The Journal of the Convergence on Culture Technology (JCCT)
Vol. 7, No. 4, pp.745-750, November 30, 2021. pISSN 2384-0358, eISSN 2384-0366

http://dx.doi.org/10.17703/JCCT.2021.7.4.745

JCCT 2021-11-91

# 텍스트 데이터 워드클라우드 분석을 위한 데이터 정제기법에 관한 연구

# A Study on Data Cleansing Techniques for Word Cloud Analysis of Text Data

이원조*

**Won-Jo Lee***

**요 약** 비정형 텍스트 데이터의 빅데이터 시각화 분석에서 원시 데이터는 대부분 대용량이고 비정형으로 정제하지 않고 분석기법을 적용할 수 없는 상태이다. 따라서 수집된 원시 데이터는 1차 휴리스틱 정제과정을 통해서 불필요한 데이터들을 제거하고 2차 머시인 정제과정을 통해서 불용어를 제거한다. 그리고 어휘의 빈도수를 계산하여 워드클라우드 기법으로 시각화하고 핵심 이슈들을 추출하여 정보화하고 그 결과를 분석한다. 본 연구에서는 파이썬 워드클라우드에서 외부 불용어 Set(DB)를 사용한 새로운 불용어 정제기법을 제안하고 실무 사례분석을 통하여 이 기법의 문제점과 효용성을 도출한다. 그리고 이 검증 결과를 통해 제안된 정제기법을 적용한 워드클라우드 분석의 실무적용에 대한 효용성을 제시한다.

**주요어** : 빅 데이터, 텍스트 분석, 워드클라우드, 파이썬, 중지어, 시각화, 데이터 정제

**Abstract** In Big data visualization analysis of unstructured text data, raw data is mostly large-capacity, and analysis techniques cannot be applied without cleansing it unstructured. Therefore, from the collected raw data, unnecessary data is removed through the first heuristic cleansing process and Stopwords are removed through the second machine cleansing process. Then, the frequency of the vocabulary is calculated, visualized using the word cloud technique, and key issues are extracted and informationalized, and the results are analyzed. In this study, we propose a new Stopword cleansing technique using an external Stopword set (DB) in Python word cloud, and derive the problems and effectiveness of this technique through practical case analysis. And, through this verification result, the utility of the practical application of word cloud analysis applying the proposed cleansing technique is presented.

**Key words :** Big Data, Text Analysis, Word Cloud, Python, Stop Words, Visualization, Data Cleansing

## Ⅰ. 서 론

최근 기술의 발전 추세는 상승나선(Positive Spiral) 효과와 같이 더욱 가속화되고 있다. 따라서 이러한 기술의 변화에 대비하기 위한 정보를 선점하기 위해서 빅데이터 분석에 대한 연구가 활발하게 진행되고 있다. 그러나 빅데이터 분석을 위한 대부분의 원시 데이터는 대용량이고 비정형 데이터 형태로 존재하기 때문에 분석에는

*정회원, 울산과학대학교 산업경영공학과 부교수 (제1저자)
(울산대 전자계산학과 공학박사/울산과학대학교 IT응용기술학부 20년/현재 산업경영공학과 부교수)
접수일: 2021년 9월 30일, 수정완료일: 2021년 10월 10일
게재확정일: 2021년 10월 18일

Received: September 30, 2021 / Revised: October 10, 2021
Accepted: October 18, 2021
*Corresponding Author: wjlee@uc.ac.kr
Dept. of Industrial Management Eng., Ulsan College, Korea

적합하지 않다. 그래서 효율적인 데이터 유형변환을 위한 데이터 정제기법에 대한 기업이나 개인들의 관심이 높아지고 있다. 그러나 정제기법이 정립되어 있지 않아 정제과정이 어렵고 전문적인 지식을 요구하기 때문에 시각화 분석결과에 대한 신뢰도가 낮다. 그래서 대용량 비정형 데이터의 시각화 분석결과의 신뢰도를 높이기 위해서는 비정형 데이터 정제기법의 표준정립과 워드클라우드 시각화 분석결과 해석을 위한 표준절차의 정립이 요구된다[1][2][3].

