The Journal of the Convergence on Culture Technology (JCCT)
Vol. 9, No. 6, pp.935-940, November 30, 2023. pISSN 2384-0358, eISSN 2384-0366

http://dx.doi.org/10.17703/JCCT.2023.9.6.935

JCCT 2023-11-112

# 불용어 시소러스를 이용한 비정형 텍스트 데이터 후처리 방법론에 관한 연구

# A Study on Unstructured text data Post-processing Methodology using Stopword Thesaurus

이원조*

**Won-Jo Lee***

**요 약** 인공지능과 빅데이터 분석을 위해 웹 스크래핑으로 수집된 대부분의 텍스트 데이터들은 일반적으로 대용량이고 비정형이기 때문에 빅데이터 분석을 위해서는 정제과정이 요구된다. 그 과정은 휴리스틱 전처리 정제단계와 후처리 머시인 정제단계를 통해서 분석이 가능한 정형 데이터가 된다. 따라서 본 연구에서는 후처리 머시인 정제과정에서 한국어 딕셔너리와 불용어 딕셔너리를 이용하여 워드크라우드 분석을 위한 빈도분석을 위해 어휘들을 추출하게 되는데 이 과정에서 제거되지 않은 불용어를 효율적으로 제거하기 위한 “사용자 정의 불용어 시소러스” 적용에 대한 방법론을 제안하고 R의 워드클라우드 기법으로 기존의 “불용어 딕셔너리” 방법의 문제점을 보완하기 위해 제안된 “사용자 정의 불용어 시소러스” 기법을 이용한 사례분석을 통해서 제안된 정제방법의 장단점을 비교 검증하여 제시하고 제안된 방법론의 실무적용에 대한 효용성을 제안한다.

**주요어** : 빅데이터 분석, 비정형 텍스트 데이터, 워드클라우드, 시각화분석, 전처리 정제, 후처리 정제, 불용어, 시소러스

**Abstract** Most text data collected through web scraping for artificial intelligence and big data analysis is generally large and unstructured, so a purification process is required for big data analysis. The process becomes structured data that can be analyzed through a heuristic pre-processing refining step and a post-processing machine refining step. Therefore, in this study, in the post-processing machine refining process, the Korean dictionary and the stopword dictionary are used to extract vocabularies for frequency analysis for word cloud analysis. In this process, “user-defined stopwords” are used to efficiently remove stopwords that were not removed. We propose a methodology for applying the “thesaurus” and examine the pros and cons of the proposed refining method through a case analysis using the “user-defined stop word thesaurus” technique proposed to complement the problems of the existing “stop word dictionary” method with R’s word cloud technique. We present comparative verification and suggest the effectiveness of practical application of the proposed methodology.

**Key words :** Bigdata analysis, unstructured text data, word cloud, visualization analysis, pre-processing refinement, post-processing refinement, stop words, thesaurus

*정회원, 울산과학대학교 스마트제조공학과 부교수(제1저자)
(울산대학교 전자계산학과 공학박사)
접수일: 2023년 10월 3일, 수정완료일: 2023년 10월 25일
게재확정일: 2023년 11월 10일

Received: October 3, 2023 / Revised: October 25, 2023
Accepted: November 10, 2023
*Corresponding Author: wjlee@uc.ac.kr
Dept. of Smart Manufacturing, Ulsan College, Korea

## Ⅰ. 서 론

최근 4차 산업혁명과 함께 도래한 거대한 디지털 전환의 물결은 누구도 거역할 수 없는 시대적 과제가 되었다. 개인이나 기업은 이를 수용해야만 하는 현실에 직면해 있다. 따라서 이러한 급격한 변화로 많은 사람들과 기업들은 혼돈에 빠져있다. 이러한 변화의 물결은 우리 사회의 전반적인 구조의 변화를 요구하고 있으며, 이 변화를 수용하느냐 마느냐의 선택의 기회마저 허락하지 않는다. 오직 어떻게 얼마나 빠르게 수용하는가가 개인이나 기업의 생존을 가늠하는 척도가 되고 있다. 4차 산업혁명 가장 핵심적인 기술은 인공지능(AI)과 빅데이터 분석이다. 이들의 활용도 향상 위해서는 다양한 정형, 비정형 데이터들의 수집과 결과의 신뢰도를 높일 수 있는 정제기법이 요구된다. 최근에는 SNS와 웹 스크래핑(Web scraping)으로 수집된 대부분의 텍스트 데이터들은 일반적으로 대용량이고 비정형 데이터이기 때문에 정제기법의 중요성이 높아지고 있다. 그리고 기업들은 소비자들의 인터넷 검색 패턴을 분석하여 소비자 요구사항을 분석하고 향후 추세분석 예측을 통해서 마케팅과 신제품 개발, 고객 서비스에 고객의 요구사항을 반영할 수 있고 기업의 지속성장을 위한 의사결정을 보다 명확하게 할 수 있게 되었다.

