The Journal of the Convergence on Culture Technology (JCCT)
Vol. 8, No. 6, pp.891-897, November 30, 2022. pISSN 2384-0358, eISSN 2384-0366

http://dx.doi.org/10.17703/JCCT.2022.8.6.891

JCCT 2022-11-110

# 비정형 텍스트 데이터 정제를 위한 불용어 코퍼스의 활용에 관한 연구

# A Study on the Use of Stopword Corpus for Cleansing Unstructured Text Data

이원조*

**Won-Jo Lee***

**요 약** 빅데이터 분석에서 원시 텍스트 데이터는 대부분 다양한 비정형 데이터 형태로 존재하기 때문에 휴리스틱 전처리 정제와 컴퓨터를 이용한 후처리 정제과정을 거쳐야 분석이 가능한 정형 데이터 형태가 된다. 따라서 본 연구에서는 텍스트 데이터 분석 기법의 하나인 R 프로그램의 워드클라우드를 적용하기 위해서 수집된 원시 데이터 전처리를 통해 불필요한 요소들을 정제하고 후처리 과정에서 불용어를 제거한다. 그리고 단어들의 출현 빈도수를 계산하고 출현빈도가 높은 단어들을 핵심 이슈들로 표현해 주는 워드클라우드 분석의 사례 연구를 하였다. 이번 연구는 R의 워드클라우드 기법으로 기존의 불용어 처리 방법인 “내포된 불용어 소스코드” 방법의 문제점을 개선하기 위하여 “일반적인 불용어 코퍼스”와 “사용자 정의 불용어 코퍼스”의 활용 방안을 제안하고 사례 분석을 통해서 제안된 “비정형 데이터 정제과정 모델”의 장단점을 비교 검증하여 제시하고 “제안된 외부 코퍼스 정제기법”을 이용한 워드클라우드 시각화 분석의 실무적용에 대한 효용성을 제시한다.

**주요어** : 빅데이터, 텍스트 데이터, 데이터 정제, 워드클라우드, 시각화 분석, 불용어, 코퍼스

**Abstract** In big data analysis, raw text data mostly exists in various unstructured data forms, so it becomes a structured data form that can be analyzed only after undergoing heuristic pre-processing and computer post-processing cleansing. Therefore, in this study, unnecessary elements are purified through pre-processing of the collected raw data in order to apply the wordcloud of R program, which is one of the text data analysis techniques, and stopwords are removed in the post-processing process. Then, a case study of wordcloud analysis was conducted, which calculates the frequency of occurrence of words and expresses words with high frequency as key issues. In this study, to improve the problems of the “nested stopword source code” method, which is the existing stopword processing method, using the word cloud technique of R, we propose the use of “general stopword corpus” and “user-defined stopword corpus” and conduct case analysis. The advantages and disadvantages of the proposed “unstructured data cleansing process model” are comparatively verified and presented, and the practical application of word cloud visualization analysis using the “proposed external corpus cleansing technique” is presented.

**Key words :** Big Data, Text Data, Data Cleansing, Wordcloud, Visualization Analysis, Stopwords, Corpus

*정회원, 울산과학대학교 산업경영공학과 부교수 (단독저자)
(울산대학교 전자계산학과 공학박사/울산과학대학교 컴퓨터 IT학부 22년/현재 산업경영공학과 부교수)
접수일: 2022년 10월 21일, 수정완료일: 2022년 11월 3일
게재확정일: 2022년 11월 8일

Received: October 21, 2022 / Revised: November 3, 2022
Accepted: November 8, 2022
*Corresponding Author: wjlee@uc.ac.kr
Dept. of Industrial Management Eng., Ulsan College, Korea

