J. of Practical Agriculture & Fisheries Research ISSN 2234-1129
2021 June. : 23(1): 89-104

언어네트워크분석을 활용한 한국농수산대학 신입생 자기소개서 분석
- TF-IDF 분석을 기초로 -

Analyzing Self-Introduction Letter of Freshmen at Korea National College of Agricultural and Fisheries by Using Semantic Network Analysis : Based on TF-IDF Analysis

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

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

김종숙
J. S. Kim
국립한국농수산대학 [1]
농수산비지니스학과
jskimsy@korea.kr

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

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

Abstract

Based on the TF-IDF weighted value that evaluates the importance of words that play a key role, the semantic network analysis(SNA) was conducted on the self-introduction letter of freshman at Korea National College of Agriculture and Fisheries(KNCAF) in 2020.

The top three words calculated by TF-IDF weights were agriculture, mathematics, study (Q. 1), clubs, plants, friends (Q. 2), friends, clubs, opinions, (Q. 3), mushrooms, insects, and fathers (Q. 4). In the relationship between words, the words with high betweenness centrality are reason, high school, attending (Q. 1), garbage, high school, school (Q. 2), importance, misunderstanding, completion (Q.3), processing, feed, and farmhouse (Q. 4). The words with high degree centrality are high school, inquiry, grades (Q. 1), garbage, cleanup, class time (Q. 2), opinion, meetings, volunteer activities (Q.3), processing, space, and practice (Q. 4). The combination of words with high frequency of simultaneous appearances, that is, high correlation, appeared as ‘certification - acquisition’, ‘problem - solution’, ‘science - life’, and ‘misunderstanding - concession’.

In cluster analysis, the number of clusters obtained by the height of cluster dendrogram was 2(Q.1), 4(Q.2, 4) and 5(Q. 3). At this time, the cohesion in Cluster was high and the heterogeneity between Clusters was clearly shown.

Key words : Semantic network analysis, Association rules analysis, Betweenness centrality, Degree centrality

*교신저자
1 Korea National College of Agriculture and Fisheries

Ⅰ. 서론

텍스트 마이닝 (Text Mining)은 텍스트 형태로 이루어진 비정형 데이터들을 자연어 처리 방식을 이용하여 정보를 추출하는 기법이다. 비정형 데이터에서 의미 있는 정보를 추출하기 위하여 자주 등장하는 키워드 고빈도 키워드를 생각해 볼 수 있다. 그러나 고빈도 키워드가 항상 중요한 키워드라고 할 수 없다. 고빈도 키워드는 중요한 키워드일 수도 있으나 동시에 흔한 키워드일 가능성도 높기 때문이다.

이런 단점을 보완하기 위하여 **텍스트 빈도수를 산출하여 가중치를 부여하여 연구의 정확도를 높이는** TF-IDF(Term Frequency-Inverse Document Frequency) 분석법이 활용되고 있다. TF-IDF는 정보검색론(Information Retrieval)에선 흔하게 접하는 가중치를 구하는 알고리즘으로서 기사, 논문, 리뷰 등 다량의 문서에서 핵심적으로 사용된 단어 그리고 의미 있는 단어들을 찾아 주제 등을 유추할 수 있는 장점이 있다.

본 연구에서는 한국농수산대학(이하 한농대) 신입생(550명)의 자기소개서에서 문항별로 정보를 추출하기 위하여 핵심적인 역할을 하는 단어의 중요도를 평가하는 TF-IDF 가중치를 기초로 한 언어네트워크분석(SNA: Semantic network analysis)을 하였다.

언어네트워크분석 방법을 사용한 이유는 자소서 내 '단어'를 텍스트의 특성을 나타내는 기본개념으로 간주하여 분석 단위로 삼으며 단어의 특성과 단어들 간의 의미적 관계에서 나타내는 속성들을 파악하는 데 유용하며, 자소서 내 출현 단어의 단순 빈도 분석보다 유의미한 구조를 발견할 수 있는 분석을 하는 데 목적이 있다. 분석 도구로는 R과 한글 형태소 분석을 위한 KoNLP 패키지를 이용하였다.

언어네트워크분석의 구체적 방법으로 단어의 중요도를 글자 크기나 색깔로 표시하는 워드 클라우드(Word Cloud)에 의한 시각화와 네트워크 다이어그램(Network diagram)을 활용한 중심성 분석(Centrality analysis)을 이용하였다. 또한 단어 기반의 계층적 클러스터링 기법을 이용한 군집분석(Cluster analysis)을 위하여 클러스터 덴드로그램(Cluster dendrogram) 방법을 이용하였다.

Ⅱ. 연구내용 및 방법

1. TF-IDF

TF-IDF(역문서 빈도) 가중치 모델은 텍스트 마이닝을 위해서 문서 내에서 단어의 중요도를 평가하기 위한 표현방식으로 지프의 법칙[1]이 이론적 배경이 되고 있다. TF-IDF 가중치 값이 큰 단어일수록 속해 있는 문서의 주제나 의미를 결정지을 가능성이 크며 이 측정치를 주요 키워드를 추출할 수 있는 척도로 활용할 수 있다.[2] TF-IDF 모델은 SNS나 뉴스 기사에서 자주 언급되는 이슈들을 분석하는 데 유용하게 쓰인다. 특정 단어가 문서 내에서 얼마나 중요한지 척도 계산, 문서 내 단어들에 척도를 계산해서 핵심어 추출, 검색엔진에서 검색 결과의 순위 결정 그리고 네트워크 관계 등을 발견할 수 있다.

빈도수가 높은 단어와 문서에 점수를 부여한

1) 단어의 출현 빈도에 관한 성질로서, 단어를 출현 빈도가 큰 순으로 늘어세워 순위번호를 1, 2,… 라고 붙이면, '순위번호×출현 빈도=일정'이라고 하는 관계가 성립한다고 하는 경험법칙. G. K. Zipf가 1936년에 제창하였다.
2) 이성직·김한준, "TF-IDF의 변형을 이용한 전자뉴스에서의 키워드 추출기법". 한국전자거래학회보, 한국전자거래학회, 2009, 제14집 제4호, p.61.

후 확률모형을 통해 주가 변동을 예측한 연구[3]는 역문서 빈도가 출현 단어와 해당 문서의 중요도를 얼마나 핵심적으로 파악하고 있는지 보여주는 사례라고 볼 수 있다.

TF(Term Frequency, 단어 빈도)란 어떤 범위 내의 문서에서 등장하는 특정 단어의 빈도수, 즉 TF는 단어 빈도를 기본으로 문서와의 연관성을 간단하게 정규화시키는 방법[4]으로 수치화한다는 것이다. 문서 내에서 많이 출현할수록 상대적으로 더 중요하다는 의미일 수 있으나 지나치게 빈도수가 높은 단어는 불용어 수준의 단어에 해당할 수 있다. 따라서 TF 분석 이후 연구자가 의도하는 주제에 따라 불용어 처리 여부를 결정해야 한다. 이를 보완하기 위해 DF(단어가 출현한 문서의 수)를 사용해야 한다.

문서 d_j에서 단어 t_i의 중요도는 다음과 같다.

$$TF_{i,j} = \frac{n_{i,j}}{\Sigma_k n_{k,j}} \qquad (1)$$

$n_{i,j}$:문서 d_j에서 단어 t_i가 나오는 횟수

$\sum_k n_{kj}$:문서 d_j에서 나오는 모든 단어 횟수

DF(Document Frequency, 문서 빈도)는 **자주 등장하는 단어가 몇 개의 문서에 등장하는지, 즉** 한 단어가 공통으로 나타난 문서의 수라고도 할 수 있으며, **전체 문서 수(d_j)에 대한 단어(t_i)의 중요도를 나타낸다.** "있다", "하다", "된다"와 같은 흔한 단어는 DF 값이 높게 나타난다. DF가 높을수록 문서에 많이 쓰이는 범용적인 단어라고 볼 수 있으며 중요하지 않은 단어일 가능성이 높다. DF 값이 클수록, TF-IDF의 가중치 값을 낮춰주기 위해서 DF 값에 역수를 취한 값이 바로 IDF가 된다.