본 연구에서는 비정형 텍스트 데이터의 워드클라우드 분석의 신뢰도를 향상하기 위한 “사용자 정의 불용어 시소러스”를 활용한 후처리 머시인(Machine) 정제 방법론을 제안하고 제안된 기법의 문제점과 유용성을 검정하기 위해서 다음과 같이 실험을 수행하였다. R의 워드클라우드 시각화 기법으로 제안된 기법을 검증하기 위해서 웹 스크래핑으로 수집된 원시 비정형 텍스트 데이터를 수작업 전처리 과정을 통해서 1차 정제하고 2차 후처리 머시인 정제과정에서 분석결과의 정밀도 향상을 위해서 한국어 어휘들을 “한국어 딕셔너리”로 추출한다. 그리고 불용어 제거를 위해서 “기존 불용어 딕셔너리”를 이용하고 그 다음으로 분석자가 휴리스틱(Heuristic) 판정하여 불용어를 제거하는 과정이 필요한데 이 과정에서 “사용자 정의 시소러스”를 불용어 정제에 적용하는 기법을 R 워드클라우드 사례분석으로 유용성을 검증하고 실무적용에 대한 효용성을 제안한다.

## Ⅱ. 관련연구

### 1. 머시인러닝(Machine learning)

머시인러닝은 축적된 데이터를 학습하고 데이터의 패턴 분석을 통해서 미래를 예측하기 위한 정보를 찾아내는 방법이다. 최근에는 성능이 우수한 컴퓨터 자원들을 저가에 활용할 수 있게 되어 데이터 크기(Volume), 속도(Velocity), 다양성(Variety)의 특징을 가지고 있는 빅데이터 연산으로 효율이고 경제적인 분석이 가능해졌다. 이러한 컴퓨터 처리환경의 변화는 다양하고 방대한 데이터들을 과학적인 알고리즘으로 실시간 분석할 수 있는 시스템의 구축이 가능하다. 이를 이용하여 조직들은 새로운 기회창출과 위험요소를 예측할 수 있게 되어 관심이 매우 높다. 또한 머시인러닝은 비정형 텍스트 데이터 분석에도 유용하게 이용되는데 텍스트 마이닝(Text mining)이라고 한다[1].

### 2. 비정형 텍스트 데이터

텍스트 데이터 분석은 수집된 텍스트 데이터를 데이터 마이닝을 통해서 의미 있는 정보를 추출하는 과정이며, 이를 기반으로 경영 의사결정의 중요한 자료로 사용된다. 이러한 과정에서 사용되는 대부분의 텍스트 데이터는 비정형이기 때문에 반드시 전처리 과정과 후처리 과정을 통해서 정형 데이터로 정제하는 과정이 요구된다. 정제의 정밀도가 텍스트 마이닝 결과의 신뢰도에 매우 중요하다. 따라서 이 과정에서 머시인 정제가 데이터의 정제과정에 이용된다. 이렇게 정형화된 텍스트 데이터는 데이터 마이닝 알고리즘이 적용된 컴퓨터 시스템에서 처리되고 분석되어 해석에 반영된다[2].

## Ⅲ. 한국어 텍스트 데이터 분석

### 1. 한국어 텍스트 데이터 분석 기법

빅데이터 분석에서 워드클라우드 시각화 분석기법은 비정형 텍스트 데이터를 분석하는 도구이다. 다양한 매체와 유형으로 수집된 비정형 텍스트 데이터를 휴리스틱 전처리 정제단계와 한글문장을 형태소 분석으로 어휘들을 추출하고 각 어휘별로 출현 빈도수를 계산한다. 이 과정에서 불용어나 중요도가 낮은 단어들을 제거하는 머시인 후처리 정제단계를 통해서 정밀도가 높고 유의미한 어휘들을 R의 워드클라우드 기