## Ⅰ. 서 론

최근 4차 산업혁명과 함께 찾아온 세상의 변화에 대한 요구는 우리들에게 선택의 기회를 허락하지는 않는다. 먼저 와버린 미래는 더욱 가속화될 것이며, 우리사회의 전반적인 구조적 변화를 일방적으로 몰아가고 있다. 이러한 변화의 속도는 나선형 효과(Spiral Effect)와 같이 급속도로 진행되고 있다. 그러나 이 시점에서 미래의 변화에 대한 예측도 상당부분이 추론일 뿐 예측이 불가능하다. 그러나 단 한 가지 분명한 사실은 이 변화의 물결에 편승하지 못하는 다수의 기업이나 개인은 도태될 것이란 것이다. 최근 우리사회의 전반적인 부문에서 빅데이터 분석의 중요성이 높아지고 활용이 증가되고 있으나 대부분의 원시 데이터들은 비정형 데이터 형태이기 때문에 빅데이터 분석의 신뢰도를 높이기 위해서는 비정형 텍스트 데이터의 정제기법에 대한 관심이 높아지고 있다. 따라서 본 연구에서는 비정형 텍스트 데이터 분석을 위한 원시 데이터를 수집하고 빅데이터 분석을 위한 수작업 전처리 과정을 통해서 텍스트 데이터를 정형화하고 불용어 제거를 위한 “사용자 정의 불용어 코퍼스”를 이용한 머시인(Machine) 후처리 정제를 반복하여 비정형 텍스트 데이터의 정밀도를 높일 수 있는 효율적인 정제모델을 제안하고 사례분석을 통해서 이 모델을 이용한 워드클라우드 시각화 분석의 실무적용에 대한 효용성을 제시한다.

본 연구의 사례는 “문재인 대통령 취임사”와 “윤석렬 대통령 취임사”의 텍스트 문서를 빅데이터 분석 툴인 R의 워드클라우드를 이용하여 대통령의 정책에 대한 핫 이슈(Hot Issue)를 도출하고 이를 비교 해설하는 사례 연구를 통해서 제안 모델의 유용성을 검증한다. 따라서 R의 워드클라우드를 이용하여 핵심 단어들을 시각화하여 분석하는 방법과 이 과정에서 발생하는 다양한 문제점을 도출하고 이에 대한 해법 모델을 제시하고 워드클라우드의 실무적용에 대한 다양한 활용 가능성을 제시한다.

## Ⅱ. 관련연구

### 1. 비정형 텍스트 데이터 분석 기술

오늘날 인터넷을 통한 정보의 유통은 다양한 형태의 원시 데이터 형태로 존재하며, 다양한 매체를 통해서 유통되고 있어 이들을 수집하여 분석하기 위해서는 분석자에 의한 휴리스틱(Heuristic) 정제와 머시인(Machine) 정제과정을 통해서 정형화하고 분석이 용이한 형태로 가공되고 저장되어야 한다. 비정형 텍스트 데이터 분석에서 인터넷 검색어 분석은 검색어를 사용하여 이용자들의 관심사를 분석할 수 있도록 지원해주는 사이트 들이 많이 있다. 이것들 중에서 대표적인 것이 구글 트렌드와 네이버 데이터 랩(Data lab)등이 있다. 그러나 본 연구와 같은 비정형 텍스트 데이터인 연설문의 핵심 이슈분석에는 적합하지는 않다. 따라서 본 연구의 구현사례에서 사용된 비정형 텍스트 데이터 분석 방법은 R 프로그램의 워드클라우드를 사용하고 빅데이터 분석에 사용되는 “세종 한글사전”을 사용하였으며, 사례 검증을 위한 소스 코드는 “모두를 위한 R데이터분석”의 워드클라우드 기법을 응용하였다[1].

### 2. 워드크라우드 분석 기법

빅데이터 분석에 많이 사용하는 데이터들은 숫자 형태의 정형화된 데이터가 일반적이다. 그러나 문자나 문장 형태의 데이터 분석에 대한 활용에 관심이 높아지고 있다. 비정형 데이터의 사례는 전자신문, SNS기사, 이메일, 문자 메시지, 전자책, 음성, 영상 등 매우 다양한 형태로 존재한다. 그러나 이러한 비정형 텍스트 데이터의 분석에는 많은 전문성과 시간을 요구한다. 워드클라우드는 텍스트 데이터를 분석하는 대표적인 분석기법으로 휴리스틱(Heuristics) 전처리 정제와 형태소 분석을 통해 단어(명사)들을 추출하고 불필요한 데이터들을 머시인(Machine)을 이용한 후처리 정제 후에 각 단어들의 출현 빈도수를 계산하여 워드클라우드 기법으로 시각화 분석하는 것이다. 여기서 출현빈도가 높은 단어는 중심에 크게 표시된다. 이는 중요도가 높거나 관심도가 큰 핵심 이슈들(Issues)로 해석한다[2].