IDF는 TF와 반대되는 개념으로 특정 단어가 나타나는 문서의 수를 의미한다. 어느 특정 단어가 제공하는 정보의 양을 나타내는 기준으로서, 용어가 모든 문서에서 일반적인지 혹은 생소한지 아닌지를 알 수 있다. IDF는 그 값이 클수록 특이한, 즉 희귀한 단어라는 의미이다. 희귀 단어는 IDF가 높게 나오다가 단어(t_i)가 나오는 문서가 늘어나면 IDF가 감소하면서 임계치인 0에 가깝게 내려간다. 대부분의 불용어는 대부분의 문서에 포함되므로 IDF는 0에 가깝게 나온다.

$$IDF_i = \log\left(\frac{|D|}{|\{d_j : t_i \in d_j\}|}\right) \qquad (2)$$

$|D|$: 전체 문서 수

$|\{d_j : t_i \in d_j\}|$: 단어 t_i가 나오는 문서의 수(= DF)

TF-IDF 가중치는 TF와 IDF를 곱한 것이며 TF-IDF 가중치 값에 따라 문서 내에서 단어(t_i)의 출현 빈도가 높고 전체 문서에서 단어(t_i)가 출현하는 문서들의 수가 적은 단어가 중요한 단어로 평가된다. TF-IDF 가중치는 하나의 문서에서 TF가 크고 전체 문서에서 DF가 작을수록 그 값이 높아지며, 값이 높아질수록 상대적으로 문서 내에서 핵심적인 단어라고 할 수 있다. 즉 TF × IDF 값이 크다면 많이 사용되는 단어이면서 여러 문서에서 공통으로 나타나는 단어가 아닌 의미 있는 단어라 할 수 있다. 따라서 이 값을 이용하여 모든 문서에 나타나는 흔한 단어들을 걸러내며, 특정 단어가 가지는 중요도를 알 수 있다.

$$TFIDF_{i,j} = TF_{i,j} \times IDF_i \qquad (3)$$

TF-IDF는 기사, 논문, 리뷰 등 다량의 문서에서 중요 키워드와 키워드의 빈도수를 산출하여 가중치를 부여하고 연구의 정확도를 높이는 방법이라 할 수 있다.

3) Fawcett·Provost, Data science for business. O'REILLY·Hanbit, 2014, pp. 312-323.

4) TF 값을 정규화시키는 방법으로 불린 빈도(Boolean Frequency), 로그 스케일 빈도(Logarithmically Scaled Frequency), 증가 빈도 (Augmented Frequency)가 있다.

2. 언어네트워크분석

언어네크워크분석(SNA)은 사회네트워크분석(Social Network Analysis)을 텍스트에 응용한 내용분석의 한 방법으로 전통적인 내용분석 방법보다 더 계량화되고 도식화된 결과물을 도출할 수 있다. 언어네트워크분석은 의사소통의 메시지나 아이디어의 의미를 조사하는 방법으로 사용되는 내용분석(Content analysis) 기법의 한계점을 보완하여 사용되는 기법이다.

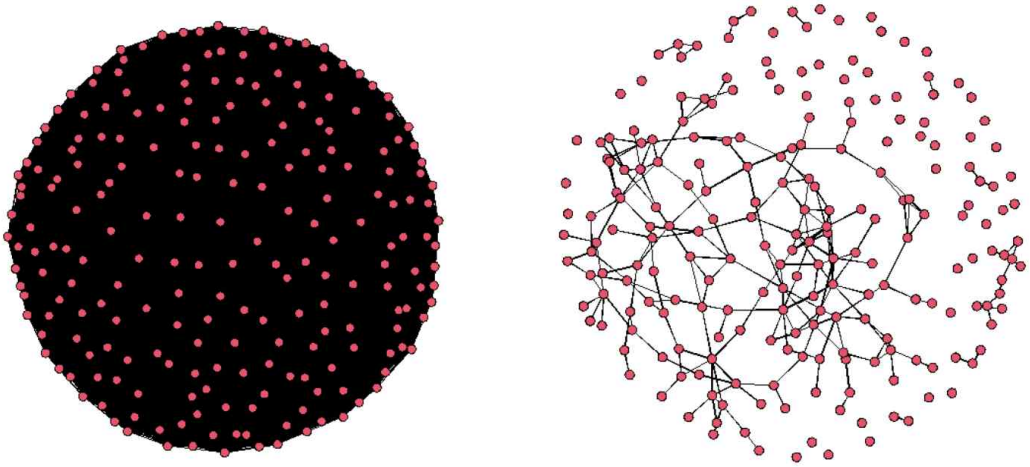
언어네트워크분석은 비정형 텍스트의 주요 개념들 및 단어 간의 관계를 시각적으로 분석하여 주요 개념 간의 관계를 보다 정확하게 분석할 수 있다. 이는 비정형의 임의의 텍스트로부터 정보를 추출하게 되지만 구조화된 형태 확보가 가능한 방법으로, 각각 주요 개념들의 커뮤니케이션 과정 및 중요도를 가시화할 수 있어 각각의 개념에 대한 의미 분석을 가능하게 해준다. 또한 기본적으로 핵심 키워드 간의 구조적 관계를 파악하는 기법으로, 미디어 자료를 분석할 때 더욱더 유의미한 결과 도출이 가능하다는 장점이 있다. 그뿐만 아니라, 언어네트워크 분석법은 개념들 사이의 관계를 시각화하여 주요 개념과 각 개념들 사이의 관계 강도까지 한눈에 볼 수 있어서 보다 효과적이다.[8)]

언어네트워크분석은 빅데이터 분석 방법 가운데 하나로 질적인 방법과 양적인 방법을 모두 적용할 수 있는 매우 큰 장점이 있는 기법이다(김학준 등, 2019). 언어네트워크분석은 행위자 간의 연결성을 중요시하는 사회네트워크분석과는 달리 단어와 개념 등 텍스트의 공유된 의미를 기반으로 시스템 구조를 분석하는 데 중점을 두는 방법이다. 이러한 방법론적인 장점에 따라 언어네트워크 분석을 활용한 많은 연구들이 다양한 분야에 적용되고 있다. 특히, 언론 보도 및 각종 토론내용 및 연설문의 내용분석이 필요한 커뮤니케이션 등과 같은 텍스트 간의 연결성을 파악해야 하는 곳에 주로 사용되고 있다.

또한, 언어네트워크분석에서 네트워크 구조의 중심성을 파악하는 방법은 매개중심성(Betweenness Centrality), 연결정도중심성(Degree Centrality), 근접중심성(Closeness Centrality), 아이겐벡터중심성(Eigenvector Centrality) 등이 있으며, 각각의 중심성이 높게 나타난 개념을 찾아, 문장이 나타내고자 하는 중요 의미 및 핵심어 등을 파악하는 데에 용이하다. 중심성(centrality) 지표는 네트워크 분석지표 중 가장 많이 사용되는 지표로, 전체 연결망 내에서 중요한 역할을 하는 노드가 무엇인지를 보여준다.

그중에서 매개중심성은 네트워크 내에서 한 노드가 담당하는 매개자 역할의 정도로서 중심성을 측정하는 방법으로서 전체 네트워크에서 해당 노드와 다른 노드 사이에 있는 최단 경로수를 측정하여 노드 간의 매개체 역할을 하는 노드를 알 수 있다. 즉 한 노드가 연결망 내의 다른 노드들 사이의 최다 경로 위에 위치하면 할수록 그 노드의 매개중심성이 높은 것을 의미하며, 매개중심성이 높은 노드는 “정보의 흐름”을 통제하는 데 큰 영향력을 가질 수 있다. 연결정도중심성은 한 노드에 직접적으로 연결된 다른 노드의 개수에 따른 중요도 측정 방법이다. 한 노드(키워드)가 얼마나 많은 다른 노드들과의 연결 관계를 맺고 있는지를 측정하는 지표로 노드의 중심 정도를 계량화 한 것이다. 일반적으로 쓰이는 Degree를 의미한다.

