R 데이터(텍스트) 마이닝 및 시각화 분석

Seongmin Mun

20th January 2022

아주대학교 AJOU UNIVERSITY

Outline

웹사이트 스크래핑/크롤링

웹사이트 스크래핑/크롤링

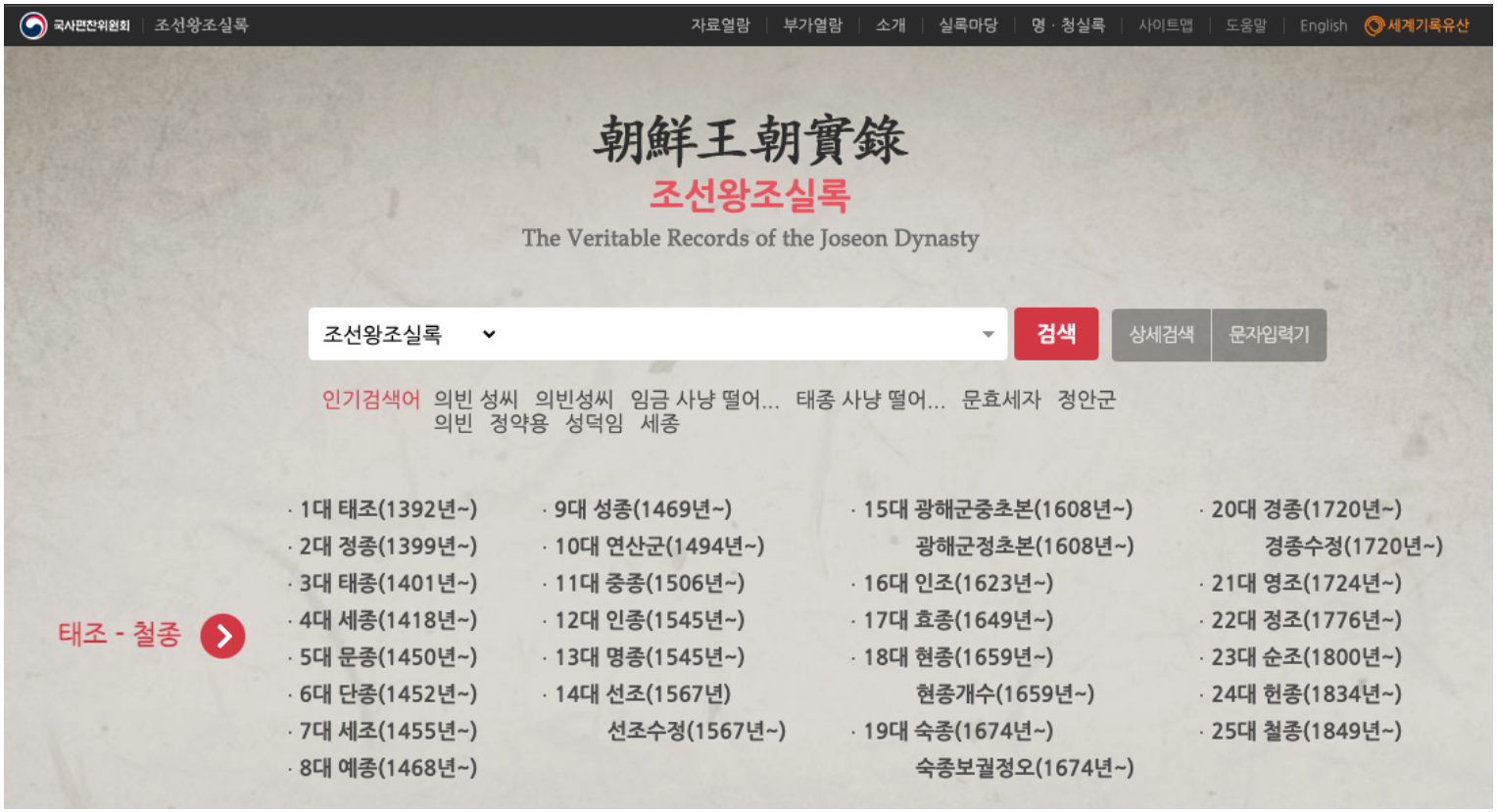
생각하기

- 크롤링은 불법인가요?
- 사례1 (이득을 얻는 행위)
- 사례2 (과도한 트랙픽 접속)
- 네이버 약관 및 개인정보 보호

생각하기

- 크롤링한 데이터 타인에게 제공하여 이득을 얻는 행위
- 크롤링으로 서버에 무리한 트래픽 요청을 하여 해당 사이트의 서비스 이용을 방해하는 행위