### 3. 한국어 불용어 제거

불용어(Stopword)는 텍스트 데이터에서 유의미한 단어 토큰만을 선별하기 위해서는 큰 의미가 없는 단어 토큰을 제거하는 작업이 필요하다. 여기서 큰 의미가 없다는 것은 출현 빈도는 높으나 중요도가 낮은 단어들을 의미한다. 따라서 이러한 단어들을 불용어(Stopword)라고 한다. 한국어에서 불용어를 제거하는 방법으로는 간단하게는 토큰화 후에 조사, 접속사 등을 제거하는 방법이

있다. 그러나 불용어를 제거하는 과정에서 조사나 접속사와 같은 단어들뿐만 아니라 명사, 형용사와 같은 단어들 중에서 불용어로 제거해야 하는 단어들이 있다. 그래서 사용자가 직접 불용어 코퍼스를 만들어야 정밀도가 높아진다[3].

## III. 비정형 텍스트 데이터 분석

### 1. 텍스트 데이터 분석 모델

한국어 비정형 텍스트 데이터의 분석과정에서 수집된 원시 데이터는 일반적으로 워드클라우드 분석에는 적합하지 않은 상태이다. 따라서 컴퓨터를 이용한 분석이 가능한 형태로 수집된 데이터를 분석자가 1차 휴리스틱(Heuristics) 전처리 정제를 한 후에 텍스트 데이터로 저장한다. 빅데이터 분석에서 데이터 정제는 전처리와 후처리로 나누어지는데 분석에 적합한 형태로의 데이터 정제가 매우 중요하다. 그 이유는 데이터 정제의 정도에 따라서 데이터의 분석이 불가능 하거나 분석결과의 신뢰성이 좌우되기 때문이다. 그림 1은 “기존 비정형 데이터 정제과정 모델”로 한글사전을 사용하여 형태소 분석에서 명사 단어들을 추출하고 휴리스틱(Heuristics) 인지된 불용어(Stopword)들을 R 프로그램 코드에 추가하여 제거하는 후처리 과정을 반복해야 하는데, 제거대상 불용어가 많을 경우에는 매우 복잡하고 번거로운 일이며, 불용어가 축적이 되지 않아 재사용이 어려운 단점이 있다. 따라서 그림 2는 “제안된 비정형 데이터 정제과정 모델”은 외부의 “일반적인 불용어 코퍼스”와 “사용자 정의 불용어 코퍼스”를 적용하여 외부 사용자 정의 불용어 축적을 통한 재사용으로 후처리 과정의 효용성과 정제의 정밀도 향상을 위한 “사용자 정의 불용어 코퍼스” 방법을 제안한다[4].

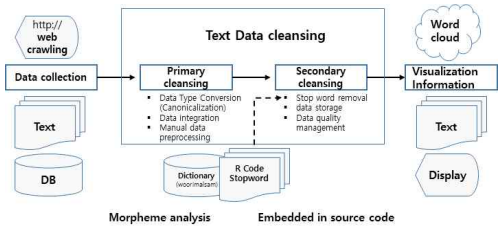


그림 1. 기존 비정형 데이터 정제과정 모델
Figure 1. Model of the existing unstructured data cleansing process diagram

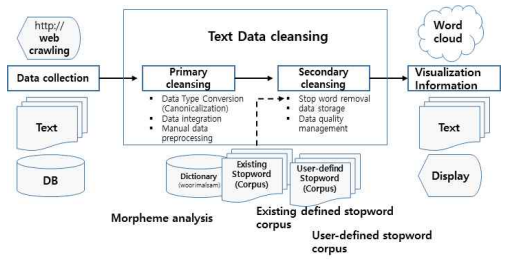


그림 2. 제안된 비정형 데이터 정제과정 모델
Figure 2. Model of the proposed unstructured data cleansing process diagram