한편 매개중심성 분석을 나타내는 단어네크워크 시각화에는 노드를 연결하는 edge가 너무 많으면 시각화가 어렵기 때문에 본 연구에서는 30개(문항 1, 3), 35개(문항 2), 50개(문항 4) 이상의 문서에 나타나는 단어를 대상으로 노드 수를 구한 후, 상관행렬의 상관계수 값의 기준 설정으로 적절한 edge수를 구하여 네트워크 맵을 작성하였다. Fig. 1은 문항 3에서 상관계수를 보정하여 edge 수를 23,871개에서 161개의 edge로 조



(a) vertices = 219, total edges= 23871 (b) vertices = 219, total edges= 161

Fig. 1. Semantic network analysis by edge number difference

정하여 시각화한 네크워크 맵이다.

3. 군집분석과 덴드로그램

군집분석은 변수 또는 개체(item)들이 속한 모집단 또는 범주에 대한 사전정보가 없는 경우에 관측값 사이의 유사성을 이용하여 변수 또는 개체들을 자연스럽게 몇 개의 그룹 또는 군집(Cluster)으로 나누는 분석 방법이다.

군집분석은 유사한 속성을 갖는 객체들끼리 묶어 그룹화 하는 과정이며 군집 간 유사도나 거리함수에 기반한다. 본 연구의 군집분석 방법은 데이터의 거리 또는 유사도를 토대로 덴드로그램(Dendrogram) 형태의 군집형성을 수행하여 군집을 결정하는 계층적 군집화(Hierarchical clustering) 방법을 이용하였다. 계층적 군집분석은 개체간의 유사도, 또는 비유사도와 같은 지표를 기본적인 개체 간 데이터로 설정하여, 차례대로 분석대상들의 군집을 구성하고 이를 전체 데이터에 확대하는 분석방법이다. 군집화 방법으로는 두 개의 군집을 융합할 때 집단 내의 분산과 집단 간의 분산비율을 최대화하는 군집을 구성하는 방법으로서 가장 명확한 군집이 만들어지며, 분류 결과도 좋은 것으로 알려진 워드법(Word's Method)을 사용하였다.[5)]

군집화에서 군집의 개수 k는 덴드로그램으로 시각화한 후에 왼쪽의 높이(Height) 축을 기준으로, 위에서 아래로 높이를 이동하면 군집의 개수가 1개, 2개, 4개, 6개, ..., 30개, 이런 식으로 변하게 된다. 이때 군집 간 높이의 차이가 큰 군집의 개수를 선택하면 군집 내 응집력은 높고, 군집 간 이질성이 큰 적절한 군집을 구할 수 있다.

R에서 계층적 군집분석을 수행하는 함수는 hclust() 함수이며, 각 개체의 거리를 산출하는 함수는 dist() 함수이다. 군집분석에서는 분류할 데이터가 너무 많으면 설명에 초점을 맞추기 어려운 경우가 발생하므로 일부 데이터를 추출하여 사용해야 할 필요가 있다. 이에 본 연구에서는 문항에 따라 80개(문항 1, 3), 100개(문항 2), 120개(문항 4) 문장 이상에서 나타나는 단어를 대상으로 군집분석을 하였다.

5) https://m.blog.naver.com/pmw9440/221597864343 R에서 계층적 군집분석 실시하기

4. 분석 방법

본 연구에 활용한 분석 자료는 2020년 한농대 신입생 550명의 자소서이며, 분석 프로그램은 R 프로그래밍 언어에 기반한 RStudio(버전 4.05)이다. 분석 방법은 텍스트 데이터로부터 Text processing 기술 및 처리 과정인 텍스트 마이닝 기법을 활용하여 문항별로 TF-IDF 가중치로 핵심단어를 추출하고 워드클라우드, 언어네크워크분석 및 군집분석 등으로 단어의 연관성 및 시각화 등을 검토하였다. RStudio에 의한 분석 순서 및 내용을 Table 1에 나타냈다.

Table 1. Analysis order and contents using RStudio

순서	분석 내용	계산 함수 및 옵션
1. 데이터 처리	① 데이터 입력	UTF-8 형식의 텍스트 파일
	② 단어 추출	
	③ 단어 정제	
	④ 단어 빈도수 산출	빈도수 상위 50위
	⑤ 워드클라우드 작성 (시각화)	
	⑥ 말뭉치 구하기	VCorpus()
2. TF-IDF	⑦ TDM 작성과 TF-IDF 계산	TermDocumentMatrix() wordLengths : 단어길이 설정 bounds : 단어가 등장하는 문서 수 설정 weightTfIdf : TF-IDF 가중치 계산
	⑧ 워드클라우드 작성 (시각화)	TF-IDF 값 상위 50위
	⑨ 단어별 연관성 단어 추출	findAssocs()
3. 언어네트워크	⑩ 상관행렬 구하기	TDM을 DTM으로 전치(변환) 단어간 상관행렬 만들기 상관행렬 크기 조정
	⑪ 매개중심성 산출 ⑫ 연결정도중심성 산출	Betweenness centrality Degree centrality
	⑬ 네트워크 맵 그리기 (시각화)	ggnet2() 매개중심성 상위 10% 노드 노란색 library : network, GGally, sna
4. 군집분석	⑭ 군집분석	거리계산 dist() 군집계산 hclust()
	⑮ dendrogram 작성 (시각화)	rect.hclust()

Ⅲ. 결과 및 고찰

1. TF-IDF

TF와 IDF를 곱으로 나타내는 TF-IDF 가중치 산출 결과(상위 20위)를 Table 2와 같으며 얻어진 빈도 순위는 키워드가 전체 문서에 나타난 빈도수의 순위를 의미한다. TF-IDF 가중치에 의한 순위와 빈도 순위를 비교하면 전혀 다른 결과를 보이는 것을 알 수 있다.

Table 2에는 나타내지 않으나 단어 '생각'은 자소서에 등장한 빈도수는 문항 1 ~ 문항 3에서 2위, 문항 4에서 1위로 모두 높게 나타났으나[10), 11)], TF-IDF 순위는 각 문항에서 57위, 88위, 56위, 129위로 나타나 자소서에 많이 등장한 단어이지만 중요도는 높지 않은 단어로 나타났다. 이

는 단순히 빈도 순위가 높은 단어가 반드시 핵심적인 단어라 할 수 없으며, TF-IDF 값이 높은 단어가 특정 문서 내에서 많이 등장하는 단어이지만 여러 문서에는 자주 등장하지 않으면서 중요도가 높은 의미 있는 단어라는 것을 알 수 있다.

문항 1에서는 '농업', '수학', '공부', '문제', '친구' 등이 고교재학 기간 중 학업에 기울인 노력과 학습 경험을 표현하는 데 중요한 핵심단어인 것으로 나타났으며, 특히 '수학', '자격증', '성적', '영어', '과학', '활동', '과목', '사람', '발표' 등의 낮은 빈도 순위의 단어들은 흔하게 사용되지 않은 단어이지만 TF-IDF 순위가 높게 나타나 본 문항을 서술하는 중요 핵심단어인 것을 알 수 있다.

문항 2는 재학 기간에 의미를 두고 노력했던 교내 활동을 기술하는 문항으로서 단어 '활동'이 빈도 순위 1위로 조사되었으나 TF-IDF 순위는 18위로써 많은 학생들이 자주 사용한 단어로 나타났다. 문항 2의 중요 핵심단어는 '동아리', '식물', '친구', '농업', '작물'로 나타났으며, 특히 '식물', '작물', '공부', '쓰레기', '발표', '실험', '자격증', '판매', '의견', '고등학교', '토론' 등의 빈도 순위를 보면 자소서에 많이 등장하지 않은 단어였으나 TF-IDF 순위를 보면 중요 핵심단어로 사용된 것을 알 수 있다.

