

{ Data Crawling }

(Kyungnam Univ. Bigdata center Kim JiSeong)

1) Data Crawling

사용 이유 ? SNS Data, NEWS Data 를 개인 데이터 수집 (..?)
항상 내가 필요한 데이터는 항상 잘 없다

창원시 화재 빈도

전체(0) 파일데이터(0) 오픈API(0) 표준데이터(0)

전체 0건을 찾았습니다.

| 📁 파일데이터 [0건] | 정확도 ▼ | 날 | 짜 | 제 | 목 | 조회수 | 다운로드 |
|---------------|-------|---|---|---|---|-----|------|
| 🔍 검색결과가 없습니다. | | | | | | | |

| ⚙️ 오픈API [0건] | 정확도 ▼ | 날 | 짜 | 제 | 목 | 조회수 | 활용신청 |
|---------------|-------|---|---|---|---|-----|------|
| 🔍 검색결과가 없습니다. | | | | | | | |

(Data Crawling)

2) How to Crawling ?

HTML node가 ..., XML 이 어쩌고 ...?? @_@
좀더 쉽고 강력한 방법 ?! **Selenium**

[Selenium 사용을 위한 준비 (1)]

https://github.com/KimJiSeong1994/crawling_lecture

(Data Crawling)

2) How to Crawling ?

HTML node가 ..., XML 이 어찌고 ...?? @_@
좀더 쉽고 강력한 방법 ?! **Selenium**

[Selenium 사용을 위한 준비 (2)]

1) cmd(명령 프롬프트) 열기

2) cd C:\W\selenium

3) java -Dwebdriver.gec.ko.driver="geckodriver.exe"
- jar selenium-server-standalone-3.141.59.jar
- port 4445

(Data Crawling)

2-1) Selenium

#----- Step 1) chrome open -----

```
install.packages(c("RSelenium"))  
library(Rselenium)
```

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

```
remDr$open() # chrome driver 열기
```

(Data Crawling)

2-1) Selenium

#----- Step 2) navigate page -----

```
remDr$open()
```

```
remDr$navigate( url 경로 ) # url 경로에 원하는 사이트 주소 입력
```

#----- Step 3) get blog url -----

```
remDr$findElements(using = "class", value = "f_url") %>%
```

```
map(function(x) x$getText()) %>%
```

```
unlist() -> url_daum # blog 주소를 받아 접속하기 위해
```

(Data Crawling)

2-1) Selenium

#----- Step 4) get blog text -----

```
text_df <- data.frame(NULL) # 빈 데이터프레임 만들기
for (k in 1:length(url_daum)) {
  if(!str_detect(url_daum[k], "blog.naver.com")) {
    remDr$navigate(paste0("https://", url_daum[k]))
    remDr$findElements(using = "css", value = "P") %>%
      map(function(x) x$getText()) %>%
      unlist() %>%
      .[str_detect(., www)] -> text_vec # 문자 이외의 행 제거

    text_vec <- data.frame(text = text_vec) # text_vec reset
    text_df <- rbind(text_df, text_vec)
  }
}
```

(Data Crawling)

2-1) Selenium

#----- Step 5) total loop -----

(Data Crawling)