J. of Practical Agriculture & Fisheries Research ISSN 2234-1129
2021 June. : 23(2): 71-86

웹 크롤링에 의한 네이버 뉴스에서의 한국농수산대학
- 키워드 분석과 의미연결망분석 -

Korea National College of Agriculture and Fisheries in Naver News by Web Crolling : Based on Keyword Analysis and Semantic Network Analysis

주진수
J. S. Joo
국립한국농수산대학 [1]
농어업·농어촌연구소
nongsusan@af.ac.kr

이소영
S. Y. Lee
국립한국농수산대학 [1]
농수산비지니스학과
lsy2000@korea.kr

김승희
S. H. Kim
국립한국농수산대학 [1]
과수학과
vitis@korea.kr

박노복 *
N. B. Park*
국립한국농수산대학 [1]
화훼학과
noubogpark@naver.com

Abstract

This study was conducted to find information on the university's image from words related to 'Korea National College of Agriculture and Fisheries (KNCAF)' in Naver News. For this purpose, word frequency analysis, TF-IDF evaluation and semantic network analysis were performed using web crawling technology.

In word frequency analysis, 'agriculture', 'education', 'support', 'farmer', 'youth', 'university', 'business', 'rural', 'CEO' were important words. In the TF-IDF evaluation, the key words were 'farmer', 'dron', 'agricultural and livestock food department', 'Jeonbuk', 'young farmer', 'agriculture', 'Chonju', 'university', 'device', 'spreading'.

In the semantic network analysis, the Bigrams showed high correlations in the order of 'youth' - 'farmer', 'digital' - 'agriculture', 'farming' - 'settlement', 'agriculture' - 'rural', 'digital' - 'turnover'.

As a result of evaluating the importance of keywords as five central index, 'agriculture' ranked first. And the keywords in the second place of the centrality index were 'farmers' (Cc, Cb), 'education' (Cd, Cp) and 'future' (Ce). The sperman's rank correlation coefficient by centrality index showed the most similar rank between Degree centrality and Pagerank centrality.

The KNCAF articles of Naver News were used as important words such as 'agriculture', 'education', 'support', 'farmer', 'youth' in terms of word frequency. However, in the evaluation including document frequency, the words such as 'farmer', 'dron', 'Ministry of Agriculture, Food and Rural Affairs', 'Jeonbuk', and 'young farmers' were found to be key words. The centrality analysis considering the network connectivity between words was suitable for evaluation by Cd and Cp. And the words with strong centrality were 'agriculture', 'education', 'future', 'farmer', 'digital', 'support', 'utilization'.

Key words : Web crawling, Semantic network analysis, Pagerank centrality(Cp), Degree centrality(Cd), Eigenvector centrality(Ce), Sperman's rank correlation coefficient

*교신저자
1 Korea National College of Agriculture and Fisheries

Ⅰ. 서론

우리는 인터넷만 접속하면 무궁무진한 정보들과 만날 수 있는 정보의 홍수 속에 살고 있다. 하지만 사람의 능력은 한계가 있어서 이러한 가공되지 않은 정보와 많은 양의 데이터에서 의미 있는 정보를 찾는 것은 많은 시간과 노력이 필요하다.

디지털 정보 바다 속의 가치 '빅데이터'의 특징[1]은 데이터의 '양(Volume)', 데이터 생성 '속도(Velocity)' 그리고 정형, 반 정형, 비정형 데이터 등 형태의 '다양성(Variety)' 등 3V로 요약하는 것이 일반적이었다. 그러나 근래에는 여기에 빅데이터의 새로운 속성[2]으로 데이터의 신뢰성에 대한 정확성(Veracity), 데이터가 맥락에 따라 의미가 달라진다는 가변성(Variability), 사용 대상자의 이해를 높여주는 시각화(Visualization) 등이 추가되었다.

빅데이터 관련 기술은 데이터를 수집·저장하는 데이터 '처리 기술'과 데이터를 분석·시각화하는 데이터 '분석 기술'로 구성된다. 이 가운데 데이터 분석 기술의 발달은 기존 데이터 분석에서는 불가능했던 비선형적 상관관계 규명, 감성분석 등 비정형화된 분석까지도 가능하게 만들었다[3].

현대사회에서 빅데이터에 대한 지속적인 관심과 실험적인 시도 등의 활용 여건이 계속 개선되는 이유는 CPU, 저장장치, 메모리 등 저장매체의 가격 하락으로 정보 저장 및 처리 등 데이터 관리비용의 감소와 HADOOP 및 R 같은 오픈 소스 기술의 발달 등으로 데이터 처리·분석 기술의 발달에 있다. 또한 여러 기업에서 관리되지 않고 버려지는 데이터에 관심을 두고 사업적 가치를 발굴하여 실제 사업에 적용한 사례가 많이 등장했기 때문이다.

본 연구에서는 오픈 소스 R을 이용한 웹 크롤링 분석 기술을 바탕으로 네이버 뉴스에서 한국농수산대학(이하 한농대)을 주제어로 하는 관련 뉴스를 수집·분석하여 네이버 뉴스 내에서 한농대 이미지를 나타내는 단어 및 단어 간의 관계 등을 연구하였다.

Ⅱ. 연구내용

1. 웹 크롤링

크롤링의 대상은 바로 온라인상 데이터이며, 크롤링은 인터넷상의 웹페이지(html, 문서) 등에서 컴퓨터 프로그램을 이용해 자료를 수집해서 분류하고 저장하는 것을 뜻한다. 그러나 엄밀하게 말하면 많은 사람들이 인터넷에서 존재하는 데이터를 컴퓨터 프로그램을 통하여 자동화된 방법으로 웹에서 데이터를 수집하는 모든 작업을 의미하는 스크래핑과 혼용하여 사용하고 있다. 그러나 크롤링과 스크래핑은 기본적으로 서로 다른 개념이라 할 수 있으나 인터넷에서 프로그램을 이용해 자료를 추출하는 작업을 의미하는 공통 개념이라 할 수 있어 전문 개발자가 아니라면 정확하게 구분할 필요는 없다.

크롤링한 대용량 데이터는 엑셀 파일로 저장하고, 아래에서 설명하는 분석 도구(tool)들을 이용하여 수집한 데이터의 분석을 통하여 데이터에 담겨 있는 정보를 추출하였다.

1) https://terms.naver.com/entry.nhn?docId=3386304&cid=58370&categoryId=58370&expCategoryId=58370 네이버 지식백과. 빅데이터의 특징 참고
2) https://terms.naver.com/entry.nhn?docId=3386305&cid=58370&categoryId=58370&expCategoryId=58370 2018.12.19 빅데이터의 공통적 속성과 새로운 V 참고
3) 경제금융용어 700선

2. 키워드 빈도 및 중요도 분석

프로그램 R을 이용하여 크롤링한 한농대 관련 네이버 뉴스의 기사와 URL 등의 데이터를 엑셀 파일로 저장하고 키워드의 빈도 분석 및 중요도 분석 그리고 텍스트 마이닝을 통한 단어 의미연결망분석 등을 수행하였다.