법으로 시각화하여 핵심 이슈를 해석을 통해서 도출하는 데이터 분석기법이다. 워드클라우드 시각화 분석에서 각 각의 어휘는 칼라로 구분되고 출현 빈도수가 높은 어휘는 일반적으로 중앙에 위치하고 크게 표시된다. 여기서 빈도수가 높은 어휘는 중요도가 높고 유의미한 어휘로 해석한다. 다음은 후처리 정제단계에서 어휘의 출현 빈도수에 따른 연구자의 중요도 해석구분을 위한 규정으로 다음과 같이 첫째, 빈도가 높고 중요도가 낮은 단어 둘째, 빈도와 중요도가 높은 단어 셋째, 빈도가 낮고 중요도가 낮은 단어 넷째, 빈도가 낮고 중요도가 높은 단어 다섯째, 빈도는 높지만 자격이 없는 값으로 기 정립된 5가지 유형 구분을 사용한다.  상기의 첫째, 셋째, 다섯째 항은 제거대상 어휘로 분류하여 불용어 시소러스에 저장한다. 그리고 둘째 항과 넷째 항은 수렴하고 기존의 불용어 딕셔너리에는 없으나 제거할 어휘들은 별도의 불용어 시소러스에 저장하고 불용어 딕셔너리와 불용어 시소러스를 병합하여 후처리 정제처리에 이용한다[2].

## 2. 제안 텍스트 데이터 정제모델

기존 텍스트 데이터 정제모델로 수집된 다양한 형식의 데이터들을 R 프로그램의 소스코드 내부에 포함하여 불용어를 제거하는 “소스코드 내 불용어 기입” 방법을 사용하고 있다. 그러나 제안된 모델은 그림 1과 같이 수집된 다양한 형식의 데이터들을 R 프로그램의 외부에 사용자가 불용어 시소러스를 생성하여 불용어를 제거하는 “사용자 정의 불용어 시소러스”를 이용하여 불용어를 제거하는 방법의 사용을 제안한다[2][3].

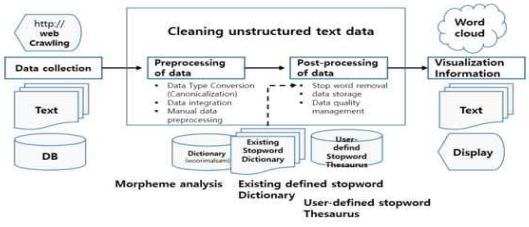


그림 1. 제안 텍스트 데이터의 데이터 정제모델
Figure 1. Data purification model for proposed text data

## 3. 텍스트 데이터 전처리 정제

본 연구에서 사용할 비정형 텍스트 데이터는 대한민국 청와대 홈페이지에서 웹 스크래핑하여 텍스트 파일 형식으로 수집하고 전처리 휴리스틱(Heuristic) 정제과정을 거치고 텍스트 파일로 형태로 저장하게 되는데 이 과정에서 주의해야 할 사항은 다음과 같다. 첫 번째, 맨 마지막 문장이 끝나면 반드시 줄바꿈을 하고 저장해야 한다. 그렇게 하지 않으면 후처리 머시인(Machine) 정제과정에서 텍스트 파일을 읽을 때 오류가 발생하게 된다. 두 번째, 파일 저장시 [파일] 메뉴에서 [다른 이름으로 저장]을 선택하고 [인코딩]을 [UTP-8]을 선택하고 [저장] 한다. 그리고 R 프로그램에서 분석용 파일을 읽어올 때 폴드명이나 파일명은 영문으로 사용하는 것이 오류를 줄일 수 있는 방법이다.

## 4. 텍스트 데이터 정제과정

### 1) 워드클라우드를 위한 환경설정

텍스트 데이터의 R 워드클라우드 분석을 위한 기본 환경설정을 위해서는 "wordcloud", "RColorBrewer", "KoNLP" 등의 패키지가 사용되는데, 그림 2는 R 워드클라우드 분석을 위한 기본 환경설정 소스코드이다. 다음 실험용 소스들은 “모두를 위한 R 데이터분석 입문 한빛아카데미”의 소스코드를 인용으로 작성되었다.

```
Sys.setenv(JAVA_HOME='C:/Program Files/Java/jdk-19')
#-----------------------------------
library(wordcloud)                  #
library(KoNLP)                      #
library(RColorBrewer)
library(wordcloud2)                 #
#-----------------------------------
```

그림 2. R 워드클라우드 분석을 위한 기본 환경설정
Figure 2. Basic environment settings for R word cloud analysis

### 2) 한국어 딕셔너리를 이용한 어휘 추출

분석대상 비정형 텍스트 데이터에서 한국어 딕셔너리(woorimalsam)를 이용하여 명사 어휘들을 간단하게 추출할 수 있는데, 다음 그림 3은 한국어 딕셔너리를 이용한 어휘 추출 소스코드이다.

