

## 재난 문자 크롤링 및 시각화 애플리케이션 설계 및 구현

이원주<sup>o</sup>, 박봉균\*, 박문규\*

<sup>o</sup>인하공업전문대학 컴퓨터정보공학과,

\*인하공업전문대학 컴퓨터정보공학과

e-mail: wonjoo2@inhac.ac.kr<sup>o</sup>, P01074553164@gmail.com\*, ansrb0906@naver.com\*

## A Design and Implementation of Disaster Text Crawling and Visualization Application

Won Joo Lee<sup>o</sup>, Bong Kyun Park\*, Mun Kyu Park\*

<sup>o</sup>Dept. of Computer Science and Engineering, InHa Technical College,

\*Dept. of Computer Science and Engineering, InHa Technical College

### ● 요약 ●

본 논문에서는 Python과 Selenium 라이브러리 기반의 재난 문자 크롤링 및 데이터 시각화 애플리케이션을 설계하고 구현한다. 이 애플리케이션의 특징은 웹상에서 재난 문자 데이터를 크롤링(Crawling)하여 그 빈도수에 따라 시각화하는 것이다. 이 애플리케이션을 활용하여 국민재난안전포털에 접속하여 재난 문자 데이터를 크롤링하고, 그 데이터를 Word Cloud를 활용하여 지역별 재난 문자 빈도수를 시각화한다. 지역별 재난 문자 빈도수를 한눈에 보기 쉽게 시각화함으로써 재난문자를 잘 확인하지 않는 사람들에게 해당 지역의 재난 정보를 쉽게 전달하는 기능을 제공한다.

키워드: Python, Pycharm, Word Cloud, Selenium, Crawling

### 1. 서론

현재 코로나바이러스 감염증(COVID-19)으로 인해 많은 재난 문자를 받고 있다. 재난 문자 전송이 처음 시행되었던 시기에는 재난 문자에 대한 관심도가 높았다. 하지만 재난 문자를 받는 빈도수가 증가하면서 재난 문자를 읽지 않고 삭제해 버리는 등, 재난 문자에 대하여 피로감을 호소하고 있다. 2019년 7월부터 2020년 9월까지 재난 문자 발송 건수는 그림 1과 같다[1].

그림 1을 살펴보면 2020년 2~9월 사이에 발송된 재난 문자는 34,684건으로 2019년에 비해 51.2배 증가하였다. 이러한 재난 문자중에 코로나19와 관련된 안내 문자는 84.4%(29,303건)을 차지한다. 재난 문자 역시 4~7월에는 월 2,000여건이 발송되던 것이 8월에는 10,401건, 9월은 7,510건으로 증가하였다.



Fig. 1. 전국 재난 문자 발송 건수

이러한 재난 문자에 대한 부정적 시각을 줄이기 위해서는 재난 문자 발송 빈도와 시기 등을 분석하여 적절하게 대응할 필요가 있다. 따라서 본 논문에서는 재난문자를 수집하고, 수집한 재난 문자 데이터를 시각화 할 수 있는 애플리케이션을 설계하고 구현한다.

## II. 재난 문자 크롤링 및 시각화 애플리케이션 설계

본 논문에서는 재난 문자 크롤링 및 시각화 애플리케이션을 위해 파이썬(Python)과 Selenium 라이브러리를 사용한다. Selenium은 오픈 소스 프레임워크로 운영체제 수준에서 브라우저를 제어 할 수 있는 기능을 제공한다. Selenium 기능은 그림 2와 같다[2].

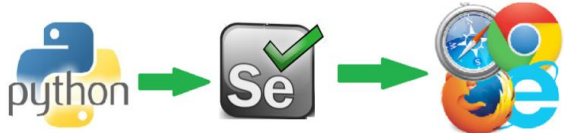


Fig. 2. Selenium 기능

그림 2에서 Selenium은 web driver API를 통하여 운영체제에 설치된 크롬(Chrome) 등의 브라우저를 제어한다. Chrome Selenium Client를 설치하고 국민재난안전포털[3]에 접속해 특정 HTML Tag 에 데이터를 수집한다.