먼저 키워드 빈도 분석은 특정 문서 집단 내에서 많이 언급되는 주제어를 추출하고 이들이 언급되는 빈도에 따라 중요도를 '단어 빈도(Term Frequency: TF)'로 나타냈다. 그러나 TF 값이 큰 단어일수록 중요도가 높다고 판단할 수 있지만 '문서 빈도(Document Frequency: DF)' 값이 큰 단어일 수 있기 때문에 키워드 중요도 분석은 DF 값의 역수인 IDF(역 문서빈도, inverse document frequency)에 단순 단어 빈도(TF) 값을 곱한 TF-IDF 가중치를 사용한다.

TF-IDF는 단순한 단어 빈도 처리가 아닌 단어의 출현 확률을 기준으로 출현 빈도를 재가공한 지표로 특정 문서 내에서 단어 빈도가 높을수록, 그리고 전체 문서들 중 그 단어를 포함한 문서가 적을수록 그 값이 크게 나타난다. 이러한 TF-IDF를 통해 기사 텍스트 데이터 전체에서 공통으로 혹은 무의미하게 반복적으로 나타나는 특정 단어를 제거할 수 있다.

3. 의미연결망분석

텍스트 마이닝은 TF와 TF-IDF 분석을 발전시킨 것으로 뉴스 기사 분석 분야에서 두드러지게 유용한 분석 방법이다. 텍스트 마이닝 분석의 한 종류인 '의미연결망분석(semantic network analysis)'을 통하여 기사 텍스트의 문맥에 따라 쟁점을 파악하고 텍스트 간 연계를 분석하였다.

의미연결망분석은 사회연결망분석 기법을 텍스트 내 단어의 관계에 적용한 것으로 의미연결망 분석에서는 일정한 범위 내에서 어휘가 동시에 등장하면 서로 연결된 것으로 간주하고 이 연결 관계들을 분석하였다.

의미연결망분석에 활용한 중심성 지표는 그래프상에서 어떤 Node가 가장 중요한지를 살피는 척도로서 지난 연구[4, 5, 6]에 사용한 근접 중심성(Closeness Centrality, Cc)[4]과 매개 중심성(Betweenness Centrality, Cb)[5] 평가지표와 더불어 다음 세 가지 평가지표를 추가하였다.

먼저 연결 중심성(Degree Centrality, Cd)은 가장 기본적이고 직관적으로 중심성을 측정하는 지표로서 각 node 별로 직접 연결된 edge 수를 나타내며 해당 Node가 가진 영향력을 확인할 수 있다. 그러나 단순히 연결된 Node의 숫자만으로 중요성을 평가하는 단점이 있다.

두 번째로 고유벡터 중심성(Eigenvector Centrality, Ce)은 연결된 상대 단어의 중요성을 반영해서 계산하는 방법으로서 주변 node의 중심성까지 고려하므로 중요한 단어와 많이 연결됐다면 고유벡터 중심성은 높아지게 된다.

마지막으로 페이지랭크 중심성(Pagerank Centrality, Cp)은 월드 와이드 웹과 같은 하이퍼링크 구조를 가지는 문서에 상대적 중요도에 따라 가중치를 부여하는 방법으로 지금까지 등장한 중심성 알고리즘 중 가장 성공한 알고리즘이라고 볼 수 있다.

Ⅲ. 분석 프로그래밍

1. URL 설정

URL은 인터넷상에서 '통신 규칙://인터넷 호스트 주소/경로 이름'으로 이루어진 웹 페이지의

4) 중요한 Node일수록 다른 Node까지 도달하는 경로가 짧을 것이라는 가정
5) Node들 간의 최단 경로를 가지고 계산

주소라 할 수 있다. 크롤링의 여기서부터 시작하며, 원하는 정보를 얻으려면 얻고 싶은 웹상의 내용(Contents)이 어디에 있는지 알아야 한다.

<Fig. 1>에 나타낸 URL은 본 연구에서 주제어로 설정한 '한국농수산대학'을 네이버에서 검색한 후 뉴스 카테고리로 들어간 상태의 URL이다. URL 내용은 'where=' 뒤에 '뉴스' 카테고리라는 것을 나타내며, 'query=' 뒤에 검색 키워드 '한국농수산대학'으로 구성되어 있다.

Fig. 1. URL of search results for KNCAF in Naver News

URL은 체계적인 규칙으로 만들어진 우리나라 도로명 주소처럼 일정한 규칙이 있다. '한국농수산대학'을 검색한 후 규칙 확인을 위해 검색한 창의 다음 페이지로 넘어가 보면 URL의 모든 구조는 똑같고 마지막 부분 'start=숫자' 에서 '숫자' 값만 변하는 것을 알 수 있다. 이런 구조적 특성 때문에 R에서 반복문을 이용하여 쉽게 모든 페이지의 URL을 가져올 수 있다. <Fig. 2>는 검색결과의 두 번째 페이지 URL로 검색 기간을 2020년 8월 21일부터 2021년 8월 20일로 설정한 조건이 나타나 있다.

```
https://search.naver.com/search.naver?where=news&sm=tab_pge&query=%ED%95%9C%EA%B5%AD%EB%86%8D%EC%88%98%EC%82%B0%EB%8C%80%ED%95%99&sort=0&photo=0&field=0&pd=3&ds=2020.08.21.&de=2021.08.20&cluster_rank=50&mynews=0&office_type=0&office_section_code=0&news_office_checked=&nso=so:r,p:from20200821to20210820,a:all&start=11
```

Fig. 2. URL on page 2 of search results for KNCAF in Naver News

2. HTML 파악

HTML(Hypertext Markup Language)은 웹에서 볼 수 있는 문서를 만들기 위한 일종의 표준 언어로 '태그(Tag)'를 통해 웹 브라우저로 보이는 모양을 나타내며 '트리(Tree) 구조'로 되어 있다. HTML에서의 트리 구조와 태그는 구글 크롬 기준 F12를 누르면 '소스 코드'를 확인할 수 있다.

'한농대'로 검색한 후 F12를 누르면 <Fig. 3> 처럼 오른쪽 창에 트리 구조를 갖는 HTML이 생겨난 것을 볼 수 있으며, 원하는 부분이 어떤 '태그'에 속하는지 알 수 있다. 이 창을 보면 들여쓰기가 되어 있는데, 들여쓰기가 되어 있는 부분들이 하나의 '트리 구조'를 형성하고 있다.

그리고 '<' 다음에 li, div, dl, dd, span, a 등 등 여러 가지 태그가 있다. 태그는 '<>'로 시작하고 '</>'로 반드시 끝나므로 '트리 구조'와 '태그'를 이용해서 원하는 정보를 추출할 수 있다.