조선왕조실록



http://sillok.history.go.kr/main/main.do

초기 설정

```
> #memory & previous_works
> gc()
          used (Mb) gc trigger  (Mb) limit (Mb) max used  (Mb)
Ncells 1850769 98.9    3587530 191.6         NA  3587530 191.6
Vcells 4109824 31.4    8388608  64.0      16384  8388608  64.0
> rm(list=ls())
> #Encoding_mac
> Sys.setlocale(category = "LC_CTYPE", locale = "ko_KR.UTF-8")
[1] "ko_KR.UTF-8"
> #Window운영체제
> #options(Encoding = "UTF-8")
```

RSelenium 설치

```
> install.packages("RSelenium")
URL 'https://cran.rstudio.com/bin/macosx/contrib/4.1/RSelenium_1.7.7.tgz'을 시도합니다
Content type 'application/x-gzip' length 3345778 bytes (3.2 MB)
==================================================
downloaded 3.2 MB


The downloaded binary packages are in
        /var/folders/f9/tcyhkyhx32q6nxcypnp1rl2c0000gn/T//RtmpdxHA1N/downloaded_packages
> # install.packages("RSelenium")
> library(RSelenium)
```

가상 드라이브 사용

```
> pJS <- wdman::phantomjs(port = 4567L)
checking phantomjs versions:
BEGIN: PREDOWNLOAD
BEGIN: DOWNLOAD
BEGIN: POSTDOWNLOAD
> remDr <- remoteDriver(remoteServerAddr = 'localhost',
+                       port =4567L, # 포트번호 입력
+                       browserName = "chrome")
> remDr$open()
[1] "Connecting to remote server"
$browserName
[1] "phantomjs"

$version
[1] "2.1.1"

$driverName
[1] "ghostdriver"

$driverVersion
[1] "1.2.0"
```

조선왕조실록접속-1

```
> #메인 페이지로 접속하기
> remDr$navigate("http://sillok.history.go.kr/main/main.do")
> remDr$screenshot(display = T)
> Sys.sleep(1)
>
>
```