### 2. 비정형 텍스트 데이터 수집

본 연구의 비정형 텍스트 빅데이터 정제 사례에 사용하게 될 문서는 대한민국 청와대 홈페이지에서 “문재인 대통령 취임사”와 “윤석렬 대통령 취임사” 원문을 수집하고 수집된 문서는 일반적으로 테스트 파일의 형태로 저장하고 저장된 파일은 윈도우의 “메모장”을 사용하여 “열기”를 하여 수작업 전처리를 통해서 1차 가공한다. 그리고 마지막 문장은 다음 라인으로 넘기고 “파일”메뉴의 “다른 이름으로 저장”으로 저장시 “인코딩”옵션에서 “UTF-8”을 선택하고 파일의 이름은 영문으로 저장한다. 이 방법으로 저장하지 않으면 오류가 발생할 수 있다[5][6][7][8][9][16].

### 3. 비정형 텍스트 데이터 정제

1) 워드클라우드 환경설정

R 프로그램에서 텍스트 데이터의 워드클라우드 분석을 위해서는 “wordcloud”, “RColorBrewer”, “KoNLP” 등의 환경설정을 위한 패키지의 설치가 필요한데 다음의 명령어들로 간단하게 설치가 가능하다. 그러나 한국어 텍스트 분석을 위한 “KoNLP” 패키지 설치와 부가적인 환경설정은 인내를 요구하는 매우 번거로운 과정이다.

```
install.packages(“wordcloud”)
install.packages(“RColorBrewer”)
install.packages(“KoNLP”)
```

2) 한글사전을 사용한 명사추출

다음 그림 3은 세종한글사전(useSejongDic())을 사용하여 외부파일 읽기(readLines()) 분석 대상문서에서 명사 단어를 추출하는 R의 소스코드이고 이 코드의 실행

결과로 단어(명사)들이 추출할 수 있다.

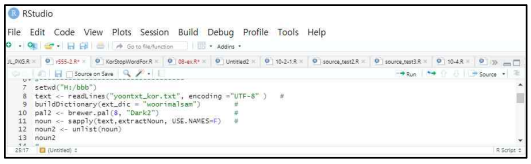


그림 3. 한글사전을 사용한 명사단어 추출 소스코드
Figure 3. Source code for extracting noun words using Korean dictionary

3) 사전에 단어 추가방법

빈도수가 높고 유의미한 단어이나 사전에 등록되어 있지 않아서 워드클라우드 시각화에 표시되지 않은 단어들은 그림 4와 같이 한글사전에 추가하여 표현이 가능하다.

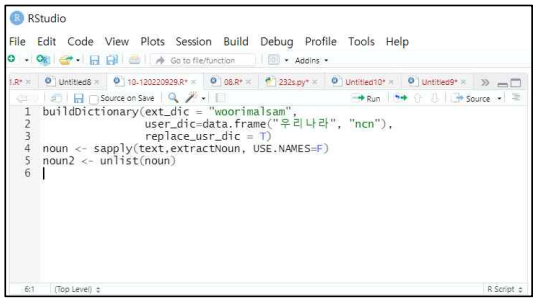


그림 4. 사전에 없는 빈도수 높은 단어의 추가 소스코드
Figure 4. Source code for adding high frequency words that are not in the dictionary

4) 불용어 제거방법

불용어를 제거하는 방법은 “내포된 불용어 소스코드”, “일반적인 불용어 코퍼스”, “사용자 정의 불용어 코퍼스”등으로 나눌 수 있는데, 그림 5는 “내포된 불용어 소스코드”를 사용하여 제거하는 방법이고 그림 6은