문항 3은 학교생활 중 '배려, 나눔, 협력, 갈등, 관리' 등을 실천한 사례를 서술하는 문항으로서 '친구', '동아리', '의견', '학생', '선생님' 등의 단어가 중요한 핵심단어로 나타났다. 또한 '동아리', '청소', '봉사', '갈등', '조원', '연습', '봉사활동', '협력', '실습' 등 자소서에 많이 사용되지 않은 낮은 빈도 순위 단어의 TF-IDF 순위가 높게 나타나 본 문항을 서술하는 중요한 의미를 지닌 단어로 나타났다.

Table 2. TF-IDF analysis results for each question

	문항 1			문항 2			문항 3			문항 4		
순위	키워드	TF-IDF	빈도 순위	키워드	TF-IDF	빈도 순위	키워드	TF-IDF	빈도 순위	키워드	TF-IDF	빈도 순위
1	농업	10.148	7	동아리	7.566	4	친구	11.262	1	버섯	8.666	39
2	수학	9.504	24	식물	7.172	19	동아리	11.231	10	곤충	8.240	49
3	공부	8.868	1	친구	6.577	3	의견	9.746	4	아버지	8.158	10
4	문제	8.857	4	농업	6.355	12	학생	8.854	9	농업	8.009	2
5	친구	8.148	6	작물	5.827	22	학급	8.523	15	농장	7.186	4
6	자격증	7.465	33	공부	5.499	24	선생님	8.247	8	양식	6.363	67
7	수업	7.236	10	쓰레기	5.336	87	사람	7.973	6	부모님	6.212	16
8	이해	7.187	8	발표	5.283	21	청소	7.526	28	농사	6.114	17
9	성적	7.174	18	학생	5.279	14	활동	7.501	5	한우	5.754	61
10	영어	7.120	32	실험	5.179	58	봉사	7.460	34	조경	5.700	76
11	과학	7.110	27	문제	5.134	15	갈등	7.229	11	재배	5.581	12
12	내용	6.718	12	사람	5.096	7	생활	7.081	19	작물	5.328	19
13	노력	6.694	3	선생님	5.091	10	조원	7.050	43	산림	5.070	108
14	시간	6.570	5	자격증	4.938	79	학교	6.917	7	목장	5.064	103
15	활동	6.511	20	판매	4.617	48	문제	6.798	13	사람	4.782	8
16	학교	6.487	16	의견	4.593	32	도움	6.702	16	화훼	4.632	93
17	과목	6.468	19	고등학교	4.532	31	연습	6.619	39	생산	4.367	15
18	사람	6.452	28	활동	4.528	1	봉사활동	6.495	41	공부	4.310	6
19	선생님	6.435	9	학교	4.499	5	협력	6.494	18	운영	4.273	25
20	발표	6.419	31	토론	4.460	55	실습	6.464	60	기술	4.122	11

다른 문항과는 다르게 문항 3의 단어 '친구'는 TF-IDF 순위와 빈도 순위 모두 1위로 나타났는데 이는 다른 문항의 1위 단어보다 2배 이상 높은 빈도수(3,008)를 보이며 TF 값이 매우 높아진 것이 원인이다. 문항 3의 '친구'는 DF 값이 매우 높은(IDF 값이 매우 낮은), 즉 다수의 학생들이 많이 사용한 범용적인 단어이면서 핵심적인 단어라고 할 수 있다.

문항 4에서는 '버섯', '곤충', '아버지', '농업', '농장' 등의 단어가 중요 핵심단어로 나타났으며, 특히 낮은 빈도 순위의 '버섯', '곤충', '양식', '한우', '조경', '산림', '목장', '화훼' 등의 단어는 자소서에 많이 사용되지 않았으나 TF-IDF 값이 높게 나타나 한농대 지원동기와 학업계획 및 향후 영어·영농에 대한 진로계획을 서술하는 핵심단어로 사용된 것으로 나타났다.

Table 3은 단어 빈도와 문서 빈도를 이용한 TF-IDF 가중치 산출 결과(상위 50위)를 시각적으로 표현한 문항별 워드 클라우드(Word Cloud)이다. 단어의 중요도를 글자 크기나 색깔로 표시하는데, TF-IDF 가중치가 높은 단어는 크게 표시되기 때문에 한눈에 핵심 내용을 파악할 수 있다.

Table 4는 Table 2에서 TF-IDF 순위가 높은 각 단어들을 대상으로 함수 findAssocs()을 사용하여 해당 단어와 연관성 있는 단어를 구한 결과이다. 즉 자소서에서 해당 단어와 일정 확률(연구자 지정) 이상 함께 등장하는 단어들을 구한 결과로서 상위 단어와 연관성이 많은 단어라 할 수 있다. Table 4에서 '농업', '동아리', '친구'는 복수의 문항에서 보이지만 문항별 연관 단어들을 보면 매우 다른 결과를 보이고 있어 각 문항이 요구하는 특성에 부합하는 단어들이 사용된 것을 알 수 있다.

Table 3. Word cloud by TF-IDF analysis

문항 1	문항 2
	
문항 3	문항 4
	

Table 4. The associative words obtained by the findAssocs() function among the specific keywords

문항	키워드	연관 단어
1	농업	품종, 발전, 농촌, 미래, 작물, 경영, 농업기계, 모내기, 6차산업, 병해충, 생산물, 유기농업
	수학	문제, 공식, 풀이, 계산, 유형, 암기과목, 오답노트, 자신감, 개념, 성적, 과목, 행사
	공부	성적, 과목, 집중, 기말고사, 시험, 암기, 국어, 학원, 고사
	문제	수학, 유형, 풀이, 기본개념, 해결, 오답노트, 난이도, 몰입, 속도, 공식, 기초, 개념, 나열,
	자격증	취득, 합격, 시험, 산림기능사, 필기시험, 종자기능사, 실기시험, 유기농업기능사, 취업
2	동아리	부원, 가입, 활동, 생육, 배드민턴동아리, 선배, 하교, 기장, 부회장, 텃밭, 악기, 회원
	식물	악영향, 이름, 동식물, 화분, 잡초제거, 무관심, 피해, 공기정화식물, 사진, 장비, 친환경
	친구	선생님, 학급, 청소시간, 응원, 용기, 이야기, 친절, 눈물, 성숙, 고등학교, 자신감, 체육대회,
	농업	농촌, 작물, 농민, 미래, 발전, 경쟁력, 농업계고등학교, 농업기술, 농업경영인, 4차산업
	작물	수확, 재배, 텃밭, 방제, 심기, 농업, 호박, 살포, 고랑, 해충, 관리, 병해충, 애정, 날씨
3	친구	친밀, 마음, 부족, 사이, 사과, 장난, 중학교, 표정, 실습
	동아리	부원, 활동, 가입, 멤버, 운영, 회원, 고추, 열정, 동아리활동, 수확, 실습장, 부스운영, 열정
	의견	조율, 충돌, 각자, 결정, 안건, 회의, 동의, 주장, 토론, 제시, 결과, 피드백, 반대, 찬성
	학생	학생회, 회장, 교수님, 안건, 학교, 출마, 선도부원, 학과, 건의, 교복, 농업인, 선도부, 전교
	선생님	담임, 복도, 하교, 예의, 종례, 교문, 벌점, 허락, 고등학생, 교장, 교실, 교체, 적용, 뿌듯
	봉사활동	봉사, 할머니, 거동, 노인, 어깨, 불편, 요양원, 장애인, 말벗, 반성, 사랑, 자원봉사, 즐거움
4	버섯	버섯학과, 버섯재배, 배지, 종균, 버섯종균기능사, 원목, 표고버섯, 약용, 배양, 느타리버섯
	곤충	산업곤충학과, 장수풍뎅이, 곤충산업, 귀뚜라미, 채집, 애완동물, 식용, 먹이, 사육, 산업곤충
	아버지	휴가, 후회, 이사, 장어, 제안, 진로계획, 양계, 인문계, 공무원, 양식업, 주말, 고민, 생물
	농업	농업인, IT, 청년농부, 농법, 마이스터, 식량작물학과, 작물, 6차산업, 드론, 마늘, 미래농업
	농장	여름방학, 운영, 양돈학과, 가금류, 돼지, 삼촌, 경영, 심화과정, 청년농업인, 가축, 양돈
	한우	한우학과, 농축산업, 한우농가, 고급육, 송아지, 축산기사, 조사료, 축사, 축산업, 농축산

2. 언어네트워크분석

Fig. 2 ~ Fig. 5는 중심성분석 결과를 시각화한 문항별 네트워크 다이어그램(Network diagram)이다. 분석을 위한 TDM 메트릭스 계산에서는 시각화를 위하여 앞에서 설명한 바와 같이 적절한 문서 수를 설정하였으며, 추가로 상관행렬에 계수 0.18(문항 1, 3), 0.2(문항 2), 0.17 (문항 4)를 각각 곱하여 edge 수를 조절하였다.