R에서 HTML을 파악하고 필요한 부분을 추출

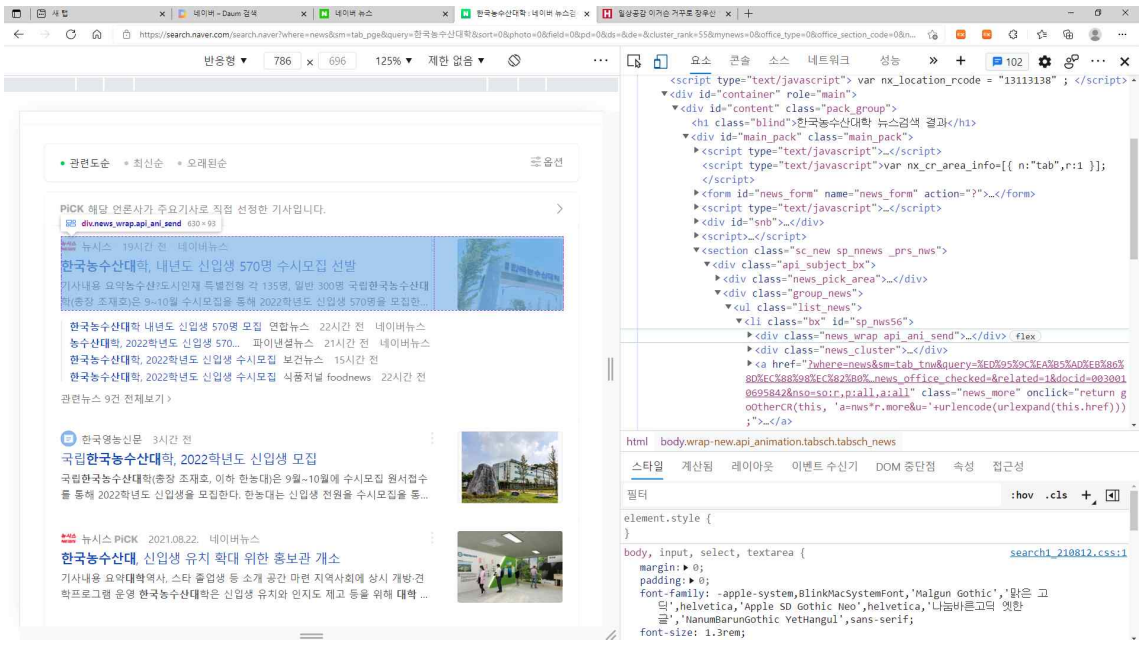


Fig. 3. Search results and developers' window for KNCAF in Naver News

하기 위해서는 패키지 'rvest'를 설치해야 한다. 또한 'tidyverse' 패키지를 함께 설치하여 함수들과 파이프라인(%>%)을 사용하도록 하였다.

3. 프로그래밍 순서

기본 URL 설정과 HTML 구조 파악을 완료한 후에 read_html()과 반복문 for()를 이용해서 검색 기간으로 설정한 검색결과의 페이지 2, 3, 4,... 등 전체 페이지에 대한 URL을 수집하였다. 본 연구의 HTML 트리구조를 보면 'ul.list_news>li>div>div>div>div>a.info' 순서로 구성되어 있으므로 <Fig. 4>와 같이 URL을 수집하였다. 여기서 html_nodes()는 원하는 Node를 뽑아올 때 사용하며, Node를 이동할 때에는 '>', class를 표현할 때에는 '.' 기호를 이용하였다. 또한 grep("naver", news_links)는 네이버 사이트만 추출하기 위한 설정이다. 검색 기간을 1년으로 설정한 '한국농수산대학' 관련 네이버 뉴스는 모두 318개가 수집되었다.

```
for (url in urls) {
  html <- read_html(url)
  news_links <- c(news_links, html %>%
                    html_nodes('ul.list_news>li>div>div>div>div>a.info') %>%
                    html_attr('href') )  #주소 차례로 더하기
  #                 노드를 이동할 때에는 '>' 기호를 사용하며 class 를 표현할 때에는 '.'
  news_links2 <- c(news_links2,news_links[grep("naver", news_links)] ) #naver 사이트만 추출
  news_links  <- NULL
}
```

Fig. 4. The code of R to get news URL in Naver News

뉴스 본문의 수집은 뉴스 기사 링크 중 하나를 복사해 들어가서 HTML 구조를 보면 div 태그에 뉴스 기사의 본문이 들어가 있음을 확인할 수 있었다(Fig. 5). 마찬가지로 read_html()을 이용해서 HTML에 연결하고, html_nodes()와 html_text() 함수로 태그 안에 있는 내용을 추출하였다. 추출한 결과는 csv 파일 형식으로 저장하였다. 기사 본문과 제목을 추출하기 위한 프로그램 소스를 <Fig. 6>에 나타냈다.

(a) Div tag of article title

(b) Div tag of article body

Fig. 5. HTML structure of article title and body

```
62 for (link in news_links2) {
63   html <- read_html(link)
64   #본문
65   Contents <- c(Contents, html %>%
66                   html_nodes('._article_body_contents') %>%
67                   html_text() )
68   #제목
69   Title <- c(Title, html %>%
70                html_nodes('#articleTitle') %>% #여기서 #은 id( <h3 id> )
71                html_text() )
72 }
```

Fig. 6. The code of R to get the text content and article title in Naver News

키워드의 빈도 분석 및 시각화를 위하여 read.csv()을 이용해서 파일을 불러온 후 편리한 분석을 위해 tidyverse 패키지를 설치하였다. 데이터는 table 형태로 만들고 str_remove() 함수를 이용해서 전처리 후 텍스트 파일 형식으로 저장하였다.

한편 키워드의 추출 함수는 SimplePose09()를 사용하여 9개의 품사로 단어를 구분하여 추출한 후 체언만을 활용한 분석을 하였다. KoNLP 패키지에 있는 SimplePose09()는 한국어 표현을 분해해서 이해하기 위한 함수로서 KIAST 품사 태그셋에 따라 텍스트에 품사를 붙일 수 있는 함수이다. <Fig. 7>에 나타낸 바와 같이 각각의 단어들이 품사 태그를 단 채로 추출하였다.

분석 단계에서는 검색 키워드가 '한국농수산대학'이었으므로 한농대 교명 관련 단어(기사에 9개

의 명칭 사용됨)를 포함하지 않았으며 분석에 불필요한 언론사 명칭 및 기자 이름 등도 제외하였다. 또한 '이상', '만원', '올해', '무단', 재배포, 무단전재 등 의미 없는 단어들은 모두 str_remove(), str_replace_all() 및 filter() 함수 등으로 분석에 포함하지 않았다. 분석에는 useNIADic() 사전을 사용하였으며, mergeUserDic() 함수를 이용하여 사전에 없는 단어를 추가하였다.

