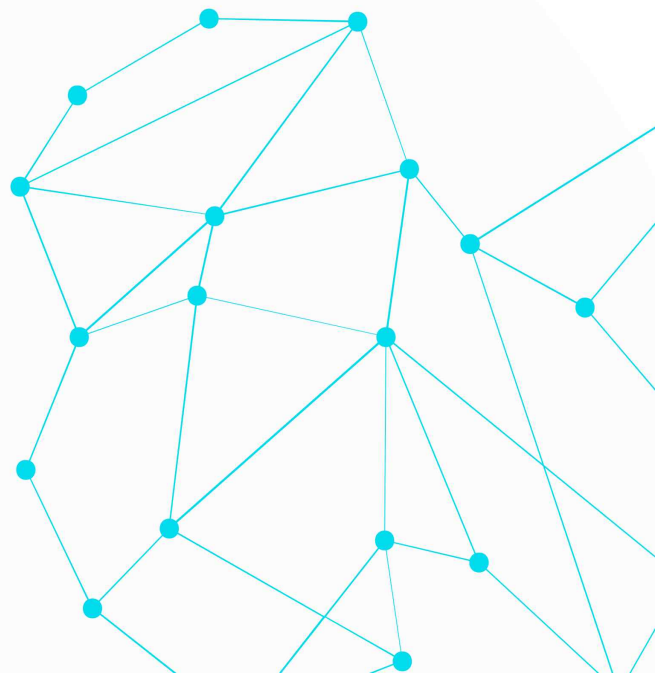


TEXTOM

# 텍스툼 분석 보고서 분석 의뢰 참고 자료



## 목 차

<b>1. 분석 설계</b>	<b>3</b>
1) 개념적 정의	
2) 분석 도구 (TEXTOM)	
3) 분석 절차	
<b>2. 데이터 수집</b>	<b>4</b>
1) 데이터 수집 범위	
2) 수집 키워드 선정	
<b>3. 데이터 정제/전처리</b>	<b>5</b>
1) 키워드 및 기계학습 기반 스팸 데이터 처리	
2) 텍사노미 구축	
<b>4. 데이터 분석</b>	<b>6</b>
1) 자료량 분석	
2) 단어 분석	
3) 의미 네트워크 분석(Semantic Network Analysis)	
4) 중심성 분석(Centrality Analysis)	
5) CONCOR 분석 (CONvergence of iterated CORrelations Analysis)	
6) 클러스터링 분석 (Clustering Analysis)	
7) 토픽 분석	
8) 감성 분석 (Sentiment Analysis)	

## (TEXTOM 분석 보고서) 분석 의뢰 참고 자료

(2025.ver)