중심성분석에서는 노드 간 최단 경로 수를 측정하는 방법인 매개중심성 값을 구하고 상위 10% 노드를 노랑색으로 나타냈다. 노드 크기는 노드에 연결된 다른 노드의 개수에 따른 중요도 측정 방법인 연결정도중심성 값으로 표현하였다. 또한 노드와 노드를 이어주는 edge의 굵기는 상관도(Edge weight)가 클수록, 즉 상관관계가 클수록 노드 간의 edge를 굵게 나타냈다. 상관도는 상관계수 행렬에서 상관관계를 구한 값으로 0에 근접하지 않으면 상관성이 있다고 볼 수 있으며, Table 5에 문항별 상위 5위까지의 분석 결과를 나타낸다.

먼저 Table 5를 보면 '자격증 - 취득' 단어 간 상관관계가 문항 1, 문항 2, 문항 4 모두에서 가장 높게 나타났으며, '문제 - 해결' 단어 조합은

문항 2, 문항 3, 문항 4에서 상관관계가 높게 사용되는 것을 알 수 있다. 문항별 단어 간 조합을 보면 문항 1의 학업 역량, 문항 2의 전공 적합, 문항 3의 인성 및 갈등 해결, 문항 4의 대학 지원동기, 학업 및 진로 계획 등을 적합하게 서술하는 단어들이 서로 묶여 있어 상관관계가 높은 것을 알 수 있다.

Table 5. Edge weight between keywords by item

문항	순위	키워드	Edge weight	문항	순위	키워드	Edge weight
1	1	자격증 - 취득	0.832834	3	1	문제 - 해결	0.569529
	2	과학 - 생명	0.659376		2	오해 - 양보	0.545038
	3	작물 - 재배	0.529543		3	학교축제 - 축제	0.536491
	4	단어 - 해석	0.460611		4	공부 - 성적	0.511506
	5	영어 - 단어	0.454088		5	담임 - 선생님	0.485418
2	1	자격증 - 취득	0.871278	4	1	자격증 - 취득	0.751598
	2	문제 - 해결	0.608384		2	문제 - 해결	0.624075
	3	효과 - 체험	0.476860		3	2학년 - 1학년	0.591965
	4	작물 - 수확	0.470924		4	3학년 - 1학년	0.531169
	5	1학년 - 2학년	0.462182		5	2학년 - 3학년	0.478286

Fig. 2는 문항 1에 대한 분석 결과이다. 노랑색으로 나타낸 매개중심성 상위 10% 키워드는 '이유', '고등학교', '재학', '확인', '사용', '실험', '수확' 등으로 나타났다. 연결된 노드 개수에 의한 중요도 측정으로 노드의 크기를 결정하는 연결정도중심성은 '고등학교'가 가장 높게 나타났으며, '탐구', '성적', '실험', '과학' 순으로 나타났다.

Fig. 3은 문항 2에 대한 분석 결과이다. 매개중심성 상위 10% 키워드는 '쓰레기', '고등학교', '학교', '수업시간', '청소', '회의', '친구', '성공' 등으로 나타났으며, 연결정도중심성 단어는 '쓰레기', '정리', '수업시간', '농장', '실력', '재학' 등의 순으로 나타났다.

Fig. 4는 문항 3에 대한 분석 결과이다. 매개중심성 상위 10% 키워드는 '중요', '오해', '완성', '자리', '발생', '관리', '해결', '갈등' 등으로 나타났으며, 연결정도중심성 단어는 '의견', '회의', '봉사활동', '조원', '주제', '중요', '태도', '학생회' 등의 순으로 나타났다.

Fig. 5는 문항 4에 대한 분석 결과로서 매개중심성 상위 10% 키워드는 '가공', '사료', '농가', '환경', '축산' 등으로 나타났으며, 연결정도중심성 단어는 '가공', '공간', '실습', '축산', '환경', '개발', '자연' 순으로 높게 나타났다.

Fig. 2 ~ Fig. 5에서 매개중심성 상위 5위까지 키워드를 붉은색 원으로 표시하였으며, 연결정도중심성이 매우 높은 키워드 조합을 파란색 타원으로 나타냈다. 이들 네트워크 다이어그램을 보면 매개중심성이 높을수록, 즉 단어의 사용 횟수가 많을수록 네트워크의 중앙에 위치하고, 두 노드 사이의 관계가 강할수록 서로 근거리에 위치하는 경향이 있는 것을 알 수 있다. 또한 연결정도중심성이 클수록 노드의 크기가 크게 나타나며, 또한 노드를 연결하는 선은 단어들이 함께 사용되었을 때 하나의 선으로 연결되는 것으로 단어들의 동시 출현 빈도가 높을수록 선이 굵게 표시되는 것을 알 수 있다.

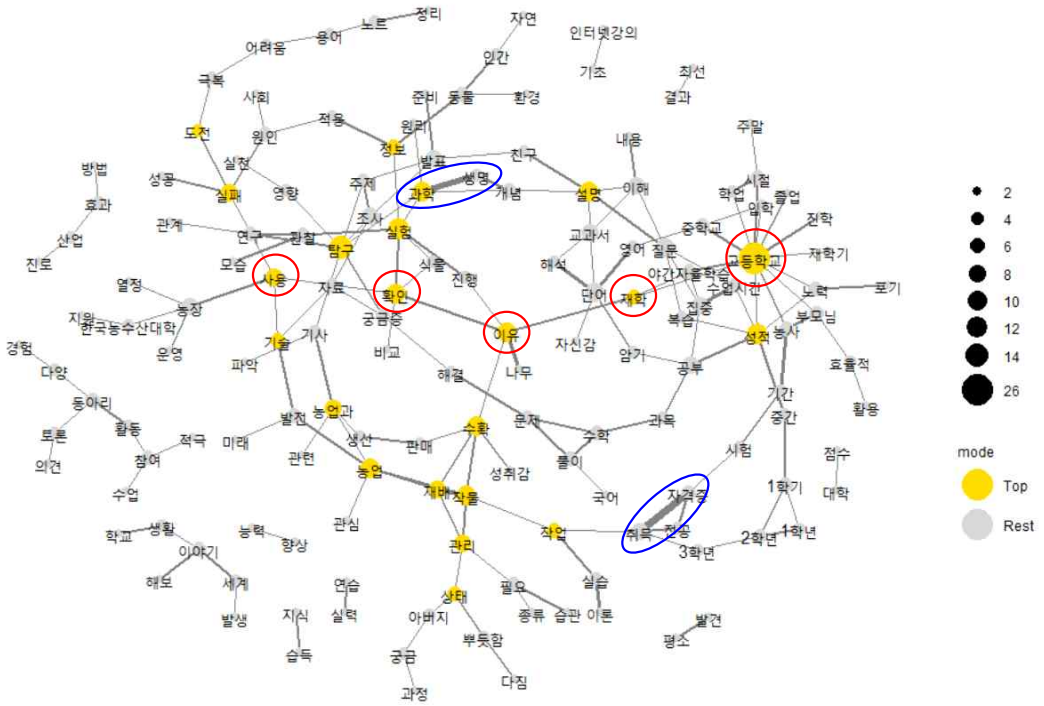


Fig. 2. Semantic network map by betweenness centrality and degree centrality to Q. 1