본 연구에서는 파이썬 프로그램의 워드클라우드 기법을 이용하는 비정형 텍스트 데이터 분석에서 기존의 불용어 Set(DB)에 존재하지 않는 불용어 들을 제거하기 위해서는 임의로 프로그램 소스코드에 불용어를 추가하는 방식을 사용하고 있다. 그러나 이러한 방식은 제거 대상 불용어가 많을 경우 적용이 어렵고 불용어 수집을 통한 축적과 재사용이 어렵다. 그래서 별도의 외부 불용어 Set(DB)으로 분리하여 후처리 정제기법에 활용하는 방식을 제안한다. 따라서 제안된 정제기법에 대한 효용성 검증을 위해서 파이썬 프로그램의 텍스트 데이터 시각화 기법인 워드클라우드를 사용하여 미국의 “트럼프 대통령의 취임사”와 “바이든 대통령의 취임사” 연설문을 빅데이터 워드클라우드 분석에 적용하고 시각화 결과를 통해서 미국의 국가전략에 대한 핫 이슈(Hot Issue)를 도출한다. 그리고 사례분석의 결과를 통해서 제안된 정제기법의 효용성과 문제점을 검증하고 실무적용 방안을 제시한다[1][2].

## II. 관련 연구

### 1. 비정형 텍스트 데이터 분석

오늘날 우리는 정보의 홍수시대에 살고 있다고 한다. 특히 인터넷의 발달로 뉴스, SNS, E-mail, 쇼핑몰, 유튜브 등을 통한 정보의 유통은 다양한 형식의 원시 데이터로 존재하고 유통되고 수집된다. 이들 중 사람이 읽을 수 있는 텍스트 형태로 유통되는 데이터가 가장 많고 중요한 정보교환의 수단으로 사용되고 있다. 그러나 이들의 대부분은 대용량이고 비정형 형태로 존재하기 때문에 인간에 의한 수작업 정제와 머시인 정제과정을 거쳐 정형화 하고 분석에 적합한 형식으로 가공하여 정보로 저장되어야 분석이 가능하다. 또한 비정형 텍스트 데이터 분석에서 인터넷 검색어 분석은 검색어를 중심으로 사용자들의 관심사를 분석할 수 있도록 지원해 주는 인터넷 사이트 들이 많이 존재한다. 그래서 이들 중에서 대표적인 것은 구글 트렌드와 네이버 데이터 랩이 있는데 구글 트렌드에서는 전 세계적인 관심사를 분석해 볼 수 있고 네이버 데이터 랩에서는 국내의 주요 관심사를 분석해 볼 수 있다. 그러나 본 연구에서와 같은 연설문 형식의 비정형 텍스트 데이터 분석에는 적합하지 않다[1][2].

### 2. 워드클라우드(word cloud)

워드클라우드는 각 단어의 크기가 빈도 또는 중요성을 나타내는 텍스트 데이트 시각화 기술이다. 즉 워드클라우드는 문서의 키워드, 개념 등을 직관적으로 파악할 수 있도록 핵심 단어를 시각적으로 돋보이게 하는 기법이다. 예를 들면 많이 언급될수록 단어를 크게 표현해 한눈에 들어올 수 있게 하는 기법 등이 있다. 주로 방대한 양의 정보를 다루는 빅데이터(Big data)를 분석할 때 데이터의 특징을 도출하기 위해 사용한다[1].

### 3. 텍스트 데이터 워드클라우드 분석

파이썬의 워드클라우드는 비정형 텍스트 데이터를 분석하는 기법으로, 사전에 휴리스틱(Heuristics)에 의해 전처리 정제된 데이터를 파이썬 프로그램으로 어휘들을 추출하고 출현 빈도수를 계산한 다음 머시인 후처리 과정에서 불용어를 제거하고 워드클라우드 기법을 사용하여 시각화하여 분석하는 기법이다. 워드클라우드 분석의 결과는 출현 빈도수가 높은 단어는 크게 표시되고 각 각의 단어들은 색상으로 구분 표시된다. 여기서 출현 빈도수가 높은 단어는 일반적으로 중요도가 높고 관심도가 높은 것으로 전제한다. 다음은 빈도수 유형에 따른 단어 수작업 전처리 규정이다. 이 규정은 연구자의 이전 논문에서 정립된 것으로 이 규정을 적용한다. 어휘의 출현 빈도수 계산결과 처리 다음과 같은 5가지 유형으로 구분할 수 있다.