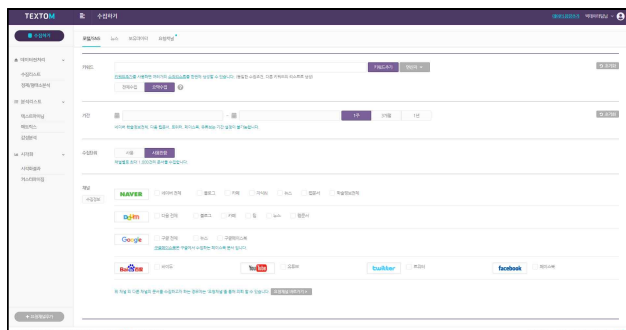
### □ 분석 설계

#### ○ 개념적 정의

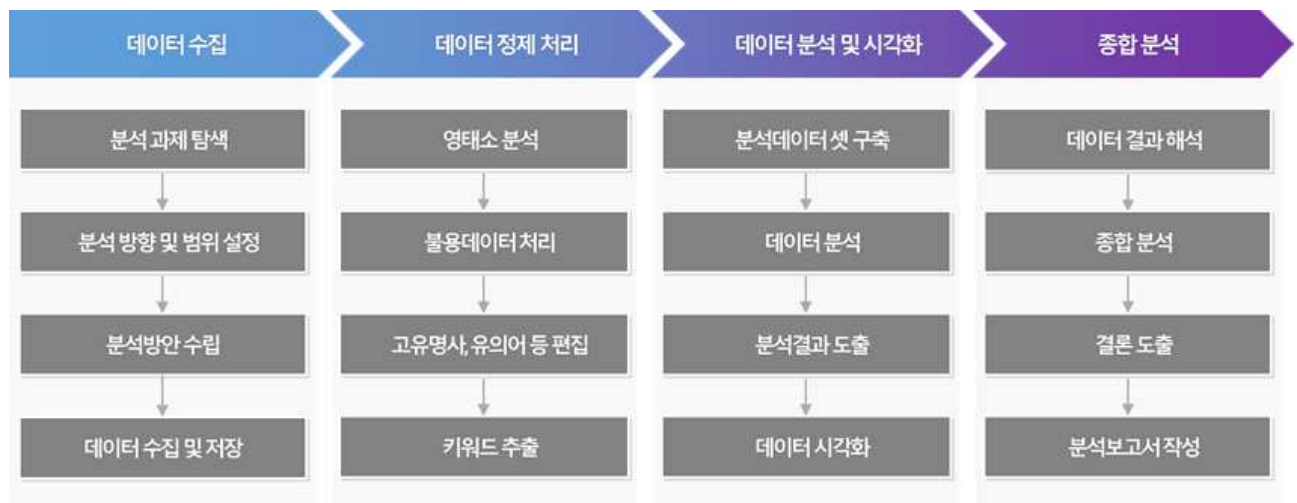
- 분석 목표를 바탕으로 고객과의 협의를 통해 분석 방향성을 확정하고, OO 분야에 대한 데이터 및 분석 범위를 도출

#### ○ 분석 도구 (TEXTOM)

- 텍스툼(TEXTOM)은 (주)터아이엠씨에서 개발한 빅데이터 일관처리 솔루션으로, 네이버, 다음, 구글, 바이두 등 국내외 포털을 비롯해 페이스북, 트위터, 유튜브 등 다양한 소셜미디어 채널의 데이터 수집 기능을 보유하고 있음
- 소셜미디어, 뉴스, 사용자 보유 데이터 등 텍스트 데이터에 대한 텍스트 마이닝과 키워드 분석, 연결망 분석, 토픽 모델링 등 다양한 분석을 수행할 수 있는 텍스트 분석 전문 플랫폼임



### ○ 분석 절차



## □ 데이터 수집

### ○ 데이터 수집 범위

- 수집 기간 : 20XX.00.00. ~ 20XX.00.00. (00개월)
- 수집 채널 : 뉴스, 블로그, 카페, 학술정보, 커뮤니티, 유튜브, Reddit, 스마트스토어 등
- 수집 내용 : 제목, 본문, 생성일, URL 등

< 텍스트 데이터 수집 채널 예시 >

구분	수집 채널		수집 내용
1	네이버	뉴스	제목, 본문, URL, 언론사명
		블로그	제목, 본문, URL
		카페	제목, 본문, URL
		웹문서	제목, 본문, URL
		지식iN	제목, 본문, URL
2	다음	뉴스	제목, 본문, URL, 언론사명
		블로그	제목, 본문, URL
		카페	제목, 본문, URL
		웹문서	제목, 본문, URL
3	구글 (한국)	뉴스	제목, 본문, URL, 언론사명
		구글페이스북	제목, 본문, URL
		웹문서	제목, 본문, URL
4	RISS	국내학술논문	제목, 연도, 주제어, 국문초록, 다국어초록, 저자 등
5	구글 (USA)	뉴스	제목, 본문, URL, 언론사명
		구글페이스북	제목, 본문, URL
		웹문서	제목, 본문, URL
6	Reddit (USA)		제목, 본문, URL 등
7	요청채널		제목, 설명, URL 등

### ○ 수집 키워드 선정

- 데이터 수집을 위한 수집 키워드 탐색 조사 진행
- 수집 키워드(안)을 대상으로 수집량 및 샘플 데이터 분석 진행하여, 최종 수집키워드 선정

< 수집 키워드 선정 예시 >

구분	대분야	수집 키워드
1	다이어트	"다이어트"
		"diet"
		"다이어트"+"식단"
2	건강기능식품	"건강기능식품"
		"건기식"
		"영양제"