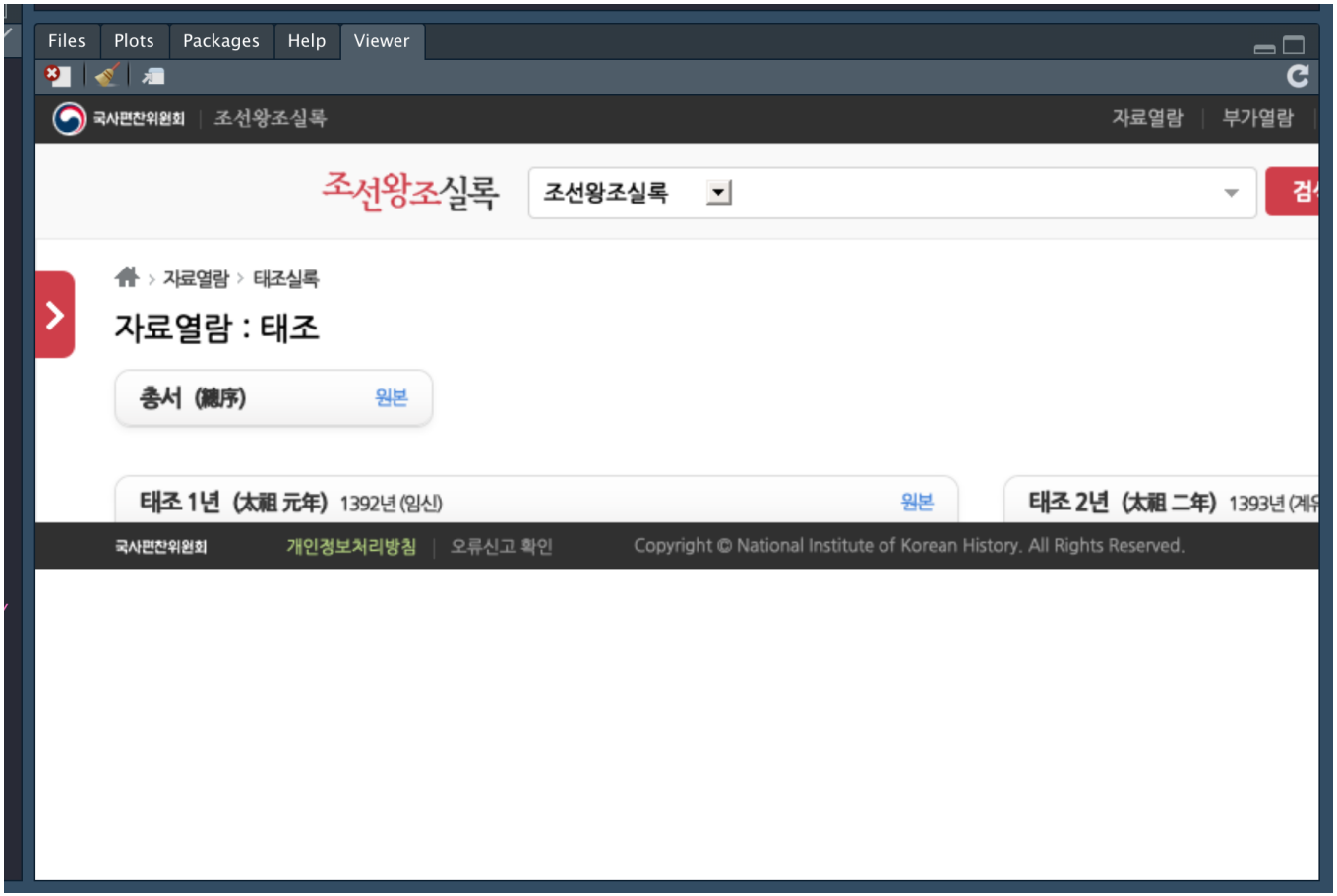
조선왕조실록접속-2


Files
Plots
Packages
Help
Viewer
국사편찬위원회 | 조선왕조실록
자료열람 | 부가열람 | 소개 | 실록마당
朝鮮王朝實錄
조선왕조실록
The Veritable Records of the Joseon Dynasty
조선왕조실록
인기검색어 의빈 성씨 의빈성씨 임금 사냥 떨어... 태종 사냥 떨어... 문효세자
의빈 정약용 성덕임 세종
· 1대 태조(1392년~)
· 2대 정종(1399년~)
· 3대 태종(1401년~)
· 4대 세종(1418년~)
· 9대 성종(1469년~)
· 10대 연산군(1494년~)
· 11대 중종(1506년~)
· 12대 인종(1545년~)
· 15대 광해군중초본
광해군정초본
· 16대 인조(1623년
· 17대 효종(1649년
태조 - 철종

태조로 접근-1

```
> #태조에 대한 텍스트 클릭하기
> webElem <- remDr$findElement(using = "xpath", value='//*[@id="m_cont_list"]/div[1]/ul[1]/li[1]/
a')
> webElem$clickElement()
> remDr$screenshot(display = T)
> Sys.sleep(1)
>
>
```

태조로 접근-2


Files
Plots
Packages
Help
Viewer
국사편찬위원회
조선왕조실록
자료열람
부가열람
조선왕조실록
조선왕조실록
자료열람 > 태조실록
자료열람 : 태조
총서 (總序)
원본
태조 1년 (太祖 元年) 1392년 (임신)
원본
태조 2년 (太祖 二年) 1393년
국사편찬위원회
개인정보처리방침
오류신고 확인
Copyright © National Institute of Korean History. All Rights Reserved.