가. 빈도수가 높고 중요성이 낮은 단어
나. 빈도수가 높고 중요성이 높은 단어
다. 빈도수도 낮고 중요성이 낮은 단어
라. 빈도수는 낮고 중요성이 높은 단어
마. 빈도수는 높으나 부적격 값

상기 파이썬 프로그램 처리에서 추출된 5가지의 유형에서 가, 다, 마항은 제거하고 나, 라항은 수렴하고

기존의 불용어 모듈(STOPWORD) Set(DB)에 존재하지 않은 불용어는 별도의 외부 불용어 Set(DB)을 생성하여 합집합(UNION)으로 통합하여 적용하고 후처리 정제를 수행한다[1][2].

## III. 비정형 텍스트 데이터 분석

### 1. 텍스트 데이터 분석 모델

본 연구에서 그림 1은 기존의 비정형 데이터 정제 과정도인데 기존 모델은 STOPWORD Set(DB)에 존재하지 않는 불용어의 제거를 위해서는 임의로 프로그램 소스코드에 불용어 들을 수작업으로 소스코드 내 추가 방식을 사용한다. 그리고 프로그램의 실행을 반복하면서 불용어들을 휴리스틱(Heuristics)으로 찾아 추가하는 후처리 정제과정을 반복하면서 시각화 명료도를 향상시킨다. 그러나 소스 코드에 불용어를 추가하는 과정의 반복은 번거롭고 제거 대상 불용어가 많을 경우 적용이 어렵고 불용어 축적을 통한 재사용이 어려운 단점이 있다. 따라서 이러한 문제점을 해결하기 위해서 그림 2 제안된 비정형 데이터 정제 과정도와 같이 별도의 외부 불용어 Set(DB)을 만들어 후처리 정제과정에 적용하는 기법을 제안한다. 이 기법이 불용어 추가 적용의 문제점 해소와 불용어 축적을 통한 재 사용성을 향상시켜 줄 것으로 기대된다[3].

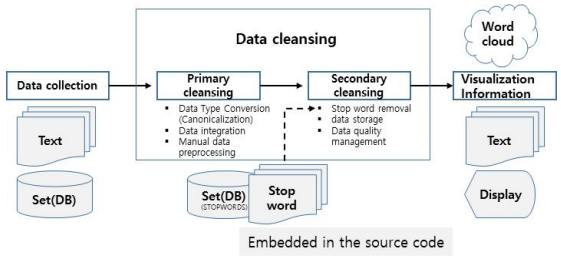


그림 1. 기존의 비정형 데이터 정제 과정도
Figure 1. Existing unstructured data cleansing process diagram

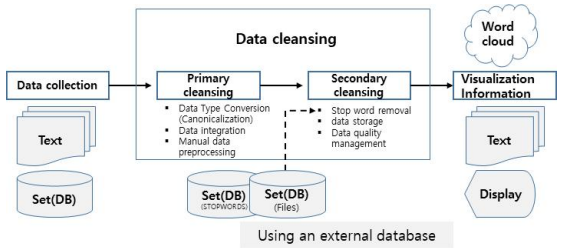


그림 2. 제안된 비정형 데이터 정제 과정도
Figure 2. Proposed unstructured data cleansing process diagram

### 2. 분석용 텍스트 데이터 수집

본 연구에서 사용할 문서는 백악관의 홈페이지에서 미국의 45대 대통령인 트럼프와 46대 대통령인 바이든의 "취임사 연설문" 원문을 수집하여 준비한다. 그리고 분석에 사용될 부분만 1차 전처리 정제하여 텍스트 파일형식으로 저장한다.

### 3. 파이썬 텍스트 데이터 정제기법

가. 워드클라우드 사용 환경설정

파이썬에서 워드클라우드를 생성 하려면 matplotlib, pandas, wordclud 등의 모듈이 필요하다[3].

