = tableContents)
```

```
dim(tableContents) # 4      4
```

```
## NULL
```

```
#
Sys.setlocale(category = 'LC_ALL', locale = 'korean')
```

```
## [1] "LC_COLLATE=Korean_Korea.949;LC_CTYPE=Korean_Korea.949;LC_MONETARY=Korean_Korea.949;LC_NUMERIC=C
```

```
#
data <- tableContents[[1]] # list      data frame
glimpse(data)
```

```
## Observations: 4
## Variables: 4
## $ X1 <chr> "  ( )", "   ", "52  ", " / "
## $ X2 <chr> "564,866", "2,229.75", "2,516.57", "   3 \n\t            ...
## $ X3 <chr> "   ( )", "   ", "52  ", "   3 \n\t                  ...
## $ X4 <chr> "6,002,015", "2,190.29", "1,984.53", "   3 \n\t          ...
```

```
# 1   : 1~2   3~4          . tidy data          2
part1 <- data[,1:2];part2 <- data[,3:4]
colnames(part1) <- c(" "," ");colnames(part2) <- c(" "," ") # rbind

data <- rbind(part1, part2) #
data <- data[-dim(data)[1],] #

#
#    : data[," "] <- data$  %>% str_remove_all(pattern = '\n\t| ')
print(data[4,2]) # \n\t
```

```
## [1] "   3 \n\t                              251 \n\t                              58 \n\t
```

```
x <- data[4,2]  %>% str_split(pattern = '\n\t ') # pattern
print(x)
```

```
## [[1]]
## [1] "   3 "
## [2] "                              251 "
## [3] "                              58 "
## [4] "                              584 "
## [5] "                              0"
```

data[4,2]에 저장돼 있는 지저분한 꼴을 4개 항목으로 따로 저장하려는 작업입니다.

```
#                separate.data
separate.data <- c()
for(i in 1:length(x[[1]])){
  separate.data[i] <- x[[1]][i] %>% str_trim()
}

#     5   ' '     content
content <- c()
for(i in 1:5){
  content[i] <- separate.data[i] %>% str_sub(start=1, end=5)
}

# ' '          ' '     value
```

```
value <- c()
for(i in 1:5){
  value[i] <- separate.data[i] %>% str_sub(start=6, end=nchar(separate.data[i]))
}

# ' ' ''
partial.data <- cbind(content, value)
print(partial.data)
```

```
##      content      value
## [1,] "   " "3"
## [2,] "   " "251"
## [3,] "   " "58"
## [4,] "   " "584"
## [5,] "   " "0"
```

```
#
data <- data[-4,]
colnames(partial.data) <- colnames(data)
data <- rbind(data, partial.data)
print(data)
```

```
##
## 1      ( )   564,866
## 2             2,229.75
## 3        52    2,516.57
## 5    ( ) 6,002,015
## 6             2,190.29
## 7        52    1,984.53
## 11                3
## 21              251
## 31               58
## 4               584
## 51                0
```

```
#
write.csv(data, "20190423 KOSPI.csv", row.names=FALSE)
```