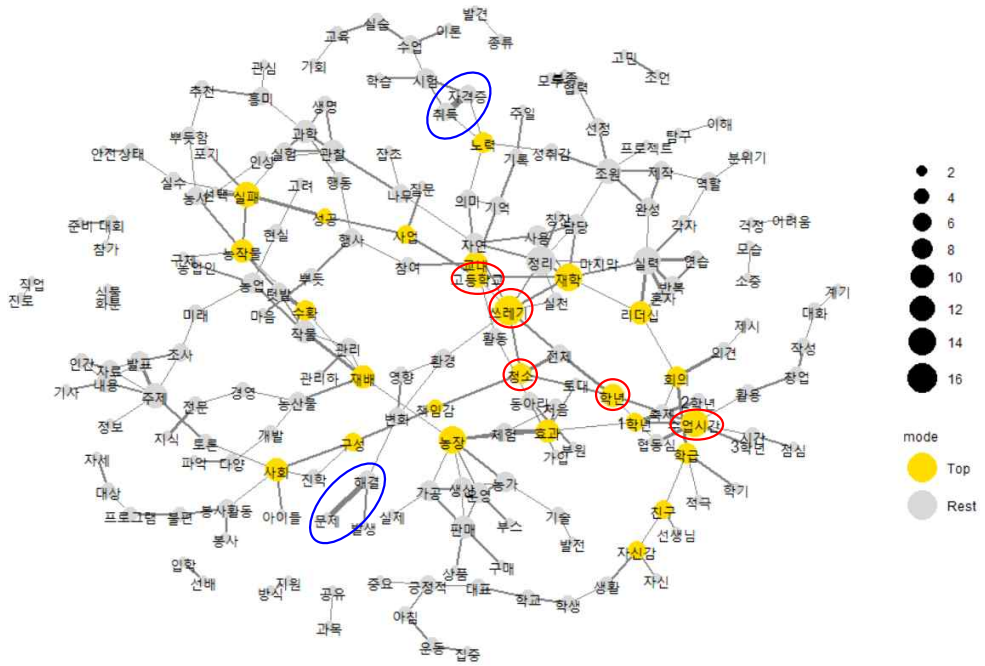


Fig. 3. Semantic network map by betweenness centrality and degree centrality to Q. 2

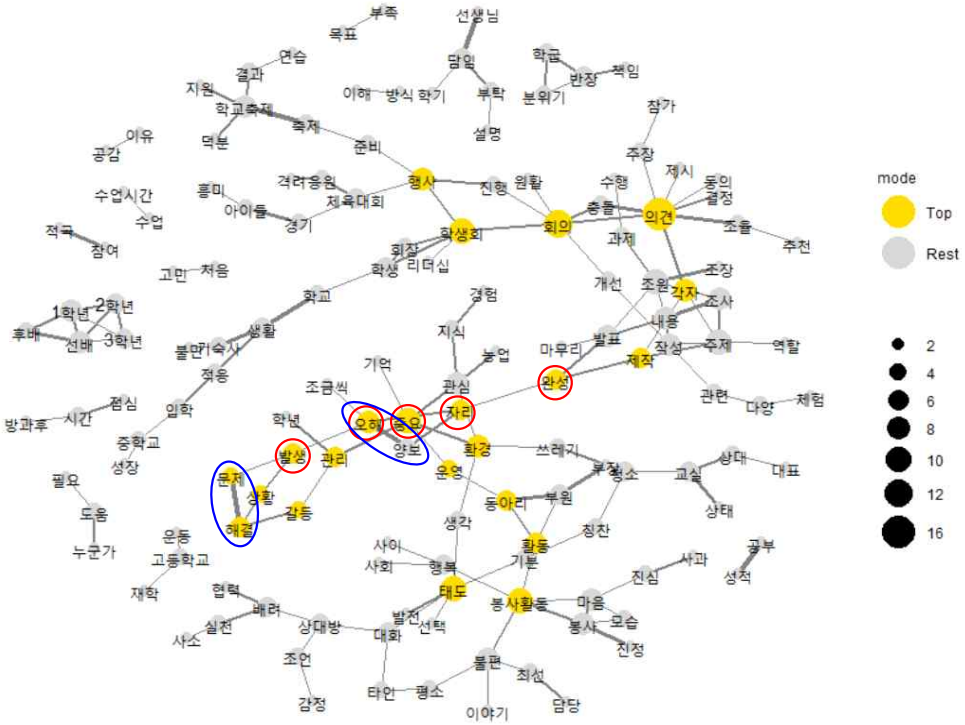


Fig. 4. Semantic network map by betweenness centrality and degree centrality to Q. 3

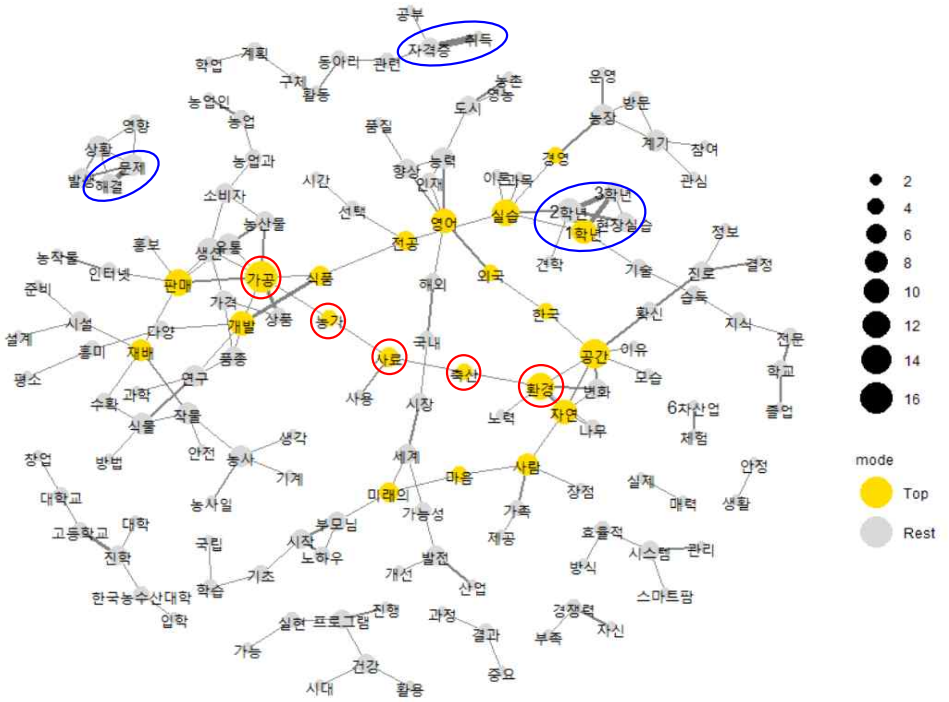


Fig. 5. Semantic network map by betweenness centrality and degree centrality to Q. 4

3. 군집분석 및 Dendrogram

자소서에 사용된 단어들에 대하여 유사성 개념에 기초한 계층적 군집분석 결과로 도출되는 덴드로그램을 문항별로 Fig. 6 ~ Fig. 9에 나타냈다. 클러스터 덴드로그램은 단어 기반의 계층적 클러스터링 기법에 의하여 단어 간 인접, 상호 관계를 계층적으로 나타낸 것이다.

