

[빅데이터연구소] 사회혁신특강: 웹 크롤링

□ 특강내용

- 일시 및 장소: 2019년 11월 26일(화), 15:00-17:30 / S1437
- 강연 제목: R을 이용한 데이터 크롤링
- 강사: 박찬엽(SKT)
- 참가 인원: 황낙주 외 30명

1) 사용한 R패키지 및 함수

- rvest, httr, dplyr
- read_html, read_nodes, html_text, html_attr

2) Background

- 서버 : 외부에서 요청하면 규칙대로 정보를 제공하는 것
- 브라우저 : 서버가 주는 것들을 사용자에게 보여주는 것
- 웹 서버는 우리에게 text(html, css, js, etc), images, 브라우저가 약속한 모양으로 우리에게 보여줌
- web api : web으로 제공되는 application programming interface
함수인데 외부 서버에 동작하여 웹 기술로 결과를 받는 것
- chrome 이용시 마우스 우측버튼 - 검사탭 선택(F12)하여 html문서 볼 수 있음
- html : xml약식으로 작성된 웹 브라우저들이 이해하는 표준 문서
- node : html에서 tag라고 불리는 것, 하나의 카테고리
- attr : attribute의 줄임말, 노드 안에 있는 원소들

3) Process

1. html 문서 데이터 가져오기
2. 필요한 노드 선택하기
3. 노드 내에서 text를 가져오기

4) code book

- 기사제목 불러오기

```
tar <- "https://news.naver.com/main/read.nhn?mode=LSD&mid=shm&sid1=102&oid=011&aid=0003657344"
```

```
read_html(tar) %>%  
  html_nodes("h3#articleTitle") %>%  
  html_text()
```

```
## [1] "윗선 향하는 '유재수 감찰 무마' 수사망...조국 등 소환할 듯"
```

- 중간글씨 불러오기

```
read_html(tar) %>%  
  html_nodes("strong.media_end_summary") %>%  
  html_text()
```

```
## [1] "이인걸·박형철 참고인으로 조사조만간 조국·백원우 소환할 듯法, 27일 유재수 영장실질심사"
```

- 기사list 불러오기

```
tar <- "https://news.naver.com/main/list.nhn?mode=LS2D&mid=shm&sid1=105&sid2=229"
```

```
read_html(tar) %>%  
  html_nodes("dt a") %>%  
  html_attr("href") %>%  
  unique() -> linkList  
linkList
```

```
## [1] "https://news.naver.com/main/read.nhn?mode=LS2D&mid=shm&sid1=105&sid2=229&oid=347&aid=0000137741"  
## [2] "https://news.naver.com/main/read.nhn?mode=LS2D&mid=shm&sid1=105&sid2=229&oid=347&aid=0000137740"
```



사진1



사진2