Jsoup를 이용한 조선왕조실록의 빅 데이터 분석

변영일* · 이충호

한밭대학교

Big Data Analysis of the Annals of the Joseon Dynasty Using Jsoup

Young-Il Bong* · Choong-Ho Lee

Hanbat National University · Hanbat National University

E-mail : qusduddlf2@naver.com / chlee@hanbat.ac.kr

요 약

조선왕조실록은 UNESCO에 등재된 중요한 기록물이다. 본 논문은 한글로 번역된 조선왕조 실록에서 단어의 빈도수를 조사하여 빅데이터를 분석하는 방법을 제안한다. 조선왕조 실록을 인터넷 사이트에서 액세스하여 단어의 빈도수를 조사하려 할 때, 그 페이지에 포함된 소스를 직접 액세스하면 HTML 문법에 필요한 키워드가 포함되어 있어 필요한 본문에서 단어 빈도수에 의한 빅데이터 분석을 하는 것이 어렵다. 본 논문에서는 Java의 Jsoup를 활용한 크롤링 기능을 사용하여 조선왕조 실록의 본문을 분석하는 방법을 제안한다. 실험에서는 조선왕조실록의 태조부분만을 추출하여 본 방법의 유효성을 검증하였다.

ABSTRACT

The Annals of the Joseon Dynasty are important records registered in UNESCO. This paper proposes a method to analyze big data by examining the frequency of words in the Annals of the Joseon Dynasty translated into Korean. When you access the Annals of the Joseon Dynasty from an Internet site and try to investigate the frequency of words, if you directly access the source included in the page, the keywords necessary for the HTML grammar are included, so that it is difficult to analyze big data based on the frequency of words in the necessary text. In this paper, we propose a method to analyze the text of the Annals of the Joseon Dynasty using Java's Jsoup crawling function. In the experiment, only the Taejo part of the Annals of the Joseon Dynasty was extracted to verify the validity of this method.

키워드

The Annals of the Joseon Dynasty, Big Data Analysis, Frequency of Words, Jsoup, Web Crawling

Ⅰ. 서 론

최근 웹 기반의 데이터를 활용하기 위하여 정보를 수집 및 검색하는 일에 시간과 비용을 과투자하는 일이 빈번히 발생하고 있다. 이러한 웹 기반의 데이터를 자동으로 수집 및 검색하는 방식으로 웹 크롤러가 있다. 웹 크롤러(web crawler)는 조직적, 자동화된 방법으로 월드 와이드 웹을 탐색하는 컴퓨터 프로그램을 지칭하는 것으로, 웹 크롤러가 하는 작업을 웹 크롤링(web crawling)이라고 한다. 웹 크롤러는 특정한 웹 페이지에서 시작해 다른 페이지로 연결된 하이퍼링크를 따라가고, 또 거기에서 다른 페이지를 연쇄적으로 따라가는 형식으로 작동한다.

본 연구에서는 JSoup을 활용한 웹 크롤링을 하여 조선왕조실록 데이터를 연대순으로 순차적으로 수집하고 수집된 데이터를 활용하여 분석하는 방법을 제안하고자 한다.

Ⅱ. 조선왕조실록 웹 크롤러

* speaker

본 논문에서는 조선왕조실록 웹 크롤링을 위하여 Jsoup을 활용하여 조선왕조실록 사이트에서 html를 분석하여 얻은 데이터를 바탕으로 구성된 하이퍼 텍스트들을 큐에 추가하도록 한다. 다음의 그림 1과 그림 2의 내용은 각각 해당 소스 코드와 결과물의 일부를 보여준다[1].

```
public void GetRealAddress()
{
    Deque<String> JScriptQueue = new ArrayDeque<String>();
    String wstr = "";
    while(!DayAddressQueue.isEmpty())
        {
        try {
            Document doc = Jsoup.connect(DayAddressQueue.pop()).get();

            if(doc == null)
                continue;

            Element getdiv = doc.getElementById("cont_area");
            Elements data = getdiv.select("li");
            Elements getherfdata = data.select("a");
            for (int i = 0; i < getherfdata.size(); i++)
            {
                JScriptQueue.add(getherfdata.get(i).attr("href"));
            }
            JScriptQueue.removeIf(str -> !(str.contains("javascript:searchView")));

            while(!JScriptQueue.isEmpty())
            {
                String[] jscriptsplit = JScriptQueue.pop().split("\'");
                wstr = "http://sillok.history.go.kr/id/"+ jscriptsplit[1] + "\n";
                RealAddressQueue.add(wstr);
                srcoutput.write(wstr.getBytes());
            }

        } catch (IOException e) {
            e.printStackTrace();
        }
    }
}
```