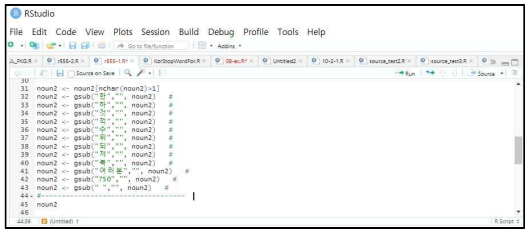


그림 5. 내포된 불용어 소스코드
Figure 5. Nested stopword source code

“일반적인 불용어 코퍼스”와 “사용자 정의 불용어 코퍼스”에서 외부 코퍼스 파일을 이용하여 불용어를 제거하는 소스코드이다. 이들 방법 이외에도 다른 방법은 여기서는 다루지 않는다[16][17].

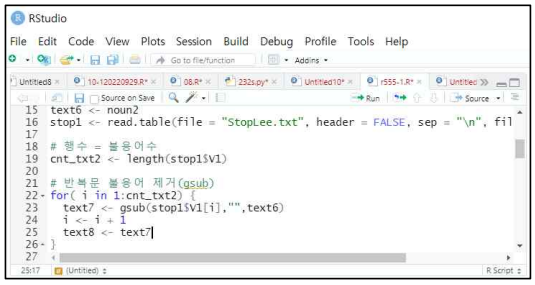


그림 6. 외부 불용어 코퍼스 사용 소스코드 (korStopwords.txt/StopLee.txt)
Figure 6. Common stopword corpus source code

5) 단어의 빈도수 계산(상위 20단어 추출)

다음 그림 7은 1차 수작업 정제와 2차 불용어가 추출된 단어들 중에서 단어의 출현빈도가 가장 높은 상위 20개 단어들을 추출하여 나타내고 이 결과를 막대그래프로 시각화하는 소스코드이다.

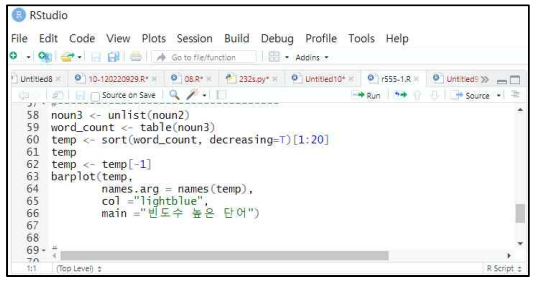


그림 7. 출현 빈도수 상위 20단어 막대그래프 시각화 소스코드
Figure 7. Top 20 words in frequency of appearance bar graph visualization source code

6) 워드클라우드 시각화 결과

다음 그림 8은 텍스트 데이터의 워드클라우드 시각화 분석을 위한 소스코드이다. 시각화 결과 핵심 단어가 나타나지 않거나 불용어가 남아 있다면 이전 단계를 반복하면서 시각화 결과의 정밀도를 높여간다. 따라서 이러한 머시인 정제 과정을 통해서 유의미한 핵심 이슈들을 추출하여 빅데이터 분석의 결과를 도출한다.

그림 8. 워드클라우드 시각화 분석 소스코드
Figure 8. Source code for word cloud visualization analysis

## Ⅳ. 워드클라우드 시각화 사례 구현

### 1. 실무구현 사례

본 연구의 실무구현 사례는 우리나라 대통령의 취임사를 워드클라우드 시각화 분석을 통해서 핫 이슈를 추출하고 대한민국의 대내외 주요 정책 기조의 변화를 비교 분석한다. 이 과정에서 불용어 제거를 위한 3가지 방법론을 적용하고 각 방법의 장단점을 평가하여 한국어 텍스트 데이터의 워드클라우드 시각화를 위한 불용어 제거 모델을 검증한다. 그림 9는 불용어 제거를 위한 3가지 방법론이며, 그림10에서 그림12는 정제 단계별 실행 결과들이고 그림13과 그림14는 워드클라우드 시각화 결과이다. 그리고 표 1은 3가지 불용어 정제 방법별 장단점 비교표이다. 이 결과와 같이 "사용자 정의 불용어 코퍼스" 방법이 비정형 텍스트 데이터 정제과정에서 유용한 것으로 평가되었다[10][11][12][13].