-pip install matplotlib pandas wordcloud

나. 분석대상 텍스트 데이터 생성

워드클라우드 이미지 생성을 하기 위해서는 WordCloud 클래스를 사용하게 되는데, 다음과 같이 생성될 이미지의 크기를 가로, 세로 값을 클래스의 생성자에 넘겨주어 이미지의 크기를 설정하고, generate() 함수를 불러 워드클라우드 분석대상 텍스트 데이터를 인자로 넘겨주어야 한다. 다음 그림 3은 워드클라우드 분석대상 텍스트 데이터를 읽어서 generate 함수로 인자를 넘겨주는 소스 코드이다[3].

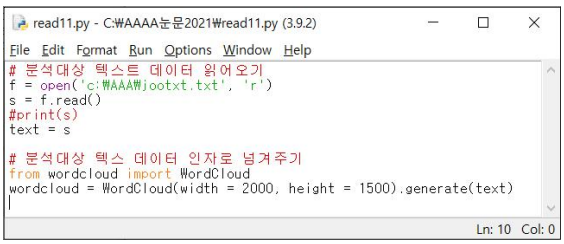


그림 3. generate 함수로 인자로 넘겨주는 실행코드
Figure 3. Executable code passed as an argument to generate function

다. 분석대상 텍스트 데이터 시각화

맷플로립(matplotlib)의 이미지 그리기 함수인 imshow() 함수를 이용하면 쉽게 표현이 가능한데, 맷플로립은 파이썬 프로그래밍을 통해서 데이터를 그래프로 시각화하는 유용한 라이브러리(Data visualization library)이다. 다음 그림 4는 맷플로립 라이브러리를 사용하여 이미지를 생성하는 소스 코드이다.

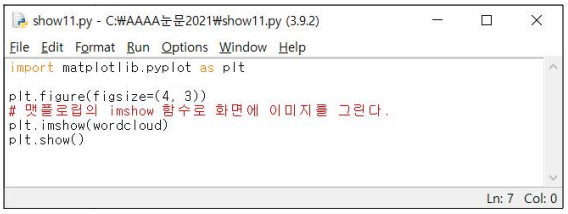


그림 4. 맷플로립을 사용한 이미지 생성을 위한 실행코드
Figure 4. Executable code for image creation using matflolip

라. 불용어(Stopwords) 전처리

불용어는 자연어 처리에서는 중지어라고도 한다. 텍스트 데이터에서 높은 빈도로 사용되지만 중요한 의미를 갖지 않는 어휘들을 말한다. 이러한 불용어를 파이썬 프로그램으로 쉽게 제거하는 방법이 있다. 이들 불용어 리스트는 wordcloud 모듈의 STOPWORDS에 Set(DB)에 데이터로 저장되어 있다. 따라서 이것을 이용하고, 따라서 새로운 불용어를 추가하고 싶다면 새로운 Set(DB)와 합집합으로 만들어 적용하면 불용어 제거가 가능하다. 다음 그림 5는 수작업 불용어 집합 생성 방법이고, 그림 6은 별도의 외부 불용어 Set(DB)의 이용 방법이다. 그리고 그림 7은 외부 불용어 Set(DB)이다 [3].

1) 수작업 불용어 집합 생성 방법은 그림 5와 같이 STOPWORD에 포함되어 있지 않는 어휘는 파이썬 소스 코드에 추가해야 하는 등의 적용에 어려움이 많은 단점이 있다.

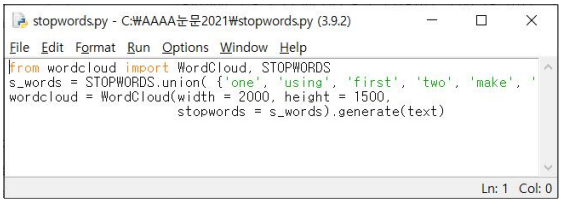


그림 5 수작업 불용어 집합 생성방법
Figure 5. How to manually create a set of Stopwords

2) 외부 불용어 Set(DB)의 이용 방법은 그림 7과 같이 STOPWORD Set(DB)에 포함되어 있지 않는 어휘는 별도의 외부 불용어 Set(DB)에 추가하면 되기 때문에 변경(추가/삭제)이 용이하고 분석자의 임의관리가 매우 쉽고 또한 불용어 어휘 축적을 통한 재사용이 가능한 장점이 있다.