그림의 세로축 'Height'는 군집화의 비유사성 지표(비유사도)를 의미하며, 최하위부터 두 개 묶여 있으면 같은 계층에 있다는 것을 의미한다. 군집의 개수 k 결정은 Ward 연결법에 의한 계층적 군집분석의 덴드로그램 시각화를 통하여 결정하였다. 덴드로그램의 높이(Height) 축을 기준으로, 위에서 아래로 높이를 이동하면서 군집 간 높이의 차이가 큰 군집의 개수 k를 선택하였다.

다만, 이 방법은 데이터 개수가 많은 경우에는 군집화 시간이 오래 걸리고, 덴드로그램으로 시각화가 거의 불가능할 수 있다. 또한 통계량을 산출하는 게 아니고 덴드로그램 시각화 결과를 보고 연구자의 판단에 의지하니 다분히 주관적일 수 있다.

그러나 유사한 단어들의 군집 간 비유사도의 차이가 큰 군집의 개수를 구하면 Fig. 6의 문항 1은 2개, Fig. 7의 문항 2와 Fig. 9의 문항 4는 4개, Fig. 8의 문항 3은 5개로 할 때 군집 내 응집력이 높고, 군집 간 이질성이 큰 적절한 군집을 구할 수 있었다. 그림에서 군집은 붉은 선으로 나누어 표시하였다. Fig. 6 ~ Fig. 9의 각 군집들은 자소서가 요구하는 학업 노력, 교내 활동, 배려 및 갈등 관리의 인성, 지원동기 및 진로계획 등 4개의 문항을 서술한 유사성 높은 단어들이 서로 가깝게 묶여 있는 것을 알 수 있다.

Ⅳ. 적요

비정형 데이터인 한국농수산대학 신입생 자기소개서에서 의미 있는 정보를 추출하기 위하여 핵심적인 역할을 하는 단어의 중요도를 평가하는 TF-IDF 가중치를 기초로 한 언어네트워크분석을 하였다.

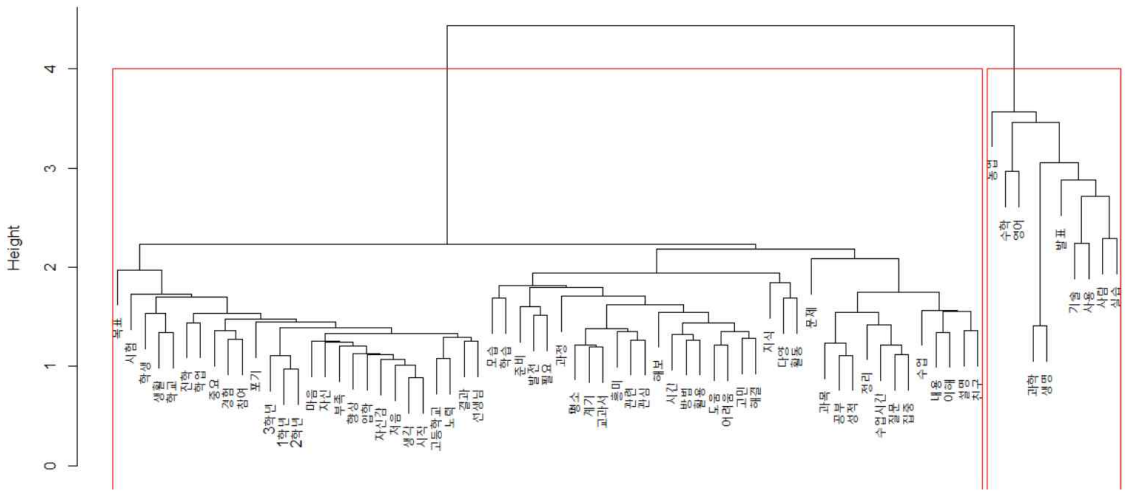


Fig. 6. Dendrogram of hierarchical clustering for question 1 (k = 2 documents)

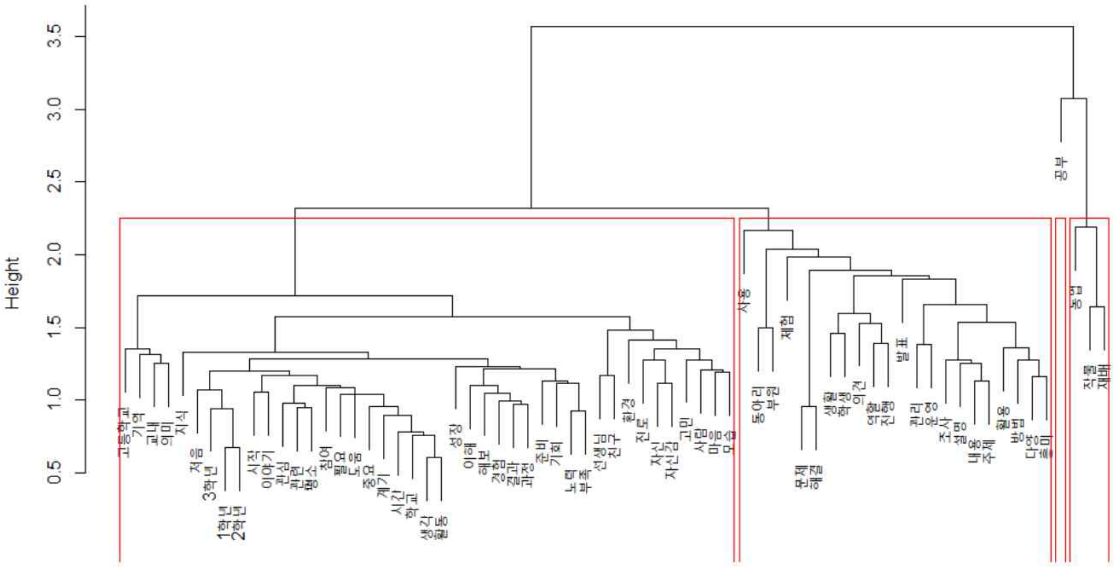


Fig. 7. Dendrogram of hierarchical clustering for question 2 (k = 4)

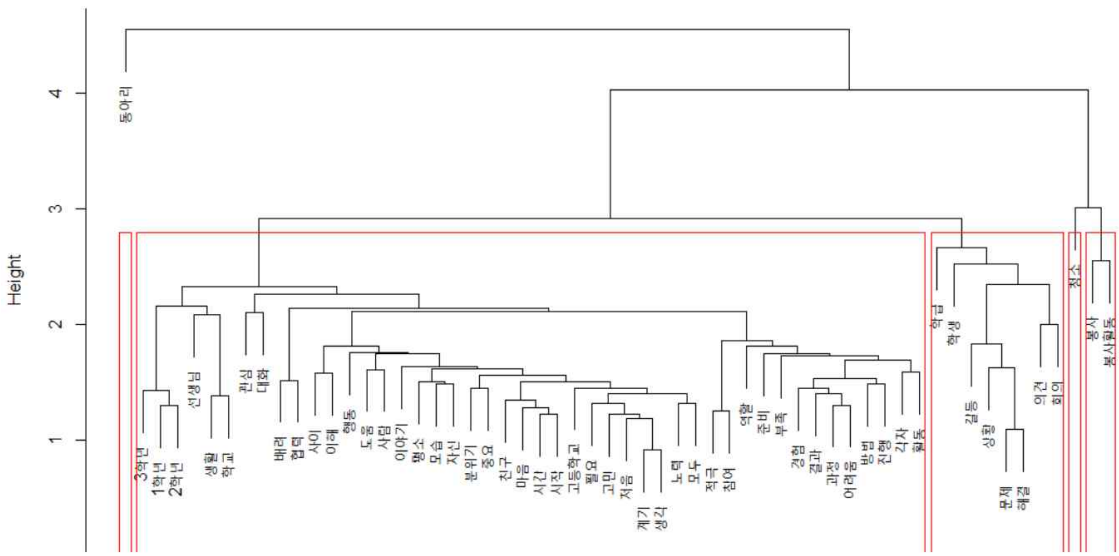


Fig. 8. Dendrogram of hierarchical clustering for question 3 (k = 5)