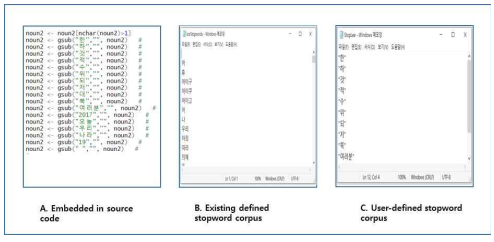


그림 9. 불용어 제거를 위한 3가지 방법론
Figure 9. 3 Methodologies to Eliminate Stopword

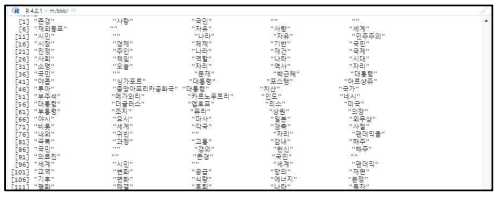

그림 10. 불용어 처리후 단어(명사) 추출
Figure 10. Result of extracting noun words after removing stopword


그림 11. 정제후 출현빈도 탑20개 단어-Yoon
Figure 11. Top 20 words with frequency of appearance after cleansing-Yoon


그림 12. 정제후 출현빈도 탑20개 단어-Moon
Figure 12. Top 20 words with frequency of appearance after cleansing-Moon


그림 13. 정제 후 워드클라우드 시각화 결과-Yoon
Figure 13. Word cloud visualization results after cleansing-Yoon


그림 14. 정제 후 워드클라우드 시각화 결과-Moon
Figure 14. Word cloud visualization results after cleansing-Moon

2. 불용어 정제 방법의 장단점 비교평가

표 1. 3가지 불용어 정제 방법별 장단점 비교
Table 1. Comparison of advantages and disadvantages of 3 stopword cleansing methods

| 구 분 | 장 점 | 단 점 |
|---|---|---|
| A. Embedded in source code | -불용어 개수가 적을 경우 적용이 용이함<br>-명확한 불용어 제거가 가능함 | -불용어 개수가 많을 경우 적용이 어렵움<br>-불용어 축적이 되지 않아 재사용이 어려움 |
| B. Existing defined stopword corpus | -기존의 불용어 코퍼스 사용이 가능함<br>-쉽게 외부 불용어 적용이 가능함 | -명확한 불용어 제거가 어려움<br>-맞춤형 불용어 제거가 어려움 |
| C. User-defined stopword corpus | -사용자 정의 불용어 축적으로 재사용이 용이함<br>-맞춤형 불용어 제거로 정제 정밀도가 향상됨 | -불용어 코퍼스 작성 과정이 요구됨<br>-분석자의 불용어 선별역량이 요구됨 |

3. 워드클라우드 시각화 분석결과

실무구현 사례의 워드클라우드 시각화 결과를 분석해 보면 문재인 대통령은 취임사에서 “국민”, “역사”, “대화”, “선거”, “약속”, “갈등”, “평화”, “권력”과 같은 단어들이 빈도수가 높은 핫 이슈들로 분석되었고 그 이외에 빈도수는 높지 않으나 분석자가 휴리스틱으로 “남과 북”은 “공동번영”과 “평화공존”을 위해서 “신뢰”를 “바탕”으로 “상호 협력”해야 한다는 핵심가치를 해설하였으며, 윤석렬 대통령은 취임사에서 “자유”, “국민”, “세계”, “평화”, “국제”, “민주주의”, “경제”, “갈등”과 같은 단어들이 빈도수가 높은 핫 이슈들로 분석되었고 그 이외에 빈도수는 높지 않으나 분석자가 휴리스틱으로 “국민과 함께”, “자유민주주의”와 “시장경제”를 수호하기 위해서 “자유”, “시장”, “공정”을 추구해야 한다는 핵심가치를 해설 하였다. 따라서 가장 중요한 이슈는 휴리스틱 개입이 필요하기 때문에 분석자의 역량이 결과의 정밀도에 중요하다는 것을 알 수 있다[14][15].