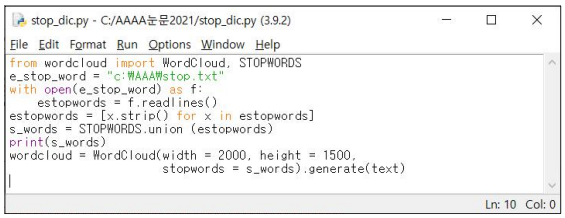


그림 6. 외부 불용어 텍스트 사전 이용 방법
Figure 6. How to use an external Stopword text dictionary

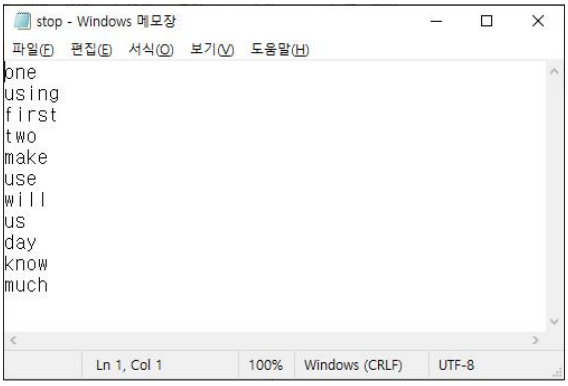


그림 7. 외부 불용어 Set(DB)
Figure 7. External Stopword Set(DB).

## 4. 워드클라우드 시각화 사례 구현

본 연구에서의 실무사례 구현은 미국의 트럼프와 바이든 대통령의 취임사를 빅데이터 워드클라우드 기법을 사용하여 사례분석을 하여 미국의 대 내외정책의 기조변화에 대한 핫 이슈 변화를 비교분석 한다. 그리고 제안된 외부 불용어 Set(DB)의 사용에 대한 문제점과 불용어 축적을 통한 재사용성의 유용성을 검증한다.

가. 외부 불용어 Set(DB) 적용 전 시각화

다음 그림 8은 트럼프 대통령의 취임사 연설문 데이터를 워드클라우드 기법으로 외부 불용어 Set(DB) 적용 전 시각화한 것이고, 그림 9는 바이든 대통령의 취임사 연설문 데이터를 워드클라우드 기법으로 외부 불용어 Set(DB) 적용 전 시각화한 것이다. 각각 불용어들이 핫 이슈로 전면에 표시되어 핵심 이슈의 추출을 통한 결과분석이 어렵거나 불가능하다.

나. 외부 불용어 Set(DB) 적용 후 시각화

그림10은 트럼프 대통령의 취임사 연설문 데이터를 워드클라우드 기법으로 시각화한 것이고, 그림11은 바이든 대통령의 취임사 연설문 데이터를 워드클라우드 기법으로 시각화한 것이다. 각 각 불용어 들이 제거되고

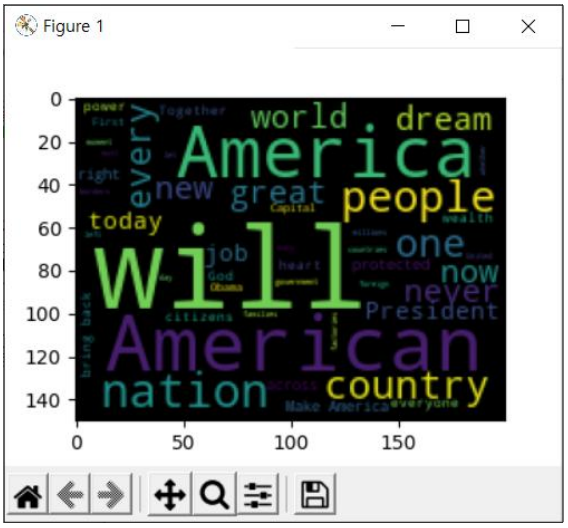


그림 8 트럼프 대통령의 취임사 연설문 시각화 결과(적용 전)
Figure 8. Visualization of President Trump's Inaugural Speech (Before application)