## □ 데이터 정제/전처리

### ○ 키워드 및 기계학습 기반 스팸 데이터 처리

- 수집된 데이터에서 분석에 불필요한 수집된 데이터에서 분석에 불필요한 데이터를 제거하기 위해 텍스트 마이닝 기법과 SVM(Support Vector Machine)을 적용한 문서 분류 기법 활용
- 스팸 및 불용 데이터 처리 기술을 적용하여, 가짜 이슈를 필터링함으로써 타당성과 신뢰성이 확보된 분석 데이터 구축

### ○ 택사노미 구축

- 분석 목적에 대응하는 핵심어 사전(택사노미, Taxonomy)을 구축하고 일괄적 적용을 통해 데이터 분석의 효율성과 신뢰도를 제고함
- OO 분야에 특화된 분류체계를 정의하고, 해당 분류에 대한 택사노미를 구축해 분석에 적용
- 분석 단어 선정 과정 및 분류 과정에서 작업자의 기준에 따라 차이가 날 수 있는 측면을 방지하고, 공통 기준 마련으로 데이터의 특성·분량에 관계없이 일관된 어휘 집단에서의 어휘 추출 진행

< 핵심어사전 구축 예시 >

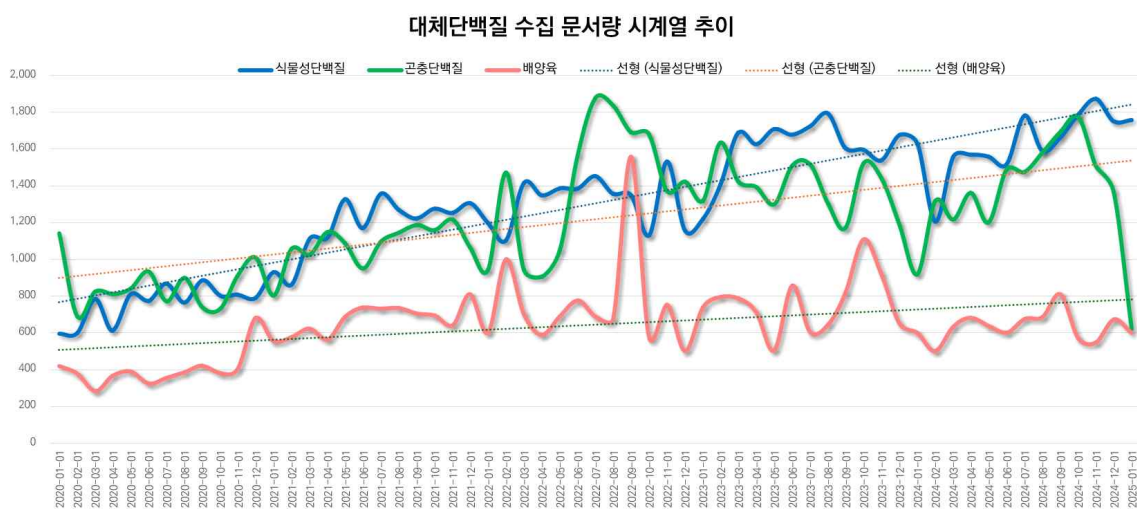
분류(카테고리)	정의
이용 대상자	이용 대상자를 지칭하는 키워드 (개인, 단체, 혼자, 가족, 연인, 친구, 학교, 유치원, 교사 등)
지역/장소	지역 혹은 장소를 지칭하는 키워드 (국립수목원, 국립세종수목원, 천리포수목원, 홍릉숲 등)
방문 목적	기관에 방문하는 목적과 관련된 키워드 (여행, 축제, 전시회, 관람, 산책, 힐링, 여행지, 치유, 감상, 휴양 등)

## □ 데이터 분석

### ○ 자료량 분석

- 수집된 빅데이터의 심층 분석에 앞서 기초 분석을 수행하여 OO 분야에 대한 현황 파악
- 채널별로 자료량을 시간의 흐름에 따라 일정한 간격마다 기록하여 OO 분야에 대한 관심도 증감 추이를 확인, 월별 자료량 분석을 통해 계절적 특성, 사건/사고 등 영향 요인을 분석