## V. 결 론

본 연구에서는 제안된 모델의 기법으로 “문재인 대통령 취임사”와 “윤석렬 대통령 취임사”의 텍스트 문서를 빅데이터 분석 툴인 R의 워드클라우드를 사용하여 전 현직 대통령의 정책에 대한 핫 이슈를 도출하는 과정상의 문제점과 효용성을 검정하였다. 도출된 문제점들은 첫째, R의 한글사전(KoNLP) 사용을 위한 환경설정이 매우 어렵고 둘째, 한글사전(세종한글사전)에 없는 단어들이 다수 존재하고 셋째, 분석자의 휴리스틱 정제 역량이 부족하고 넷째, 숙련된 분석가가 아닌 경우 해석결과의 정밀도가 저하되고 다섯째, 불용어에 대한 효율적인 제거가 어렵고 불용어의 축적을 통한 재사용이 어렵다는 것이다. 따라서 제안된 “사용자 정의 불용어 코퍼스”를 이용한 머시인 후처리 정제기법은 다섯 번째의 문제점을 보완할 수 있는 불용어 관리와 등록이 용이하여 축적된 전문 불용어 코퍼스 생성을 가능하게 되어 도출된 문제점들을 해소하는데 유용한 것으로 평가되었다. 이 방법이 분석자가 개별적으로 전문적인 불용어 코퍼스의 생성이 가능하여 비정형 텍스트 데이터의 워드클라우드 분석을 위한 불용어 제거를 용이하게 하여 시각화 결과분석의 정밀도 향상에 기여할 것으로 판단된다.

향후 연구과제는 전문 영역별로 축적된 외부 불용어 코퍼스의 작성과 한글사전에 존재하지 않는 단어들의 등록을 위한 외부 코퍼스 작성과 활용에 대한 방법을 개발하여 워드클라우드 시각화 결과의 정밀도 향상 대한 연구가 진행되어야 할 것으로 사료된다.

## References

[1] W. Lee, A Study on Data Cleansing Techniques for Word Cloud Analysis of Text Data, JCCT, vol. 7, No. 4, pp. 745-750, 2021.
[2] W. Lee, A Study on Word Cloud Techniques for Analysis of Unstructured Text Data, JCCT, vol. 6, No. 3, pp. 337-341, 2020.
[3] J. Lee, D. Yun, S. O, C. Lee, A Big Data Analysis of Civel Complaint Texts Using R Language, KIICE, 2020.
[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 southeastem Airport Using Text Mining Techniques, Smart Media Journal, Vol. 6, No. 1, 2017.
[6] Giseop Noh, An Analysis on Internet Information using Real Time Search Words, JCCT, vol. 4, No. 4, pp. 337-341, 2018.
[7] I. Chun, D. Park, Y. Kang, Python and data science, Saengneun Publishing, pp. 222-233, 2019.

[8] M. Chi , S. Lin, S. Chen, C. Lin, T. Lee, Morphable word Clouds for Time-Varying Text Data Visualization, IEEE, 2015.
[9] M. Han, Y. Kim, C. Lee, Analysis of News Regarding New southeastem Airport Using Text Mining Techniques, Smart Media Journal, Vol. 6, No. 1, 2017.
[10]Jong Suk Lee and 3 others, Big data analysis of civil complaint texts using R language, 2020.
[11]Insun Lee and 1 others, Unstructured data analysis and visualization, Korean Psychology Association, 2018.
[12]Jongyong LEE, A Study on Tourism Analysis in Uijeongbu Region Using Big Data, JCCT, vol. 6, No. 1, pp. 413-419, 2020.
[13]Sunghuk Moon, Big data environment analysis and research on ways to secure global competitiveness, JCCT, vol. 5 No. 2, pp. 361-367
[14]Web Mining, IT Glossary, Korea Information and Communication Technology Association
[15]text mining, Biochemistry Encyclopedia
[16]Sejong Oh, R data analysis for everyone, R data analysis for everyone, Hanbit Media, 2019.
[17]https://wikidocs.net/22530.