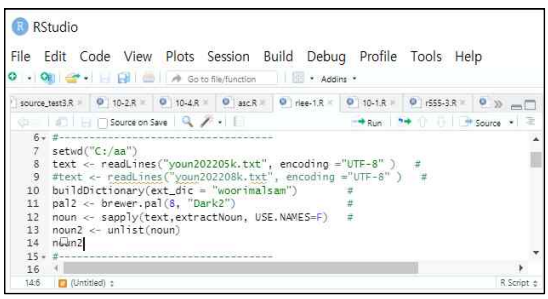


그림 3. 한국어 딕셔너리를 이용한 어휘 추출
Figure 3. Vocabulary extraction using Korean dictionary

3) 불용어 제거방법

후처리 머시인 정제과정에서 불용어를 제거하는 방법은 다음과 같이 3가지가 있는데, 첫째, 그림 4와 같이 소스코드 내 불용어 기입 방법, 둘째, 그림 5와 같이 범용 불용어 딕셔너리 방법, 셋째, 그림 6과 같이 사용자 정의 불용어 시소러스 방법을 이용한 제거가 있다.

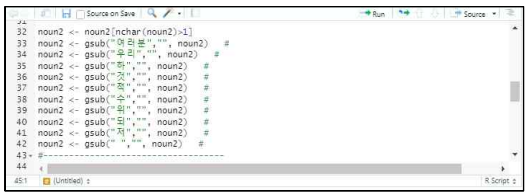


그림 4. 소스코드 내 불용어 기입 방법을 이용한 제거
Figure 4. Removal using stopword entry method in source code

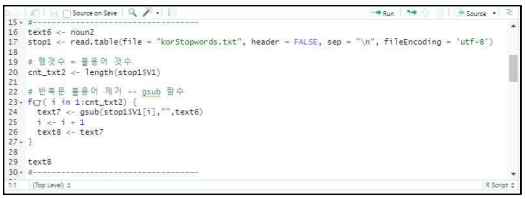


그림 5. 범용 불용어 딕셔너리 방법을 이용한 제거
Figure 5. Removal using the universal stopword dictionary method


그림 6. 사용자 정의 불용어 시소러스 방법을 이용한 제거
Figure 6. Removal of user-defined stop words using thesaurus method

4) 빈도수 상위 30개 어휘 추출

이 단계에서는 추출된 어휘들의 출현 빈도수를 확인하고 시각화 정보를 제공하여 제거 대상어휘 여부를 분석자가 쉽게 판정할 수 있도록 해준다. 그림 7에서는 출현 빈도수 상위 30개의 어휘들을 추출할 수 있다. 이 단계에서 추출되는 어휘의 수는 간단하게 지정이 가능하다.

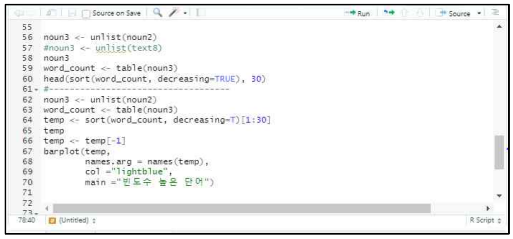


그림 7. 빈도수 상위 30개 어휘의 추출
Figure 7. Extraction of top 30 frequent vocabulary words

5. 정제방법 장단점 비교표

다음 표 1은 불용어를 제거하는 3가지 정제방법의 장단점 비교표이다. 각 각의 장단점을 보면 방법 1은 제거 대상 불용어가 적을 경우에는 편리하나 많을 경우에는 적용하기 어렵고, 방법 2는 일반적인 불용어 제거에는 유용하나 세부적인 적용에는 한계가 있으며, 방법 3은 분석자가 불용어를 시소러스에 등록하여 쉽게 적용이 가능해 세부적인 불용어 제거에는 매우 유용한 방법으로 평가된다[3~6].