그림 1. 하이퍼링크 생성 코드

```
http://sillok.history.go.kr/id/kaa_10107017_001
http://sillok.history.go.kr/id/kaa_10107017_002
http://sillok.history.go.kr/id/kaa_10107018_001
http://sillok.history.go.kr/id/kaa_10107018_002
http://sillok.history.go.kr/id/kaa_10107018_003
http://sillok.history.go.kr/id/kaa_10107018_004
http://sillok.history.go.kr/id/kaa_10107018_005
http://sillok.history.go.kr/id/kaa_10107020_001
http://sillok.history.go.kr/id/kaa_10107020_002
http://sillok.history.go.kr/id/kaa_10107020_003
http://sillok.history.go.kr/id/kaa_10107026_001
http://sillok.history.go.kr/id/kaa_10107026_002
http://sillok.history.go.kr/id/kaa_10107028_001
http://sillok.history.go.kr/id/kaa_10107028_002
http://sillok.history.go.kr/id/kaa_10107028_003
http://sillok.history.go.kr/id/kaa_10107028_004
http://sillok.history.go.kr/id/kaa_10107028_005
http://sillok.history.go.kr/id/kaa_10107030_001
http://sillok.history.go.kr/id/kaa_10107030_002
http://sillok.history.go.kr/id/kaa_10108001_001
http://sillok.history.go.kr/id/kaa_10108002_001
http://sillok.history.go.kr/id/kaa_10108002_002
http://sillok.history.go.kr/id/kaa_10108001_001
http://sillok.history.go.kr/id/kaa_10108002_001
http://sillok.history.go.kr/id/kaa_10108002_002
http://sillok.history.go.kr/id/kaa_10108002_003
http://sillok.history.go.kr/id/kaa_10108002_004
http://sillok.history.go.kr/id/kaa_10108002_005
http://sillok.history.go.kr/id/kaa_10108005_001
http://sillok.history.go.kr/id/kaa_10108007_001
http://sillok.history.go.kr/id/kaa_10108007_002
http://sillok.history.go.kr/id/kaa_10108007_003
http://sillok.history.go.kr/id/kaa_10108007_004
```

그림 2. 하이퍼링크 생성 결과

이전에 구성한큐 를 하나씩 pop하여 구성한 하이퍼 텍스트를 하나씩 순회하면서 ins_view ins_view_left w50_w50 Element 내용을 크롤링하여 국문으로된 조선왕조실록 내용만을 텍스트 파일로 저장하도록 설계하였다[2]. 크롤러는 약 2시간동안 실행한 결과 조선왕조실록의 모든 내용을 크롤링 할 수 있었다. 다음의 그림 3와 그림 4의 내용은 각각 해당 소스 코드와 결과물의 일부를 보여준다.

```
public void GetFromUrl()
{
    String url = "";

    while(!RealAddressQueue.isEmpty())
    {
        try {
            url = RealAddressQueue.pop();
            Document doc = Jsoup.connect(url).get();

            if(doc == null)
                continue;
            //ins_view ins_view_left w50_w50
            //Elements tagVal = doc.select("div p");
            Element lefttext = doc.getElementsByClass("ins_view ins_view_left w50_w50").first();
            String str = lefttext.text();

            MakeFreqData(str);              // 빈도수 계산

            str = str + "\n";
            txtoutput.write(str.getBytes());

            System.out.println(str);
        } catch (IOException e) {
            e.printStackTrace();
        }
    }
}
```