## III. 재난 문자 크롤링 및 시각화 애플리케이션 구현

Python과 Selenium을 통하여 국민재난안전포털에 접속한다. 애플리케이션 실행 시에 옵션으로 입력받은 기간으로 날짜를 설정한다. 옵션으로는 시작 날짜, 종료 날짜를 받는다. 검색 버튼을 클릭하여 재난 문자가 발생한 지역을 검색한 후, 마지막 페이지까지 각 페이지마다 리스트에 있는 문자열들을 하나씩 읽어 발생한 지역 정보만 따로 리스트에 저장한다. 검색한 모든 지역의 정보를 저장하였기 때문에 리스트에는 동일한 지역명이 중복되어 저장된 상태이다. 따라서 발생한 지역별로 그룹핑하여 해당 지역의 재난 문자 전송 빈도를 카운트한다. 카운트가 완료된 정보를 바탕으로 Word cloud를 생성한다. Word cloud는 옵션을 통해서 배경화면, Icon, Font 등을 설정할 수 있다. 해당 애플리케이션은 실행 파일로 필수 옵션을 입력받아 실행시키는 방법이다. 필수 옵션이 없을 경우에는 애플리케이션에서 실패 처리한다.

국민재난안전포털에 접속하여 크롤링한 데이터를 시각화한 결과는 그림 3과 같다.

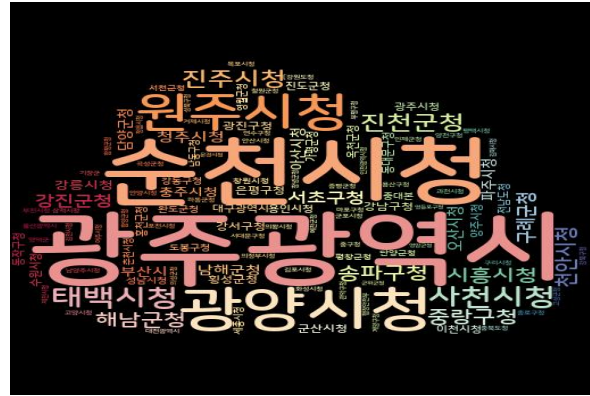


Fig. 3. 시각화 결과

그림 3을 살펴보면 광주광역시 텍스트의 크기가 가장 크면서 앞쪽에 나타나기 때문에 재난 문자 발생 빈도수가 가장 많다는 의미이다. 그리고 순천시청, 광양시청 등의 순서로 재난 문자 발생 빈도수가 많았음을 알 수 있다.

## IV. Conclusions

본 논문에서는 Python과 Selenium 라이브러리를 활용해 재난 문자 발생 빈도 데이터를 시각화하는 애플리케이션을 설계하고 구현하였다. 이 애플리케이션은 특정 기간에 날짜를 지정하면 그 기간에 발생한 재난 문자 발생 데이터를 수집하여 시각화함으로써, 텍스트로 되어 있는 다량의 데이터를 한눈에 보기 쉽게 보여준다. 크롤링 하는 과정에서 네트워크 속도나, 개인 PC의 속도에 따라 지연 현상이 발생하는 단점이 있다. 이러한 지연 현상을 해결하기 위해 지연 시간을 고려하여 3초 간격으로 페이지를 로드하도록 하였다. 그 결과 재난 문자 데이터만이 아닌 다른 데이터들도 크롤링을 통해 데이터를 수집하고 활용 할 수 있었다.

## REFERENCES

- [1] <https://news.joins.com/article/23904257>
- [2] <https://medium.com/swlh/get-started-with-selenium-webdriver-in-under-5-minutes-f9b91e2e9539>
- [3] <https://www.safekorea.go.kr/idsiSFK/neo/sfk/cs/sfc/dis/disasterMsgList.jsp?menuSeq=679>