< 자료량 분석 예시 >

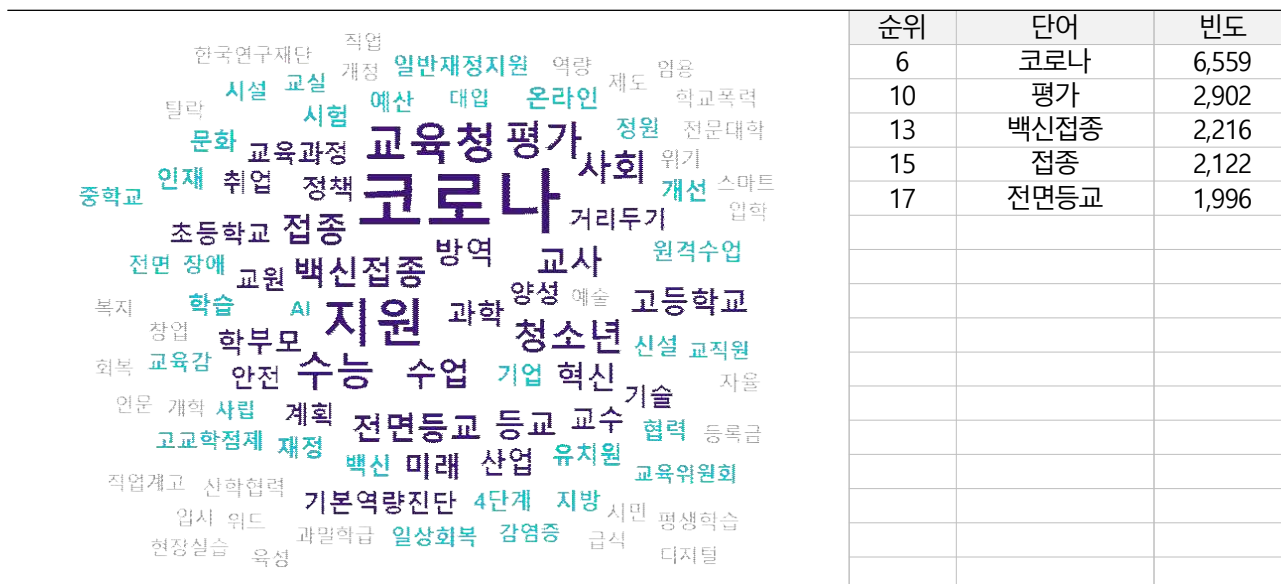


기간	2019년	2020년	2021년	합계
1월	20,421	19,405	21,170	60,996
2월	18,796	18,983	18,610	56,389
3월	20,582	19,118	20,097	59,797
4월	17,850	18,879	19,061	55,790
5월	21,450	18,367	19,628	59,445
6월	19,472	19,666	20,688	59,826
7월	20,850	19,360	20,597	60,807
8월	20,023	18,415	18,709	57,147
9월	19,651	18,994	19,502	58,147
10월	20,651	19,306	18,950	58,907
11월	19,006	19,289	19,335	57,630
12월	20,926	20,167	20,543	61,636
합계	239,678	229,949	236,890	706,517

## ○ 단어 분석

- (빈도 분석) 텍스트마이닝을 통해 수집된 텍스트 데이터에서 출현하는 단어의 빈도수, TF-IDF 수치를 활용한 시계열 분석을 통해 이슈성, 감성 등 시사점을 내포하고 있는 키워드를 도출하고 해당 키워드의 의미를 분석

&lt; 빈도 분석 예시 &gt;



- '코로나' 관련 단어의 등장이 직전 구간 대비 크게 감소(7위→18위)하고, '반도체'는 직전 구간 대비 크게 상승(24위→15위)하였음
- '입학'(10위), '초등학교'(13위), '연령'(33위) 등 초등학교 입학 연령 하향에 대한 학제개편 계획 발표와 이에 대한 비판 보도가 이어지면서 상위 40위 안에 주요 키워드가 위치함