Name	Type	Value
박일호가운데시장이	character [1]	'박일호가운데시장/N+이/J'
국립한국농수산대학과	character [1]	'국립한국농수산대학/N+과/J'
농업인	character [1]	'농업인/N'
육성	character [1]	'육성/N'
지원을	character [1]	'지원/N+을/J'
위한	character [1]	'위하/P+ㄴ/E'
업무	character [1]	'업무/N'
협약을	character [1]	'협약/N+을/J'
체결했다	character [1]	'체결/N+하/X+었다/E'

Fig. 7. Part of speech separation of words using the SimplePose09() function, a KAIST tag set

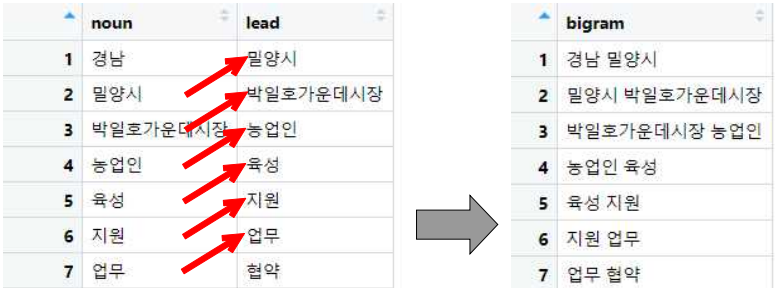
단어 정제 후 뉴스 기사에 많이 등장한 키워드 추출과 워드 클라우드를 작성하였으며, 문서 빈도를 기본으로 문서 내에서 핵심어를 구하기 위하여 tidytext 패키지의 bind_tf_idf() 함수를 이용하여 TF, IDF, TF-IDF 등을 구하였다. 또한 키워드 빈도 TF를 구한 후 단어들 사이에서 어떤 관계가 있는지 검토하기 위하여 의미연결망분석을 하였다.

의미연결망분석을 위해서는 모든 단어를 '뉴스 기사 순서대로' 불러와야 하는데 이는 의미연결망을 그린다는 것이 각각의 단어들 사이에 '연결고리'를 찾아 이어주어야 하기 때문이다. 그러나 단순한 단어들 속에서 '연결고리'를 찾기는 쉽지 않다. 그래서 연결고리를 만들어 주는 하나의 기준을 정해주는데 그것이 바로 '단어들이 같은 뉴스 기사 안에 존재하는가?' 그리고 '같은 문장 안에 존재하는가?'이다. 이 기준에 따라 연결고리를 찾아 의미연결망분석과 시각화를 하였다.

의미연결망분석에서는 'bigram' 개념을 활용하게 되는데, 이는 언어학에서 N-gram 이라는 개념으로 N에는 연속해서 등장하는 낱말의 개수를 말하는데, 2개의 단어가 연속해서 등장하는 경우를 의미한다. 의미연결망에 쓰일 데이터 bigram을 만들기 위하여 mutate(lead=lead(noun))로 코딩하여 dplyr 패키지의 lead() 함수로 noun의 바로 다음 행을 lead의 행으로 가져왔다. 그리고 unite() 함수를 이용해서 noun과 lead의 두 열을 합쳐서 bigram이라는 열을 만들어 의미연결망을 그리기 위한 데이터를 만들었다(Fig. 8).

불필요한 단어들을 삭제하기 위하여 bigram 열을 word 1, word 2 로 분리하여 word 1 과 word 2 중에서 하나의 열에 불필요한 단어가 속해 있다면 그 행 모두를 삭제(del)하였다(Fig. 9). 불필요한 단어의 판단은 except 변수(Fig. 10)에 단어들을 설정하고 case_when() 함수를 이용하였다.

의미연결망분석에는 centrality_closeness(), centrality_betweenness(), centrality_eigen(), centrality_pagerank(), centrality_degree() 등 다섯 개의 중심성 계산 함수를 이용하였다.



	noun	lead
1	경남	밀양시
2	밀양시	박일호가운데시장
3	박일호가운데시장	농업인
4	농업인	육성
5	육성	지원
6	지원	업무
7	업무	협약

	bigram
1	경남 밀양시
2	밀양시 박일호가운데시장
3	박일호가운데시장 농업인
4	농업인 육성
5	육성 지원
6	지원 업무
7	업무 협약

Fig. 8. Example of 'bigram', a semantic network data using a mutate() function

	word1	word2	n	except
28	유명	미술작품	19	ok
29	파종	살포	19	ok
30	농식품	분야	18	ok
31	스마트	농업	18	ok
32	제공	재판매	18	del
33	뉴스	뉴스	17	del
34	디지털	교육	17	ok
35	리얼타임	뉴스	17	del

Fig. 9. A refinement sample of a semantic network data using case_when() function

Table 1. Programming procedures and functions used in webcrolling

순 서	사용함수	내 용
① 기본 url 만들기	base_url <- 'https://.....	키워드 '한국농수산대학'
② 뉴스기사 url 뽑아오기	for (url in urls) { html <- read_html(url) n_links <- c(n_links, html %>% html_nodes() %>% html_attr() n_links }	반복문
③ 내용 및 타이틀 크롤링	for (link in n_links) { html <- read_html(link) 내용(타이틀)<-c(내용(타이틀), html %>% html_nodes() %>% html_text()) }	- 내용 html_nodes('._article_body_contents') - 타이틀 html_nodes('#articleTitle')
④ 저장, 불러오기, 정제하기	write.csv(), read.csv(), str_remove() write.Table(), readLines(), str_replace_all()	csv, text 파일
⑤ 단어 추출하기	SimplePose09() str_match(value,'([가-힣]+[a-zA-Z]+)/N')	9개의 품사 체언만 뽑아오기
⑥ 단어빈도 TF-IDF 계산	count() / wordcloud2 bind_tf_idf(pos_clean, doc_order, n)	시각화
⑦ 의미연결망분석	mutate(lead=lead(noun)) unite(bigram, c(noun, lead), sep = ' ') case_when(word1...., word2....) centrality_eigen() centrality_pagerank() centrality_degree()	bigram 만들기 단어 전처리 중심성 분석 시각화

```
except <- c('이하','제공','무단', '전재', '금지','속보','연예뉴스','스타','저작권자',
  '파이낸셜뉴스','뉴스','구독','매경','헤럴드','바로가기','사이트','재배포금',
  'Copyrightsc','프레시안','파이낸셜뉴스','서울경제','동아일보','한겨레','채널',
  '매일신문','전자신문','한글','서울신문','연합뉴스','헤럴드에코','밀리터리',
  '가정상담소줄리아', '투자노트','헤밀','서울연합뉴스',
  '헤럴드경제','매일경제','무궁무진한','배포금지','아시아경제','꿀잼')
#
```

Fig. 10. Variable used in case_when() function when refining the word colums of 'bigram'

Ⅳ. 결과 및 고찰

KoNLP에 포함된 SimplePose09() 함수로 한글 형태소를 분석하여 단어 빈도 분석, TF-IDF 분석 및 단어의미연결망분석을 수행한 결과는 다음과 같다.