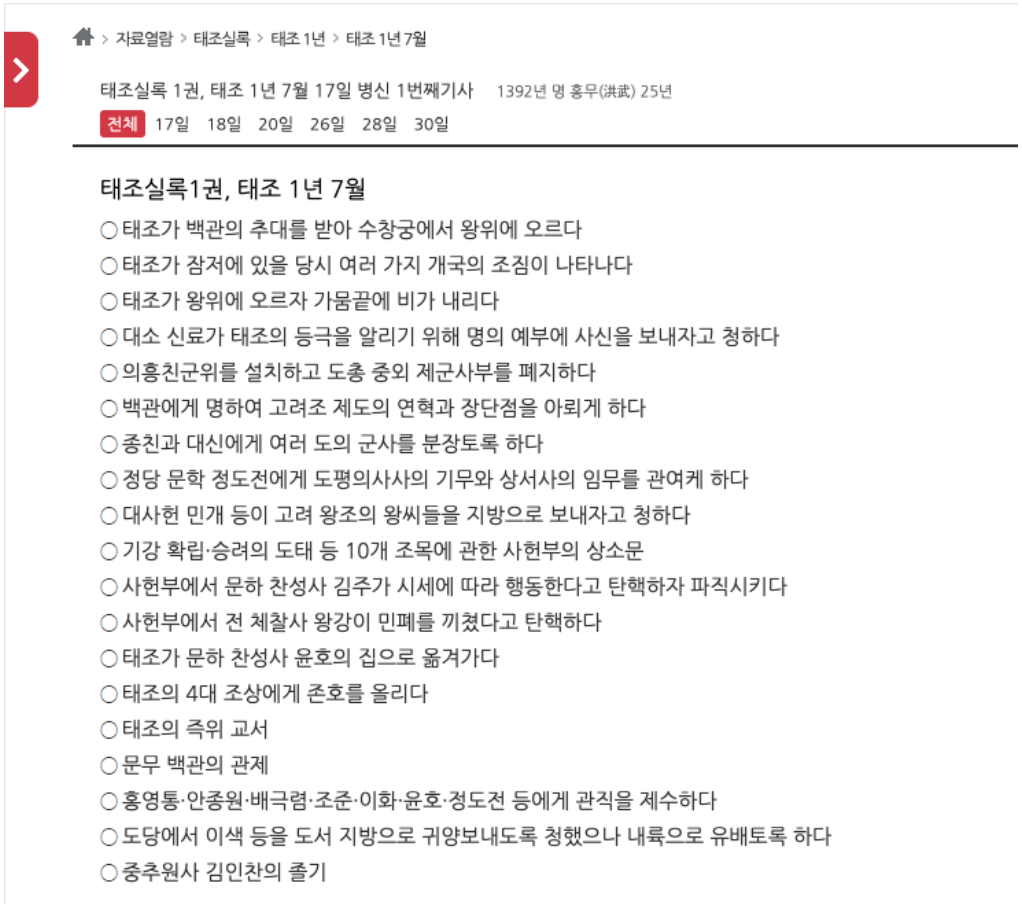
태조 1년 7월에 접근-1

```
> webElem <- remDr$findElement(using = "xpath", value='//*[@id="cont_area"]/div/div[2]/ul[2]/li
[1]/ul/li[1]/a')
> webElem$clickElement()
> remDr$screenshot(display = T)
> Sys.sleep(1)
>
>
```

태조 1년 7월에 접근-2


조선왕
자료열람 > 태조실록
자료열람 : 태조
총서 (總序)
원본
태조 1년 (太祖 元年) 1392년 (임신)
원본
7월 8월 9월 10월 11월 12월 윤12월
태조 3년 (太祖 三年) 1394년 (갑술)
원본
1월 2월 3월 4월 5월 6월 7월 8월 9월 10월 11월 12월

태조 1년 7월의 문서 목록 가져오기-1



태조 1년 7월의 문서 목록 가져오기-2

```
> # 태조 1년 7월의 문서
> #페이지 소스 읽어오기
> html <- remDr$getPageSource()[[1]]
> html <- read_html(html)
> #관련정보들 가져오기
> text_list <- html %>% html_nodes(".ins_list_main dd ul li a") %>% html_text()
> text_list
 [1] "태조가 백관의 추대를 받아 수창궁에서 왕위에 오르다"
 [2] "태조가 잠저에 있을 당시 여러 가지 개국의 조짐이 나타나다"
 [3] "태조가 왕위에 오르자 가뭄끝에 비가 내리다"
 [4] "대소 신료가 태조의 등극을 알리기 위해 명의 예부에 사신을 보내자고 청하다"
 [5] "의흥친군위를 설치하고 도총 중외 제군사부를 폐지하다"
 [6] "백관에게 명하여 고려조 제도의 연혁과 장단점을 아뢰게 하다"
 [7] "종친과 대신에게 여러 도의 군사를 분장토록 하다"
 [8] "정당 문학 정도전에게 도평의사사의 기무와 상서사의 임무를 관여케 하다"
 [9] "대사헌 민개 등이 고려 왕조의 왕씨들을 지방으로 보내자고 청하다"
[10] "기강 확립·승려의 도태 등 10개 조목에 관한 사헌부의 상소문"
[11] "사헌부에서 문하 찬성사 김주가 시세에 따라 행동한다고 탄핵하자 파직시키다"
[12] "사헌부에서 전 체찰사 왕강이 민폐를 끼쳤다고 탄핵하다"
[13] "태조가 문하 찬성사 윤호의 집으로 옮겨가다"
[14] "태조의 4대 조상에게 존호를 올리다"
[15] "태조의 즉위 교서"
[16] "문무 백관의 관제"
[17] "홍영통·안종원·배극렴·조준·이화·윤호·정도전 등에게 관직을 제수하다"
[18] "도당에서 이색 등을 도서 지방으로 귀양보내도록 청했으나 내륙으로 유배토록 하다"
[19] "중추원사 김인찬의 졸기"
```