2019년			2020년			2021년		
순위	단어	빈도	순위	단어	빈도	순위	단어	빈도
1	교통	27,987	1	교통	28,176	1	교통	30,035
2	지하철	21,489	2	지하철	21,225	2	지하철	18,174
3	버스	20,232	3	버스	16,532	3	고속열차	17,962
4	고속열차	19,356	4	역세권	15,489	4	수소	16,297
5	고속	13,382	5	고속열차	13,519	5	연결	15,450
6	역세권	13,146	6	연결	13,010	6	역세권	14,959
7	안전	12,945	7	KTX	11,352	7	버스	14,775
8	연결	11,994	8	수소	9,851	8	KTX	12,641
9	KTX	11,433	9	고속	9,828	9	고속	11,896
10	대중교통	8,884	10	안전	9,268	10	안전	10,468

### ○ 의미 네트워크 분석 (Semantic Network Analysis)

- 의미 네트워크 분석은 구조주의적 관계를 바탕으로 행위나 의미를 해석하고 파악하는 방법으로, 언어 및 해석학적으로 담론분석이라고도 함
- 네트워크 내 요소(키워드) 간의 관계 구조를 분석하여 특정 키워드의 영향력과 담론에서의 역할을 확인할 수 있음. 이를 위해 중심성 분석, CONCOR 분석, 클러스터링 분석 등을 수행함

### ○ 중심성 분석 (Centrality Analysis)

- 네트워크 구조적 특징을 수리적으로 파악하기 위해 사회연결망 이론에 근거한 다양한 중심성 지표를 활용함. 주요 네트워크 지표로는 연결중심성(Degree Centrality), 매개중심성(Betweenness centrality), 근접중심성(Closeness centrality) 등이 있음
- OO 분야의 인식에 영향을 미치는 핵심 키워드를 식별하고, 이들의 역할과 위치를 분석함. 이를 통해 담론에서 중요한 키워드가 어떤 방식으로 연결되고 있는지 파악하여, 영향력 있는 개념(키워드)과 논의의 흐름을 효과적으로 분석함

< 중심성 분석 결과 예시 >

순위	단어	연결정도	매개중심성	근접중심성	페이지랭크
1	교육	58	65.686	1.000	0.038
2	산림	56	54.877	0.967	0.038
3	숲	52	29.848	0.906	0.036
4	산림청	48	19.985	0.853	0.035
5	전문가	48	24.223	0.853	0.036
6	과정	46	22.689	0.829	0.035
7	부산	42	12.227	0.784	0.034
8	프로그램	42	13.131	0.784	0.034
9	기관	40	9.663	0.763	0.034
10	해설가	40	10.448	0.763	0.034

### ○ CONCOR 분석 (CONvergence of iterated CORrelations Analysis)

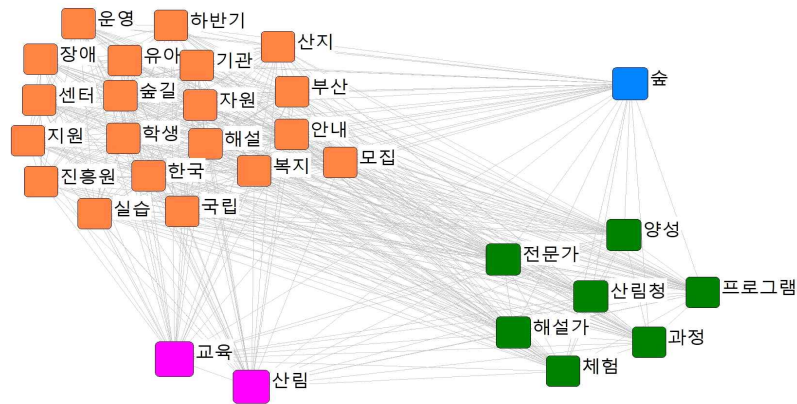
- 공출현 단어 간의 구조적 등위성을 분석하여, 노드 간의 관계 패턴의 유사성을 반복적인 상관관계 분석을 통해 유사한 역할이나 위치를 가진 노드들을 동일한 군집으로 도출하는 기법임 (활용 도구 : 텍스툼, Ucinet6 등)