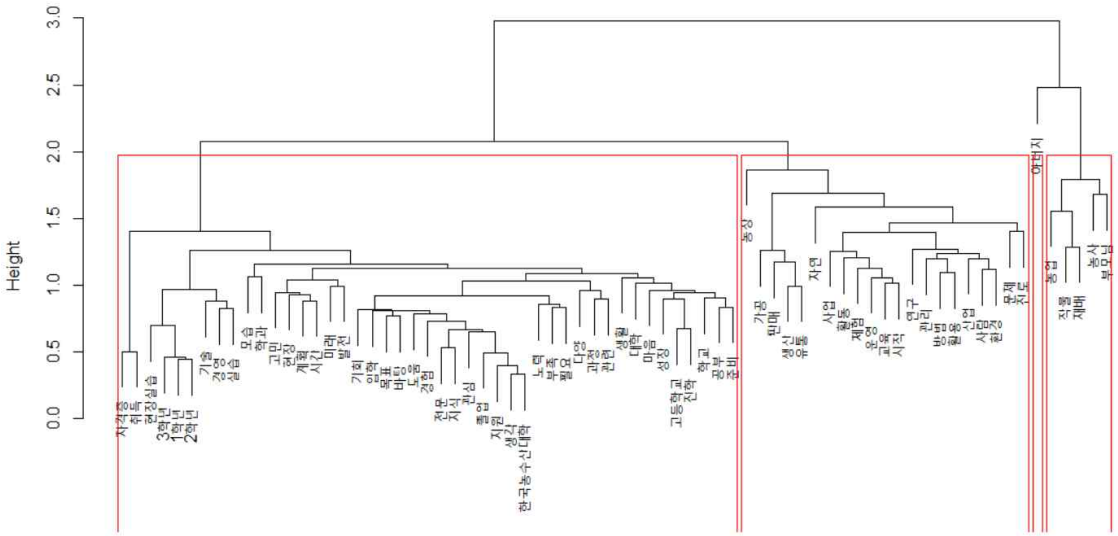


Fig. 9. Dendrogram of hierarchical clustering for question 4 (k = 4)

TF-IDF 가중치에 의한 핵심단어는 문항 1에서는 '농업', '수학', '공부', '문제', '친구', 문항 2에서는 '동아리', '식물', '친구', '농업', '작물', 문항 3에서는 '친구', '동아리', '의견', '갈등', '관리', 문항 4에서는 '버섯', '곤충', '아버지', '농업', '농장' 등으로 나타났다. 또한 빈도수는 낮은 단어이지만 핵심단어로 나타난 단어를 보면 문항 1에서는 '수학', '자격증', '성적', '영어', '과학', 문항 2에서는 '식물', '작물', '공부', '쓰레기', '발표', '실험', 문항 3에서는 '동아리', '청소', '봉사', '갈등', '봉사활동', 문항 4에서는 '버섯', '곤충', '양식', '한우', '조경' 등으로 나타났다.

단어들 간의 관계를 시각적으로 분석이 가능한 언어네트워크분석 결과 매개중심성이 높은 단어는 문항 1에서는 '이유', '고등학교', '재학', 문항 2에서는 '쓰레기', '고등학교', '학교', 문항 3에서는 '중요', '오해', '완성', 문항 4에서는 '가공', '사료', '농가'로 나타났다. 연결정도중심성은 문항 1에서는 '고등학교', '탐구', '성적', 문항 2에서는 '쓰레기', '정리', '수업시간', 문항 3에서는 '의견', '회의', '봉사활동', 문항 4에서는 '가공', '공간', '실습'으로 나타났다. 매개중심성 값이 클수록 네트워크의 중앙에 위치하고, 두 범주 사이의 관계가 강할수록 서로 근거리에 위치한다. 연결정도중심성이 클수록 노드의 크기가 크게 나타나며, 노드 연결선은 단어들의 동시 출현 빈도가 높을수록 edge가 굵게 나타났다. 동시 출현 빈도가 높은 즉 상관관계가 높은 단어 조합은 '자격증 - 취득', '문제 - 해결', '과학 - 생명', '오해 - 양보' 등으로 나타났다.

단어 기반의 계층적 클러스터링 기법에 의하여 단어 간 인접, 상호 관계를 계층적으로 나타낸 클러스터 덴드로그램으로 군집의 개수를 결정하였다. 단어들의 군집 간 비유사도의 차이가 큰 군집을 구한 결과 문항 1은 2개, 문항 2와 문항 4는 4개, 문항 3은 5개의 군집으로 분류할 경우 군집 내 응집력이 높고, 군집 간 이질성이 큰 적절한 군집을 구할 수 있었다.

Ⅴ. 참고문헌

1. 김경태, 안정국, 김동현. (2018). 빅 데이터 활용서 (Ⅰ). 시대인.
2. 김영우. (2017). 쉽게 배우는 R 데이터 분석, 이지스퍼블리싱.
3. 나종화. (2017). R 데이터마이닝, 자유아카데미.
4. 남길임, 조은영. (2017). 한국어 텍스트 감성 분석, 커뮤니케이션북스.
5. 조민호. (2019). 데이터 분석 전문가를 위한 R 데이터 분석. 정보문화사
6. 김상아, 강정배, 변찬석. (2015). 언어네트워크 분석(Semantic Network Analysis)을 이용한 국내 학습장애 연구 동향 분석. 특수교육재활과학연구 Vol. 54, No. 2, pp. 449~471.
7. 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
8. 서유빈, 이혜련, 윤유식, 김미성. (2020). 언어네트워크분석을 활용한 지역특화컨벤션 이해관계자의 인식 연구. MICE관광연구 제20권 제1호(통권 제59호) pp.71-90.
9. 서현진, 최영현, 오승택, 이규혜. (2019), RJCC 연구 키워드 네트워크. 복식문화연구 Vol.27, No.3, pp.193-205.
10. 주진수 외 5인. (2020). 한국농수산대학 신입생 자기소개서의 텍스트 마이닝과 연관규칙 분석 (1). 현장농수산연구지 Vol. 22(1), No.1: 113-130.
11. 주진수 외 5인. (2020). 한국농수산대학 신입생 자기소개서의 텍스트 마이닝과 연관규칙 분석 (2). 현장농수산연구지 Vol. 22(2), No.2: 99-114.
12. https://blog.naver.com/pasudo123/221064463377. TF-IDF
13. http://blog.naver.com/PostView.nhn?blogId=happyfox098&logNo=221210945008. 패스트캠퍼스 R 프로그래밍을 통한 텍스트 마이닝 이론과 연관 키워드 분석 4주차
14. https://bookdown.org/yuaye_kt/RTIPS/Texnetword.html. Chapter 11 텍스트 데이터-단어 네트워크맵(1)
15. https://briatte.github.io/ggnet/. ggnet2: network visualization with ggplot2
16. https://da-it-so.tistory.com/43. TF-IDF 기법 이해하기
17. https://data-traveler.tistory.com/33. R을 이용한 텍스트마이닝_TF-IDF(코드 및 설명)
18. https://iamdaisy.tistory.com/31?category=620658. 소셜네트워크 분석의 이해
19. https://rfriend.tistory.com/585 [R] 군집분석 군집의 개수 k 결정 방법
20. https://thinkwarelab.wordpress.com/2016/11/14/ir-tf-idf-%EC%97%90-%EB%8C%80%ED%95%B4-%EC%95%8C%EC%95%84%EB%B4%85%EC%8B%9C%EB%8B%A4/. [IR] tf-idf 에 대해 알아봅시다
21. http://www.datamarket.kr/xe/board_BoGi29/6479. social network analysis
22. http://www.kateto.net/wp-content/uploads/2015/06/Polnet%202015%20Network%20Viz%20Tutorial%20-%20Ognyanova.pdf. Network visualization with R

논문접수일 : 2021년 5월 13일
논문수정일 : 2021년 6월 4일
게재확정일 : 2021년 6월 14일