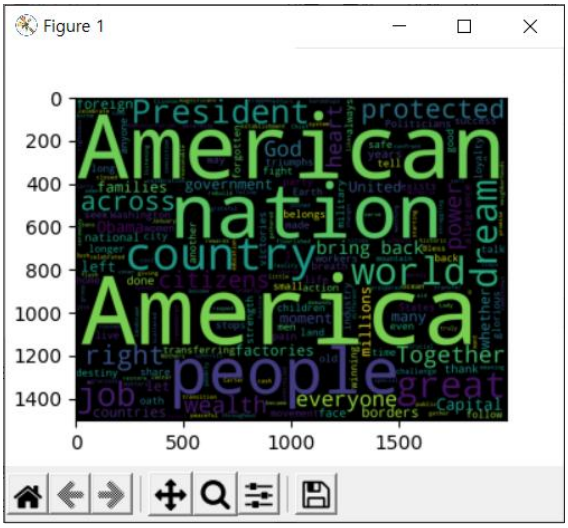


그림 10. 트럼프 대통령의 취임사 연설문 시각화 결과(적용 후)
Figure 10. Visualization of President Trump's Inaugural Speech (After application)

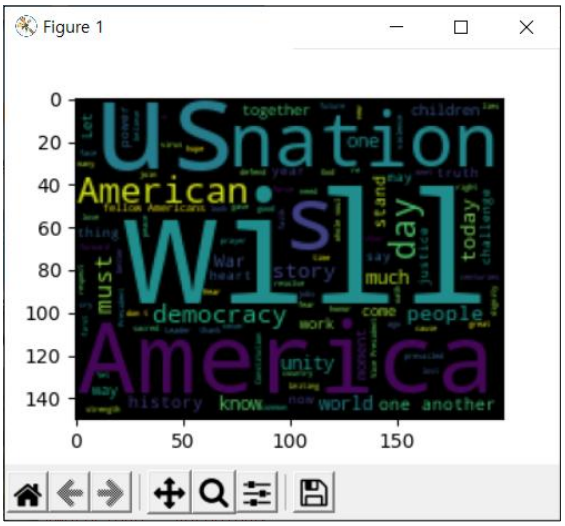


그림 9. 바이든 대통령의 취임사 연설문 시각화 결과(적용 전)
Figure 9. Visualization of President Biden's Inaugural Speech (Before application)

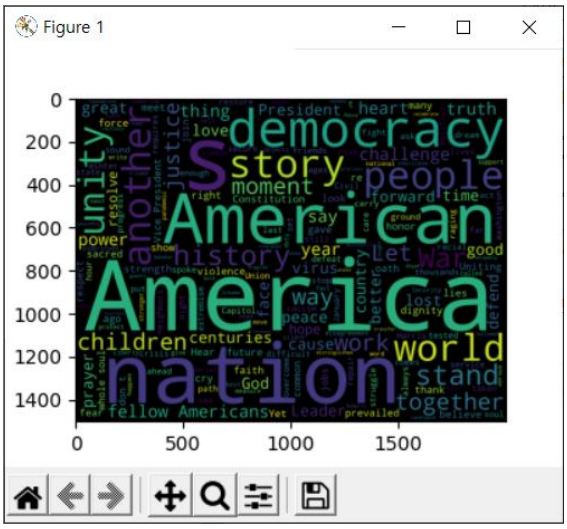


그림 11. 바이든 대통령의 취임사 연설문 시각화 결과(적용 후)
Figure 11. Visualization results of President Biden's inaugural speech(After application)

핵심 이슈가 전면에 표시되어 시각화 분석에 대한 신뢰도가 향상된 것을 볼 수 있다.

## 5. 워드클라우드 시각화 결과 분석

결과의 그림10과 그림11은 외부 불용어 Set(DB) 적용 후의 워드클라우드 시각화 결과이다. 이를 통해 시각화 결과를 분석하면 그림10은 "트럼프 대통령의 취임사 연설문"에서 추출된 빈도수가 가장 높은 단어들이고 주요한 핫 이슈는 America, American, nation, country, job, President, people, world, dream, great, protected, wealth, heart, everyone, power, God, citizens, across, right, etc. 이다. 따라서 워드클라우드 분석에서 추출된 핵심 이슈들을 정보화하여 정책으로 재해석하면 "Make America Great Again"으로 해설된다. 그리고 그림11은 "바이든 대통령의 취임사 연설문"에서 출현 빈도수가 가장 높은 단어들이고 주요한 핫 이슈는 America, American, nation, story, democracy, unity, another, world, people, stand, together, children, work, War, moment, history, etc. 이다. 워드클라우드 분석에서 추출된 핵심 이슈를 정보화하여 정책으로 재해석하면 "Bring America Together"으로 해설된다. 그리고 트럼프 대통령과 바이든 대통령의 취임사 연설문의 핵심 이슈는