표 1. 정제방법의 장단점 비교표
Table 1. Comparison table of advantages and disadvantages of purification methods

| 정제방법 | 장점 | 단점 |
|---|---|---|
| 1. 소스코드 내 불용어 기입 | -불용어 개수가 적을 경우 편리하다. | -불용어 개수가 많을 경우 적용이 어렵다.<br>-재사용이 어렵다. |
| 2. 범용 불용어 딕셔너리 | -공개된 불용어 딕셔너리의 사용이 가능하다. | -사용자 맞춤형 불용어 제거가 어렵다. |
| 3. 사용자 정의 시소러스 | -불용어 개수가 많아도 쉽게 사용이 가능하다.<br>-불용어의 재사용성과 정제 정밀도가 높다. | -분석자의 숙련된 불용어 선정역량이 요구된다. |

## Ⅳ. 워드클라우드 시각화 사례구현

1. 사례구현 방법

본 연구에서 R 워드클라우드 분석기법을 적용할 사례구현은 대한민국 제20대 윤석열 대통령의 취임사 연설문(2022.05.10.) 원문과 제78주년 광복절 기념사 연설문(2023.08.15.) 원문을 대한민국 청와대 홈페이지에서 웹

스크래핑하여 비정형 텍스터 데이터를 수집하고 이를 전처리 휴리스틱 정제과정을 통해서 정형화하고 후처리 머시인 정제과정을 통해서 분석이 가능한 정형 텍스터 데이터로 정제한다. 이 정제과정에서 상기의 불용어를 제거하는 3가지 후처리 머시인 정제방법을 R 워드클라우드 시각화 사례분석에 적용하여 제안된 “사용자 정의 불용어 시소러스” 방법의 유용성을 검증하고 이 과정에서 추출된 핵심 이슈들을 분석하여 대통령의 주요정책 변화를 해석한다[7~9].

## 2. 사용자 정의 불용어 시소러스의 생성

사용자 정의 불용어 시소러스는 기존의 방법으로 제거되지 않은 어휘들을 그림 8과 같이 “사용자 정의 불용어 시소러스 데이터 셋”에 추가 하여 생성하고 이를 그림 6과 같이 쉽게 불용어 제거에 사용할 수 있다.

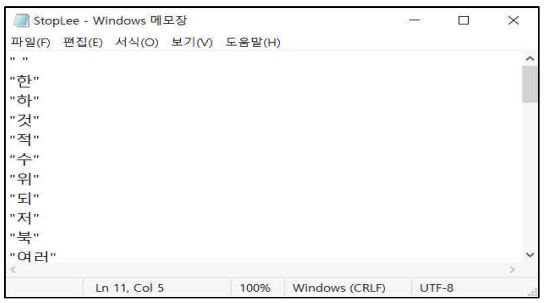


그림 8. 사용자 정의 불용어 시소러스 데이터 셋
Figure 8 Dataset from a custom stopword thesaurus

## 3. 워드클라우드 시각화 분석결과

다음 그림 9는 “제20대 윤석열 대통령의 취임사 연설문” 시각화 분석결과이고 그림10은 “윤석열 대통령의 제78주년 광복절 기념사 연설문” 시각화 분석결과이다.


그림 9. 취임사 연설문 시각화 분석결과
Figure 9. Visualization analysis results of the inauguration speech


그림10. 광복절 기념사 연설문 시각화 분석결과
Figure10. Visualization analysis results of the Liberation Day commemoration speech

## 4. 시각화 분석결과 해석

워드클라우드 시각화 사례구현 사례를 보면 그림 8과 같이 대통령의 취임사에서 “자유”, “국민”, “세계”, “평화”, “민주주의”, “국제”, “사회”와 어휘들의 출현빈도가 높은 핵심 이슈로 나타났으나, 출현빈도는 낮으나 분석자의 휴리스틱으로 “공정”, “번영”, “시장경제”, “비핵화” 등의 어휘를 핵심 이슈로 해석하였고 그림 9와 같이 대통령의 광복절 기념사에서는 “자유”, “협력”, “세계”, “안보”, “평화”, “민주주의”, “번영”, “가치동맹” 이 출현빈도가 높은핵심 이슈로 나타났으나, 출현빈도는 낮으나 분석자의 휴리스틱으로 “공정”, “연대”, “양극화 해소”, “비핵화”, “반지성주의” 등의 어휘를 핵심 이슈로 분석되어 대통령의 주요정책의 변화로 해석하였다. 따라서 불용어의 효율적인 제거가 핵심 이슈의 도출을 용이하게 지원해 결과해석의 정밀도를 높일 수 있으며, 낮은 빈도의 어휘를 핵심 이슈로 판정하는 데는 분석자의 휴리스틱 개입이 반드시 필요하기 때문에 분석자의 역량이 분석결과의 정밀도에 가장 큰 영향을 미치게 된다[10~12].