분석에는 2020년 8월 21일부터 2021년 8월 20일로 설정한 기간 동안 '한국농수산대학' 관련 네이버 뉴스 318개 가운데 중복된 뉴스나 정부의 인사 관련 뉴스 등을 제외한 262개의 뉴스 기사를 사용하였다.

1. 빈도 분석

네이버 뉴스의 한농대 관련 기사에서 추출한 단어 가운데 출현 빈도 상위 80위까지를 <Table 2>에 나타내며, <Fig. 11>은 상위 100위까지 단어를 워드 클라우드로 나타낸 결과이다. 단어 빈도분석은 언급된 빈도에 따라 단어의 중요도가 높다고 판단하므로 네이버 뉴스에서 한농대와 관련이 많은 단어로는 '농업', '교육', '지원', '농업인', '청년', '대학', '사업', '농촌', '대표' 등이 자주 사용되는 것으로 나타났다. <Table 2>에 나타난 키워드를 보면 한농대는 농어업 분야의 청년 인재, 농수산 리더, 디지털 인재를 양성하는 대학으로서 졸업 후 취·창업에 안정적 지원을 최선으로 하며, 디지털 농업 전문 인재로서의 청년농업인을 육성하는 대학 특성을 잘 나타내고 있다고 할 수 있다.

Fig. 11. Word clouds of keywords related to KNCAF in Naver News

Table 2. The frequency and ranking of words related to KNCAF in Naver News (단위 : 회)

순위	단어	빈도	순위	단어	빈도	순위	단어	빈도	순위	단어	빈도
1	농업	420	21	농어업	129	41	예정	89	61	생각	69
2	교육	304	22	창업	129	42	확대	88	62	신입생	69
3	지원	281	23	개발	119	43	선정	87	63	일반전형	69
4	농업인	233	24	진행	118	44	전문	86	64	농수산업	68
5	청년	201	25	다양한	115	45	강화	84	65	실습	68
6	대학	195	26	운영	114	46	드론	83	66	새만금	67
7	사업	172	27	정착	110	47	농가	82	67	교육과정	66
8	농촌	171	28	발전	106	48	인력	82	68	규모	66
9	대표	168	29	육성	105	49	추진	80	69	상황	66
10	졸업생	159	30	양성	104	50	공공기관	77	70	딸기	65
11	계획	157	31	농장	101	51	전북	76	71	선발	65
12	기술	152	32	인재	100	52	교수	75	72	이상	65
13	영농	152	33	문제	99	53	스마트	75	73	구축	64
14	디지털	148	34	학생	97	54	예산	74	74	도움	64
15	대상	141	35	농촌진흥청	96	55	정부	74	75	아이디어	64
16	청년농업인	139	36	스마트팜	94	56	졸업	74	76	기관	63
17	현장	137	37	사람	92	57	중심	73	77	기존	63
18	미래	134	38	지역	92	58	재배	71	78	관리	62
19	분야	133	39	기반	91	59	학과	71	79	활동	62
20	농림축산식품부	130	40	관련	89	60	활용	71	80	조성	61

2. 중요도(TF-IDF) 분석

단어 빈도(TF) 및 역 문서빈도(IDF)를 이용하여 262개의 뉴스 기사에 나타난 각 단어마다 TF-IDF를 산출하여 문서 내에 나타난 단어들의 중요도를 검토하였다.

그 예로 일부 기사에 대한 키워드 '드론'의 TF, IDF 및 TF-IDF 값의 계산결과를 <Table 3>에 나타냈다. 뉴스 기사별 '드론' 빈도수(na)와 해당 기사의 전체 단어 빈도수(nb)가 각각 다르므로 기사별 TF값이 다르며, IDF값은 DF=29, N=262으로 일정하므로 모든 기사에 같은 값으로 계산하였다. TF-IDF 값은 TF와 IDF의 곱으로 계산하였다.

이와 같은 방법으로 산출한 결과 가운데 의미연결망분석 <Table 9>에서 단어 간 연결 중심성이 높게 나타난 단어와 '드론'의 6개 기사별 TF-IDF 값과 기사별 핵심어를 구하여 <Table 4>에 나타냈다.

Table 3. Examples of TF, IDF and TF-IDF calculations for keyword 'dron' in documents

문서 번호	단어	n^a	n^b	DF	TF	IDF	TF-IDF
78	드론	7	157		0.044586		0.099701
79	드론	4	102		0.039216		0.087692
80	드론	6	121	28	0.049587	2.236140	0.110883
81	드론	7	133		0.052632		0.117692
82	드론	6	135		0.044444		0.099384

n^a : 해당 문서에서의 키워드 '드론'의 빈도수, n^b : 해당 문서에서의 키워드 전체 빈도수
$TF = (n^a/n^b)$, $IDF = \ln(N/DF)$, N : 전체 문서수

Table 4. Examples of TF-IDF values by word and keywords in documents

문서 / 단어	Doc.3	Doc.5	……	Doc.190	Doc.193	……	Doc.261	Doc.262	합계
농업	0.025653	0		0.065851	0.018044		0.00447	0.001128	1.664374
청년	0.009687	0.021019		0.009946	0.00511		0	0	1.431605
농업인	0.070374	0.063624		0.009032	0.012375		0	0	1.925427
교육	0.007547	0.008187		0.001937	0		0.003945	0	1.420341
디지털	0	0		0.045541	0		0	0	0.977469
미래	0.011602	0		0.005956	0		0.012129	0.00153	0.863377
드론	0	0		0.004991	0		0	0.007693	1.911852
핵심어	협약	밀양시		농업	농촌살리기		유리온실	새만금	농업인

<Table 5>는 전체 262개의 기사에 나타난 각 단어의 TF-IDF 값을 합한 순위를 나타낸 결과로서 '농업인', '드론', '농림축산식품부', '전북', '청년농업인', '농업' 등의 단어가 한농대와 관련된 뉴스 기사에서 핵심어 역할을 하는 것으로 나타났다.

이 가운데 상위 10위까지를 앞에 제시한 <Table 2>의 단어 빈도와 비교하여 <Table 6>에 나타냈다. 단어 빈도에서 '드론', '농림축산식품부', '전북', '청년농업인', '전주', '장치, '파종' 등의 단어 순위는 낮았으나 TF-IDF 순위에서는 10위 안에 나타나는 중요 핵심어로 평가되었다. 이 결과는 <Table 2>에서 10위 안에 나타난 '교육', '지원', '청년', '사업', '농촌' 등의 단어는 빈도가 높았으나 많은 문서에서 흔하게 사용되어 DF는 커지고 TF-IDF 값이 작아지게 되어 기사에서 핵심어 역할을 하는 단어는 아니라는 것을 의미한다.

