제11장 웹 스크래핑의 활용

I. 특정 웹 페이지의 상품 정보 스크래핑 : 단일 웹 페이지

1. 웹 사이트 접속

```
웹사이트의 화면구성 변경에 따른 URL 수정
# url <- "http://www.coupang.com/np/search?
q=%EC%97%AC%EC%84%B1%ED%81%AC%EB%A1%9C%EC%8A%A4%EB%B0%B1"
#-------------------------
# 관심있는 웹사이트의 url 할당
#-------------------------
# url <- "http://www.coupang.com/np/categories/130322"
```

2. 웹 문서 추출

```
doc <- htmlParse(url, encoding="UTF-8")
doc
```

3. 검색한 상품명 출력

```
# 상품명 추출 : id가 'productList'인 ul 태그 내의 class명이 'name'인 div 태그의 데이터 추출
#             태그 경로를 기술할 때 중간 단계를 건너뛰어 표현할 때 '//'를 사용. '[]'내에는 id
또는 class 등의 태그 구분
#-------------------------
prod_name <- xpathSApply(doc, "//ul[@id='productList']//div[@class='name']", xmlValue)

prod_name    # 추출한 상품명 출력
```

4. 불필요한 단어 제거

```
prod_name <- gsub('\n', '', prod_name)
prod_name <- gsub(' ', '', prod_name)
prod_name
```

5. 상품명에 해당하는 상품가격 추출 :

=> id가 'productList'인 ul 태그 내의 class명이 'price-value'인 strong 태그의 데이터 추출

```
price     <- xpathSApply(doc, "//ul[@id='productList']//strong[@class='price-value']",
xmlValue)
price       # 추출한 상품가격 출력
```

6. 데이터프레임 만들기 :

추출한 prod_name과 price를 상품명의 가격의 항목으로 데이터 프레임 만들기

```
df <- data.frame(상품명=prod_name, 가격=price)
df

df$상품명 <- format(df$상품명, justify = "left")
df
```

II. 다수 웹 페이지의 상품 정보 스크래핑

```
# 웹 스크래핑 URL
# library(XML)
# 웹사이트의 화면구성 변경에 따른 URL 수정
# url <- "http://www.coupang.com/np/search?
q=%EC%97%AC%EC%84%B1%ED%81%AC%EB%A1%9C%EC%8A%A4%EB%B0%B1&channel=user&component=&event
Category=SRP&sorter=&minPrice=&maxPrice=&priceRange=&filterType=&listSize=36&filter=&f
ilterKey=&isPriceRange=false&brand=&rating=0&page="
url <- "http://www.coupang.com/np/categories/130322?page="

df.products <- NULL

for (page in 1:10) {
    url2 <- paste(url, page, sep="")
    doc <- htmlParse(url2, encoding="UTF-8")
    prod_name <- xpathSApply(doc, "//ul[@id='productList']//dd[@class='name']",
xmlValue)
    prod_name <- gsub('\n', '', prod_name)
    prod_name <- gsub(' ', '', prod_name)
    price     <- xpathSApply(doc, "//ul[@id='productList']//strong[@class='price-
value']", xmlValue)
    df <- data.frame(상품명= prod_name, 가격=price)
    df.products <- rbind(df.products, df)
}
df.products

df.products$상품명 <- format(df.products$상품명, justify = "left")
df.products
```



PDF