# V. 결 론

본 연구의 사례연구는 웹 스크래핑으로 수집된 비정형 텍스트 데이터의 휴리스틱 전처리 정제과정과 R 프로그램을 사용한 머시인 후처리 정제과정에서 발생하는 문제점 들을 도출하고 제안된 “사용자 정의 시소러스”의 생성과 불용어 시소러스를 사용한 정제의 효용성을 검증한다. 따라서 이 과정에서 도출된 문제점들은 첫째,

R 워드클라우드 분석을 위한 한국어(KoNLP) 사용을 위한 환경설정이 어렵고 둘째, 한국어 딕셔너리에 없는 신조어나 전문용어, 외래어 한글표기 등은 분석자가 임의로 추가해야 하고 셋째, “소스코드 내 불용어 기입” 방법은 분석자가 임의로 불용어 여부를 판단해야 하고 제거 대상 어휘가 많을 경우에는 적용이 매우 어렵고 재사용성이 낮으며, 넷째, “범용 불용어 딕셔너리“ 방법은 사용이 편리하나 범용으로 불용어 제거의 정밀도가 낮고 다섯째, ”사용자 정의 시소러스“ 방법은 불용 어휘의 등록이 쉽고 분석자의 전문분야에 최적한 불용어 시소러스의 축적으로 재사용성 높다. 따라서 제안된 방법이 기존 방법의 불완전성을 보완할 수 있는 방법론으로 비정형 텍스트 데이터의 시각화 분석결과의 정밀도 향상에 유용할 것으로 판단된다.

향후 연구과제는 ”범용 불용어 딕셔너리“를 대체하고 전문 분야별로 축적된 ”사용자 정의 시소러스“를 제작하여 분석자가 한국어 텍스트 데이터 분석결과의 정밀도 향상에 편리하게 이용할 수 있도록 추가적인 연구가 진행되어야 할 것으로 사료된다.

## References


[1] W. Lee, A Study on the Use of Stopword Corpus for Cleansing Unstructured Text Data, JCCT, Vol. 8, No. 6, pp.891-897, 2022. DOI: 10.17703/JCCT.2022.8.6.891

[2] W. Lee, A Study on Data Cleansing Techniques for Word Cloud Analysis of Text Data, JCCT, vol. 7, No. 4, pp. 745-750, 2021. DOI: 10.17703/JCCT.2021.7.4.745

[3] W. Lee, A Study on Word Cloud Techniques for Analysis of Unstructured Text Data, JCCT, vol. 6, No. 3, pp. 337-341, 2020. DOI: 10.17703/JCCT.2020.6.4.715

[4] Kumar, P. Thakur, K. Gupta, and A. Pal, 2015, Text mining approach to analyse the relation between obesity and breast cancer data, ILNS

[5] M. Han, Y. Kim, C. Lee, Analysis of News Regarding New southeastern Airport Using Text Mining Techniques, Smart Media Journal, Vol. 6, No. 1, 2017.

[6] J. Lee, D. Yun, S. O, C. Lee, A Big Data Analysis of Civel Complaint Texts Using R Language, KIICE, 2020.

[7] Insun Lee and 1 others, Unstructured data analysis and visualization, Korean Psychology Association, 2018.

[8] Jongyong LEE, A Study on Tourism Analysis in Uijeongbu Region Using Big Data, JCCT, vol. 6, No. 1, pp. 413-419, 2020.

[9] Sunghuk Moon, Big data environment analysis and research on ways to secure global competitiveness, JCCT, vol. 5 No. 2, pp. 361-367

[10]Giseop Noh, An Analysis on Internet Information using Real Time Search Words, JCCT, vol. 4, No. 4, pp. 337-341, 2018.

[11]I. Chun, D. Park, Y. Kang, Python and data science, Saengneun Publishing, pp. 222-233, 2019.

[12]M. Chi , S. Lin, S. Chen, C. Lin, T. Lee, Morphable word Clouds for Time-Varying Text Data Visualization, IEEE, 2015.