Table 5. TF-IDF analysis of keywords related to KNCAF in Naver News

순위	단어	TF-IDF	순위	단어	TF-IDF	순위	단어	TF-IDF
1	농업인	1.925427	21	농어업	1.262981	41	육성	1.022451
2	드론	1.911852	22	사업	1.247363	42	교수	1.021992
3	농림축산식품부	1.749107	23	고구마	1.218065	43	스마트팜	1.002380
4	전북	1.698069	24	대상	1.201233	44	팩트체크	0.993906
5	청년농업인	1.696237	25	공무원	1.193962	45	분야	0.992066
6	농업	1.664374	26	업무협약	1.187381	46	프레시안	0.979568
7	전주	1.624744	27	창업	1.172417	47	디지털	0.977469
8	대학	1.556556	28	재판매	1.128709	48	농협은행	0.977107
9	장치	1.512124	29	개발	1.115001	49	농장	0.969146
10	파종	1.438850	30	인력	1.103083	50	작업	0.967698
11	청년	1.431605	31	진행	1.092701	51	밀양시	0.963506
12	대표	1.430474	32	농촌	1.085171	52	정착	0.954366
13	교육	1.420341	33	활용	1.075331	53	선정	0.949649
14	농촌진흥청	1.405963	34	신입생	1.059819	54	선발	0.948956
15	공공기관	1.398940	35	발전	1.053456	55	농촌진흥청장	0.934527
16	지원	1.396383	36	아이디어	1.047816	56	산림	0.925391
17	직위	1.353076	37	산림과학	1.032248	57	전문	0.917653
18	영농	1.302337	38	기술	1.030866	58	인재	0.910410
19	졸업생	1.280533	39	도시인재전형	1.030683	59	대회의실	0.908040
20	양성	1.268312	40	일반전형	1.028771	60	계획	0.898947

Table 6. Comparison of word frequency keywords for top 10 keywords of TF-IDF

단어	농업인	드론	농림축산식품부	전북	청년농업인	농업	전주	대학	장치	파종
TF-IDF 순위	1	2	3	4	5	6	7	8	9	10
단어빈도 순위	4	46	20	51	16	1	94	6	186	103

다음은 몇 개의 단어에 대하여 findAssocs() 함수로 연관성 높아 함께 사용될 확률이 높은 단어를 분석한 결과를 <Table 7>에 나타냈다. 키워드 '농업인'과 연관성 높은 '여성', '기술교육', '청년', '영농', '정착' 등의 단어는 함께 사용될 확률이 52% 이상으로 나타났다. '드론'과 연관된 단어들의 경우에는 함께 나타날 확률이 65% 이상이며, 키워드 '농림축산식품부'의 연관 단어들과 거의 비슷한 단어인 결과가 나타났다.

Table 7. The associative words obtained by the findAssocs() function among the specific keywords

단어	연관어
농업인	여성, 기술교육, 청년, 영농, 정착, 창업, 농촌, 컨설팅, 업무협약, 지원
드론	파종, 정밀도, 시연회, 향상, 기존, 개발, 장치, 정밀파종장치, 농림축산식품부
농림축산식품부	정밀도, 파종, 시연회, 향상, 기존, 드론, 개발, 정밀파종장치, 장치, 정밀
청년	농업인, 여성, 창업, 활성화, 정착, 기술교육, 영농, 지원, 창업농
졸업생	학교, 화훼, 수산양식, 입학, 전문, 농어업계, 산림조경, 성공사례, 책자, 콘텐츠, 특용
전북	전주, 업무협약식, 농촌진흥청, 공공빅데이터, 시연회, 창업경진대회, 학생회관
디지털	기후변화, 시대, 대응, 정보, 구축, 지속, 기후변화교육센터, 전환, 농촌진흥청, 사료살포

3. 의미연결망분석

기사에 나타난 주요 단어의 연계성을 파악하기 위한 의미연결망분석을 하였다. 먼저 네이버 뉴스에서 '한국농수산대학'을 포함한 기사에 언급된 주제어를 추출하여 서로 연결하는 단어의 바이그램 빈도수 순위를 <Table 8>에 나타냈다. <Table 8>은 추출한 34,349개의 단어 바이그램 가운데 <Fig. 10>에 나타낸 except 변수의 단어가 포함되는 바이그램을 삭제하고, 빈도수 2회 이상의 상위 6,314개 단어 조합을 대상으로 작성한 바이그램(상위 30위)이다. 바이그램 순위를 보면 '청년'-'농업인', '디지털'-'농업', '영농'-'정착', '농업'-'농촌', '디지털'-'전환' 등의 순으로 빈도가 높게 나타났다. 이 결과에 의하면 네이버 뉴스에서의 한농대 이미지는 '청년 농업인', '디지털 농업', '영농 정착', '디지털 전환', '전북 전주', '인재 양성' 등의 단어들로 표현된다고 평가할 수 있다.

약 6,300여 개의 바이그램 가운데 단어 간의 관계를 그림으로 시각화하기 위하여 상위 150위의 바이그램을 대상으로 근접 중심성(Cc), 매개 중심성(Cb), 연결 중심성(Cd), 고유벡터 중심성(Ce) 및 페이지랭크 중심성(Cp) 평가지표로 단어 중요도를 평가한 결과(상위 20위)를 <Table 9>에 나타냈다.

<Table 9>에서 키워드 '농업'은 모든 중심성 지표에서 1위로 나타났으며, 근접 중심성(Cc)과 매개 중심성(Cb)에서는 '농업인', 연결 중심성(Cd)과 페이지랭크 중심성(Cp)에서는 '교육' 그리고

Table 8. The keywords of total bigram in semantic network analysis of words related to KNCAF in Naver News

(단위 : 회)

순위	단어 1	단어 2	빈도	순위	단어 1	단어 2	빈도
1	청년	농업인	86		영농	기반	23
2	디지털	농업	42	17	농업	미래	22
3	영농	정착	36	18	개방형	직위	20
4	농업	농촌	33	19	교육과정	개편	19
5	디지털	전환	32		업무협약	체결	19
	전북	전주	32		영농기반	점수	19
7	인재	양성	30		파종	살포	19
	중장기	발전방안	30	23	농식품	분야	18
9	전문	인력	29		스마트	농업	18
	정착	지원	29	25	4차산업혁명	기술	17
11	청년농업인	육성	28		디지털	교육	17
12	허태웅	농촌진흥청장	27		부동산	투자	17
13	귀농	귀촌	24		신입생	선발	17
	수산계열	학과	24		인력	양성	17
15	농업인	육성	23		정밀	파종	17

Table 9. The results of centrality evaluation of top 150 bigram in semantic network analysis for KNCAF in Naver News

순위	Cc	Cb	Cd	Ce	Cp
1	농업	농업	농업	농업	농업
2	농업인	농업인	교육	미래	교육
3	디지털	지원	미래	디지털	활용
4	미래	정착	디지털	농수산업	영농
5	관련	영농	영농	농어업	농촌
6	농촌	교육	농업인	농업인	미래
7	인재	농촌	농촌	농촌	공공기관
8	분야	디지털	전문	관련	도시인재전형
9	스마트	미래	농수산업	인재	농업인
10	기술	인재	농어업	분야	디지털
11	지원	육성	활용	교육	전문
12	발전	관련	육성	발전	파종
13	실습장	스마트	청년	스마트	육성
14	디지털화	농수산업	인재	기술	문제
15	육성	기술	정착	실습장	농수산업
16	전문	기반	파종	디지털화	농어업
17	청년	일손부족	스마트	현장	졸업생
18	농어업	전문	문제	육성	신입생
19	교육	분야	공공기관	전환	대학
20	현장	청년	기술	리더	고등학교

고유벡터 중심성(Ce)에서는 '미래'가 중심성 2위의 키워드로 나타났다.