### ○ 클러스터링 분석 (Clustering Analysis)

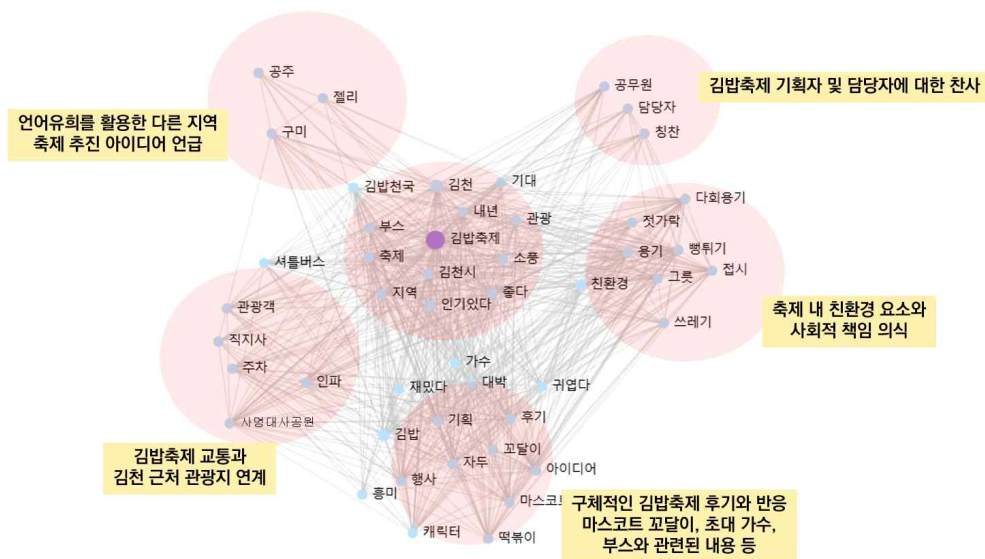
- 클러스터링 분석은 네트워크 내에서 유사한 관계 패턴을 갖는 노드를 군집화하는 기법으로, 소셜 네트워크 분석에서 많이 활용됨. 단어 빈도수를 기반으로 군집을 나누는 것이 아니라, 네트워크 내에서 노드들이 서로 연결되는 방식까지 고려함 (활용 도구 : 텍스툼, NodeXL, Python 등)