America, American, nation이 공통으로 빈도수가 높았으나, 트럼프 대통령은 job(일자리 창출), great(위대한 미국)가 다음으로 유의미한 이슈로 표시 되었고 바이든 대통령은 democracy(민주주의 수호), unity(국민통합), together(다함께 하나로)가 다음으로 유의미한 이슈로 표시 되었다. 그리고 America, American, nation 과 같이 출현 빈도수는 공통적으로 매우 높으나, 핫 이슈의 의미가 적은 단어들도 불용어 대상으로 제거해야 핵심 주제어의 추출에 대한 신뢰도가 향상됨을 알 수 있다.

## IV. 결 론

본 연구의 사례구현은 파이썬 프로그램 워드클라우드에서 외부 불용어 Set(DB)을 사용한 새로운 불용어 정제기법을 적용하였다. 이 과정에서 도출된 문제점은 첫째 불용어 STOPWORD Set(DB)에 존재하지 않는 불용어 여부에 대한 판정은 여전히 휴리스틱(Heuristics)에 의해 결정해야 하고, 둘째 분석대상 정보를 취급하는 분석자들은 일반적으로 파이썬 프로그래밍이 대체로 미숙하며, 여전히 분석역량이 부족하고, 셋째 시각화 결과의 해석역량에 따라 결과의 신뢰도가 저하 될 것으로 판단되고, 넷째 불용어가 많을 경우 반영이 어렵고 축적하여 재사용 할 수 없다. 따라서 제안된 정제기법의 효용성은 네 번째 문제점이 보완된 불용어의 반영이 쉽고 외부 불용어 Set(DB)에 축적된 어휘의 재사용이 가능하여 분석결과의 신뢰성이 향상된 것을 알 수 있다. 또한 제안된 불용어 데이터 정제기법이 실무사례 적용결과 분석정보의 신뢰도 향상에 유용한 것으로 평가되었고 이 방안이 워드클라우드 분석의 표준방식 적용지침으로써의 의미를 갖는다.

향후 연구과제는 불용어 STOPWORD Set(DB)에 존재하지 않는 어휘의 외부 불용어 Set(DB) 전용 시소러스(Thesaurus)의 개발과 연구 결과의 문제점들을 보완한 비정형 텍스트 워드클라우드 분석을 보다 더 쉽고 분석결과의 신뢰도를 향상 시킬 수 있는 표준절차의 개발에 대한 연구가 진행되어야 할 것으로 사료된다.

## References

[1] W. Lee, A Study on Word Cloud Techniques for Analysis of Unstructured Text Data, JCCT, vol. 6, No. 3, pp. 337-341, 2021.
[2] J. Lee, D. Yun, S. O, C. Lee, A Big Data Analysis of Civel Complaint Texts Using R Language, KIICE, 2020.
[3] I. Chun, D. Park, Y. Kang, Python and data science, Saengneun Publishing, pp. 222-233, 2019.
[4] M. Chi, S. Lin, S. Chen, C. Lin, T. Lee, Morphable word Clouds for Time-Varying Text Data Visualization, IEEE, 2015.
[5] Kumar, P. Thakur, K. Gupta, and A. Pal, 2015, Text mining approach to analyse the relation between obesity and breast cancer data, ILNS
[6] 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.
[7] Jong Suk Lee and 3 others, Big data analysis of civil complaint texts using R language, 2020.
[8] Insun Lee and 1 others, Unstructured data analysis and visualization, Korean Psychology Association, 2018.
[9] Dongnyeok Sim, Research on ICT issue detection and analysis methodology using text data, 2020.
[10]Software Engineering Center Webzine Materials, Big data purification process, 2019.
[11]Giseop Noh, An Analysis on Internet Information using Real Time Search Words, JCCT, vol. 4, No. 4, pp. 337-341, 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]Dictionary of current affairs.