중심성 지표 가운데 연결된 Node의 수가 많을수록 연결 정도 중심성이 높다고 판단하는 연결 중심성(Cd)에서는 '농업', '교육', '미래', '디지털', '영농', '농업인', '농촌' 등의 키워드가 높은 연결 중심성을 나타냈다. 그러나 단순히 연결된 선의 개수만 가지고 연결의 중심을 파악하는 건 한계

가 있다.

이러한 점을 보완하고 있는 고유벡터 중심성(Ce) 지표는 해당 Node에 직접 연결된 다른 Node들의 개수뿐만 아니라 자신과 연결된 다른 Node들의 영향력을 합해 중심성을 평가하는 방식으로 '농업', '미래', '디지털', '농수산업', '농어업', '농업인', '농촌' 등의 키워드가 높은 영향력을 갖는 결과를 나타냈다. 다른 중심성 지표보다 '교육' 순위가 낮아지고 '관련', '분야', '실습장', '발전', '전환' 등 키워드가 나타났으며, '농수산업', '농어업' 키워드 순위가 높아졌다. 고유벡터 중심성은 한 Node가 중심성이 아주 높게 나올 경우 이와 연결된 다른 Node도 덩달아 지나치게 중심성이 올라갈 수 있다는 단점이 있다.

이 점을 고려하여 중심성을 다른 Node로 전달할 때 일정 부분 제한하는 페이지랭크 중심성(Cp) 결과에서는 '농업', '교육', '활용', '영농', '농촌', '미래' 등의 키워드의 영향력이 높은 결과를 나타냈다. 다른 중심성 지표보다 '디지털', '농업인' 순위가 낮아지고 '활용', '공공기관', '도시인재전형', '졸업생', '신입생', '대학', '고등학교' 등 키워드가 나타났다. 이는 다양한 뉴스 기사 정보를 반영하여 링크의 절대적인 수보다는 중요도가 높은 기사부터의 연결 여부를 더 중요하게 판단한 결과라고 할 수 있다.

페이지랭크 중심성(Cp)에 대한 시각화 결과를 <Fig. 12>에 나타냈다. 중심성 상위에 나타난 키워드(붉은 실선)를 중심으로 하는 연관 단어들의 네트워크(파랑 점선)가 잘 나타나고 있음을 볼 수 있다.

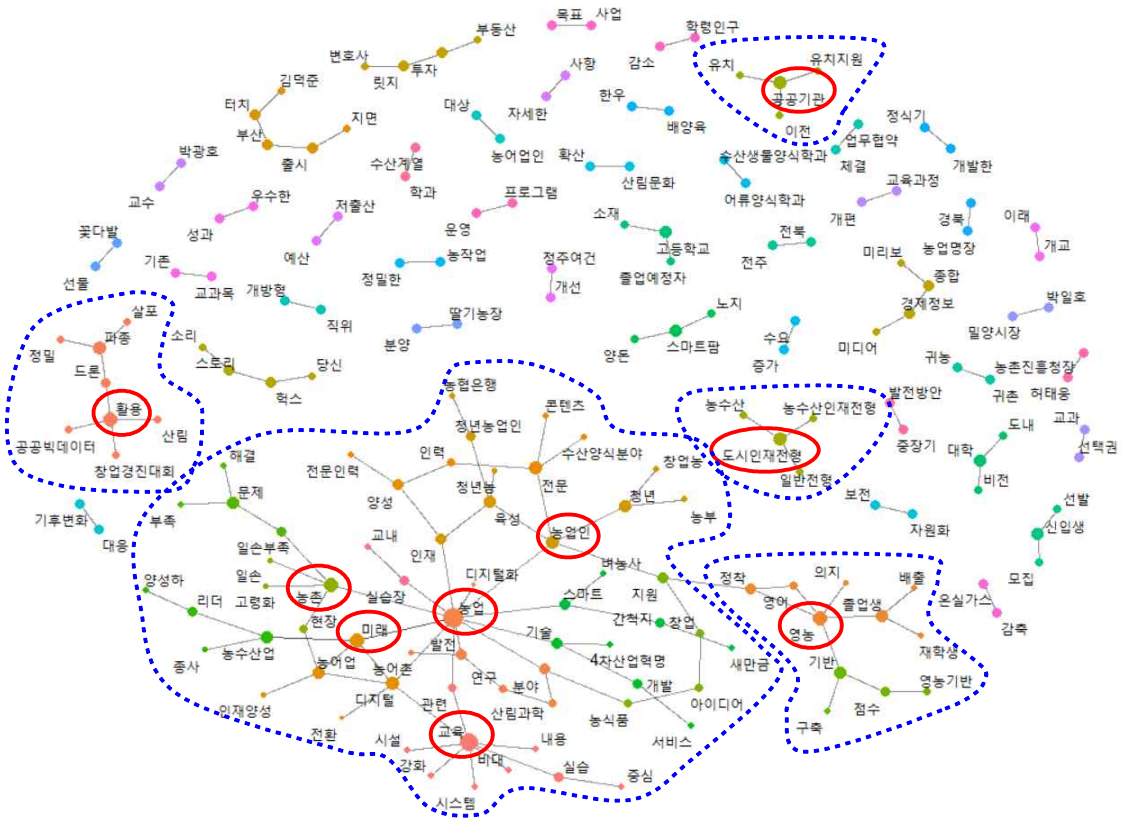


Fig. 12. Visualization of pagerank centrality in semantic network analysis of KNCAF articles

상위 150위 바이그램을 대상으로 산출한 중심성 지표별 키워드의 순위가 얼마나 비슷한가를 알아보기 위해서 변수 순위 사이의 통계적 의존성을 측정하는 스피어먼 순위 상관계수를 구하여 <Fig. 13>에 나타냈다. 스피어먼 순위 상관계수를 보면 연결 중심성(Cd)과 페이지랭크 중심성(Cp)이 0.89 전후의 상관관계를 보여서 가장 유사한 순위를 산출한 것으로 나타났다. 한편, 근접 중심성(Cc)과 페이지랭크 중심성(Cp)과의 상관관계는 0.01로서 순위의 유사성이 가장 낮은 것으로 나타났다.