그림 3. 크롤링 코드

태조가 수창궁(壽昌宮)에서 왕위에 올랐다. 이보다 먼저 이달 12일에 공양왕(恭讓王)이 장차 태조의 사제(賜
僎霖)·한상경(韓尙敬)·황거정(黃居正)·임연중(任彦忠)·장사정(張思鋌)·민여익(閔汝翼) 등 대소신료(大小臣僚)
여 왕위를 사양하고 물러났습니다. 이에 이색(李穡)·조민수(曹敏修) 등이 신우(辛禑)의 처부(妻父)인 이임(李
사사의 은혜를 베풀어 인망(人望)을 수습하고자 하여, 다만 법을 범한 사람이 있으면 반드시 모두 용서해
)를 받았다. 육조(六曹)의 판서(判書) 이상의 관원에게 명하여 전상(殿上)에 오르게 하고는 이르기를, "내가
임금이 장저(潛邸)에 있을 때, 꿈에 신인(神人)이 금자(金尺)를 가지고 하늘에서 내려와 주면서 말하기를, "
장차 손안[掌握]에 돌아갈 것이니 전쟁의 괴로움을 꺼리지 말고 능히 나라를 지키는 공을 이루게 하시오."
비가 내리었다. 이보다 앞서 오랫동안 가물었는데, 임금이 왕위에 오르자 억수같이 비가 내리니, 백성의 마
도평의사사(都評議使司) 및 대소 신료(大小臣僚)와 한량(閑良)·기로(耆老) 등이 지밀직사사(知密直司事) 조
하는 요(搖)가 그래도 허물을 고치지 아니하고 또 살육(殺戮)할 것을 꾀하므로, 온 나라 신민(臣民)들이 실
의흥친군위(義興親軍衛)를 설치하고 도총 중외 제군사부(都摠中外諸軍事府)를 폐지하였다. 【태백산사고본
백관(百官)에게 명하여 고려 왕조의 정령(政令)·법제(法制)의 장점·단점과 변천되어 온 내력의 사목(事目)을
중진(宗親)과 대신(大臣)에게 명하여 여러 도(道)의 군사를 나누어 거느리게 하였다. 【태백산사고본】 1책 1
전 정당 문학(政堂文學) 정도전(鄭道傳)을 명하여 도평의사사 기무(機務)에 참의(參議)하게 하고 상서사 사(
사헌부 대사헌(大司憲) 민개(閔開) 등이 고려 왕조의 왕씨(王氏)를 벗에 두기를 청하니, 임금이 말하였다. "
사헌부에서 또 상소(上疏)하였다. "살가 생각하옵건대, 전하(殿下)께서 하늘의 뜻에 순응하여 혁명(革命)을
를 잘 다스리는 사람은 그 편안함과 위태한 것은 보지 않고 기강(紀綱)이 서지 않은 것을 걱정하는 것입니

그림 4. 크롤링 결과

III. 워드 클라우드

text 변수에 파일을 지정하여 파일을 읽어들이고 sapply 함수를 활용하여 extractNoun을 수행하였다. 그 후 gsub함수를 활용하여 명사지만 분석 시 필요없는 글자들을 제거하였다. 그림 5와 그림 6의 내용은 각각 해당 소스 코드와 조선왕조실록(태조부분)을 R언어를 활용하여 워드 클라우드로 만든 결과물의 일부를 보여준다[3].

```
library('rJava')
library(KoNLP)
library(RColorBrewer)
library(wordcloud)

useSystemDic()
useSejongDic()
useNIADic()

pal2 <- brewer.pal(8,"Dark2")

text <- readLines(file.choose())
noun<-sapply(text,extractNoun,USE.NAMES=F)

noun2<-unlist(noun)
word_count<-table(noun2)
head(sort(word_count,decreasing=TRUE),10)

noun2 <- gsub("【태백산사고본】", "", noun2)
noun2 <- gsub("【국편영인본】", "", noun2)
noun2 <- gsub("【본류】", "", noun2)
noun2 <- gsub("1책", "", noun2)
noun2 <- gsub("2책", "", noun2)
noun2 <- gsub("3책", "", noun2)
noun2 <- gsub("A면", "", noun2)
noun2 <- gsub("B면", "", noun2)
noun2 <- gsub("\\d","", noun2)
noun2 <- Filter(function(x){nchar(x) >= 2}, noun2)

word_count <- table(noun2)
wordcloud(names(word_count),freq=word_count,scale=c(6,0.3),min.freq=3, random.order=F,rot.per=.1,colors=pal2)
```

그림 5. 워드 클라우드 코드

그림 6. 워드 클라우드

Ⅳ. 결 론

본 연구에서는 조선왕조실록을 효율적이고 간편하게 데이터를 수집을 할 수 있도록 웹 크롤러를 제안하고 워드클라우드를 구성하였다.

해당 웹 크롤러는 조선왕조실록 사이트에서 조선왕조실록 사이트를 자동으로 크롤링하기 위한 하이퍼텍스트를 큐에 저장하도록 되어있다. 큐에 저장한 하이퍼텍스트에서 지정된 주소를 크롤링함으로써 조선왕조실록 전체를 크롤링하였고, 크롤링한 결과를 텍스트 파일로 저장하는 기능을 하도록 설계하였다. 또한, 크롤링한 실록 중 태조 시대 텍스트를 분석하기 위하여 명사를 추출하고, 의미없는 단어들을 제외하여 최종적으로 의미 있는 단어들을 중심으로 워드 클라우드로 시각화하였다.

향후 연구에는 조선왕조실록 사이트의 주석부분을 크롤링할 때 배제할 수 있도록 개선하는 연구가 필요하다.

References

[1] Annals of the Joseon Dynasty. Available : http://sillok.history.go.kr/

[2] Class Element. Available : https://jsoup.org/apidocs/org/jsoup/nodes/Element.html

[3] 정용식, 강희구, 빅데이터 분석의 첫 걸음 R로 배우는 코딩, 한국, 생능출판사, 2018.