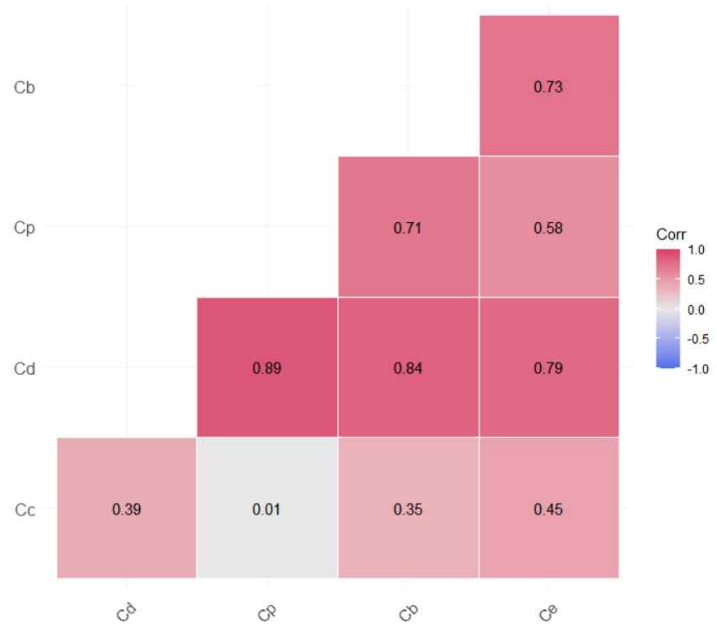


Fig. 13. Spearman's rank correlation coefficient for the ranking of centrality evaluation index in semantic network analysis

V. 적요

빅데이터 분석기술인 웹 크롤링 기술을 이용하여 네이버 뉴스 데이터 내에 담겨 있는 '한농대'에 대한 이미지 단어를 추출하였다.

뉴스 기사에서 언급된 빈도에 따라 중요한 단어로 평가는 단어빈도 분석에서는 청년농업인을 육성하는 한농대의 특성을 잘 설명하는 '농업', '교육', '지원', '농업인', '청년', '대학', '사업', '농촌', '대표' 등의 단어가 자주 사용되는 것으로 나타났다. 또한 '디지털', '스마트', '드론', '졸업생', '창업', '새만금', '교육과정' 등 디지털 농업 전문 인재를 육성하기 위한 학교의 교육, 지원, 비전 등과 관련한 단어들이 추출되었다.

모든 기사 데이터의 단어 빈도(TF) 및 역 문서 빈도(IDF)를 이용한 TF-IDF 가중치의 전체 순위는 '농업인', '드론', '농림축산식품부', '전북', '청년농업인', '농업', '전주', '대학', '장치', '파종' 등의 단어가 한농대와 관련된 뉴스 기사에서 중요한 핵심어 역할을 하는 것으로 나타났다. 단어

빈도에서 '드론', '농림축산식품부', '전북', '청년농업인', '전주', '장치, '파종' 등은 순위가 매우 낮았으나 TF-IDF 가중치 순위에서는 한농대를 표현하는 핵심어로 나타났다. TF-IDF 평가에서 '교육', '지원', '청년', '사업', '농촌' 등의 키워드는 단어빈도가 높으면서 많은 문서에서 자주 등장하는 키워드로서 핵심어 역할은 크지 않은 것으로 나타났다.

단어 간 연계성을 파악하기 위한 의미연결망 분석에서 추출한 바이그램은 '청년'-'농업인', '디지털'-'농업', '영농'-'정착', '농업'-'농촌', '디지털'-'전환' 등의 순으로 빈도가 높게 나타났다. 중심성 지표로 키워드의 영향력을 평가한 결과 모든 지표에서 '농업'이 1위로 나타났으며, 2위에는 '농업인'(근접 중심성, 매개 중심성), '교육'(연결 중심성, 페이지랭크 중심성) 및 '미래'(고유벡터 중심성)으로 나타났다. 스피어먼 순위 상관계수에 의한 중심성 지표별 키워드의 순위의 유사성은 연결 중심성과 페이지랭크 중심성이 0.89 전후의 가장 높은 상관관계를 보였다.

이상으로 네이버 뉴스의 한농대 관련 기사에서 단어 빈도로 보면 '농업', '교육', '지원', '농업인', '청년', '대학', '사업', '농촌', '대표' 등이 중요한 단어로 평가되었으나, 문서빈도를 함께 고려한 평가에서는 '농업인', '드론', '농림축산식품부', '전북', '청년농업인', '농업', '전주', '대학', '장치', '파종' 등의 단어가 핵심어 역할을 하는 것으로 나타났다. 한편 단어나 문서의 빈도가 아니라 단어 간 네트워크 연계성을 고려한 중심성 분석에서는 연결 중심성과 페이지랭크 중심성에 의한 평가가 적합한 것으로 나타났으며, '농업', '교육', '미래', '농업인', '디지털', '지원', '활용' 등이 중심성이 강한 단어로 나타났다.

Ⅴ. 참고문헌

1. 김영우. (2017). 쉽게 배우는 R 데이터 분석, 이지스퍼블리싱.
2. 조민호. (2019). 데이터 분석 전문가를 위한 R 데이터 분석. 정보문화사
3. 박경진, 정덕호, 하민수, 이준기. (2014). 언어네트워크분석에 기초한 과학학습의 목적에 대한 고등학교 교사와 학생들의 인식. Journal of the Korean Association for Science Education, 34(6), 571~581
4. 주진수 외 5인. (2020). 한국농수산대학 신입생 자기소개서의 텍스트 마이닝과 연관규칙 분석 (1). 현장농수산연구지 Vol. 22(1), No.1: 113-130.
5. 주진수 외 5인. (2020). 한국농수산대학 신입생 자기소개서의 텍스트 마이닝과 연관규칙 분석 (2). 현장농수산연구지 Vol. 22(2), No.2: 99-114.
6. 주진수 외 5인. (2021). 언어네트워크분석을 활용한 한국농수산대학 신입생 자기소개서 분석. 현장농수산연구지 Vol. 23(1), No.1: 89-104.
7. https://bookdown.org/yuaye_kt/RTIPS/Texnetword.html. Chapter 11 텍스트 데이터-단어 네트워크맵(1)
8. https://briatte.github.io/ggnet/. ggnet2: network visualization with ggplot2
9. https://da-it-so.tistory.com/43. TF-IDF 기법 이해하기
10. https://data-traveler.tistory.com/33. R을 이용한 텍스트마이닝_TF-IDF(코드 및 설명)
11. https://iamdaisy.tistory.com/31?category=620658. 소셜네트워크 분석의 이해

논문접수일 : 2021년 10월 29일
논문수정일 : 2021년 11월 14일
게재확정일 : 2021년 12월 8일