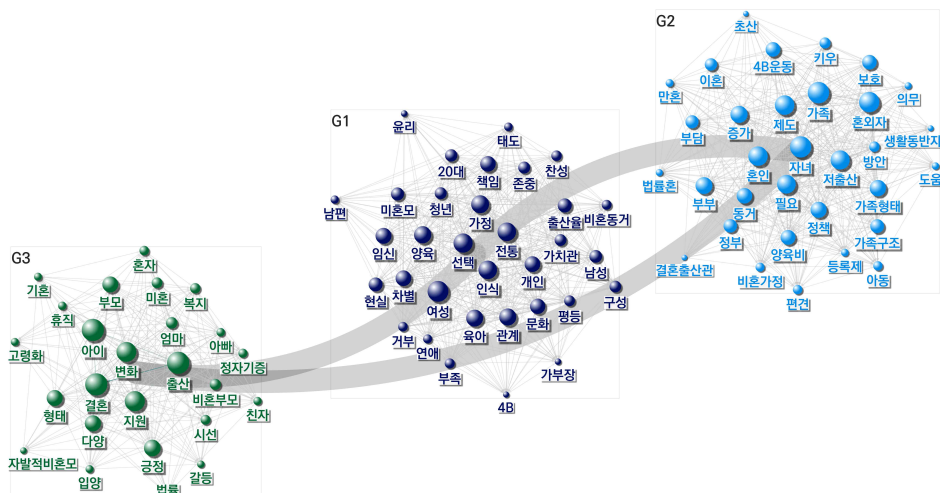
< CONCOR 분석 시각화 예시 (Ucinet6) >



< CONCOR 분석 시각화 예시 (TEXTOM) >



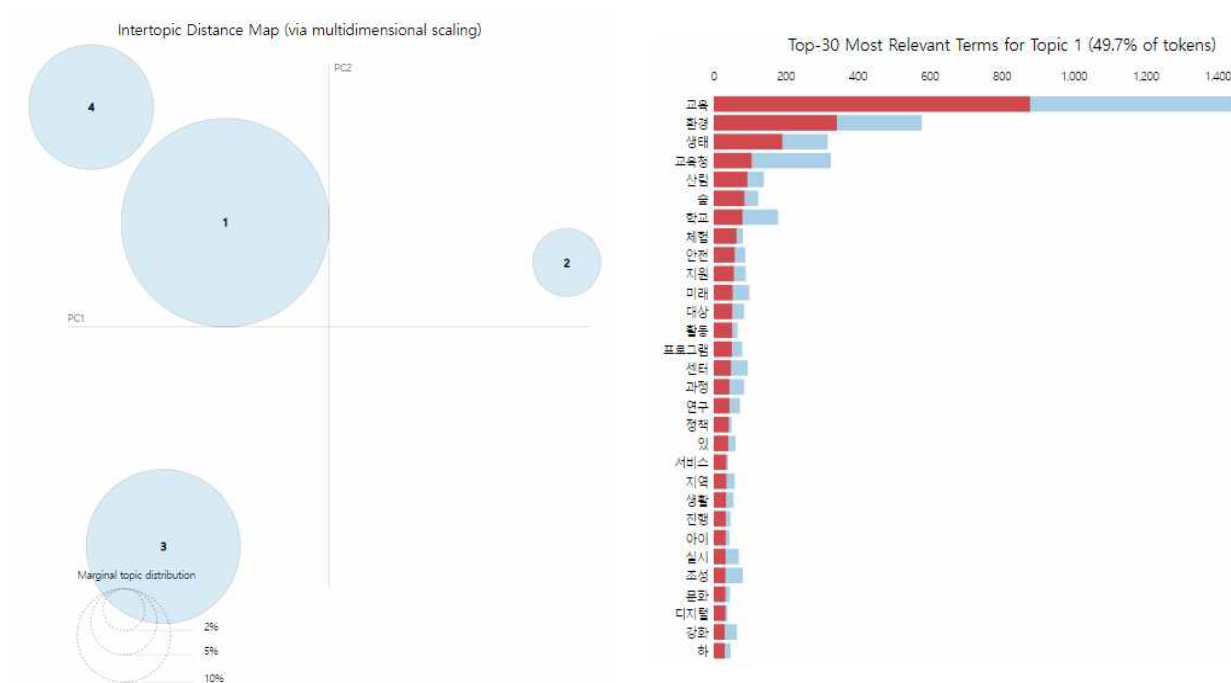
< 클러스터링 분석 시각화 예시 (NodeXL) >



## ○ 토픽 분석

- LDA 토픽모델링은 대량의 문서에서 자동으로 주제를 추출하는 알고리즘으로, 문서 내 단어 출현 확률을 반복적으로 계산하여, 특정 주제와 관련성이 높은 단어들을 도출하고 문서의 맥락을 분석함
- 람다( $\lambda$ )값 조정을 활용해 토픽별 구성 키워드를 재배열함으로써 토픽 분별력을 높여 유의미한 함의 도출할 수 있음

< LDA 토픽모델링 시각화 예시 >



< LDA 토픽모델링 결과 예시 >

	토픽	주요 키워드
1	교육 체험 토픽 ( $\lambda = 1.0$ / 49.7%)	교육, 환경, 생태, 산림, 숲, 체험 등
2	코로나19 관련 토픽 ( $\lambda = 0.8$ / 5.3%)	교육, 코로나, 위기관리, 교육청, 학교 등
3	교육기관 관련 토픽 ( $\lambda = 1.0$ / 27.2%)	교육, 교육청, 학교, 경기도, 사업, 교육원 등
4	교육 프로그램 전문가 토픽 ( $\lambda = 0.6$ / 17.8%)	교육, 환경, 양성, 숲, 모집, 숲길등산지도사 등

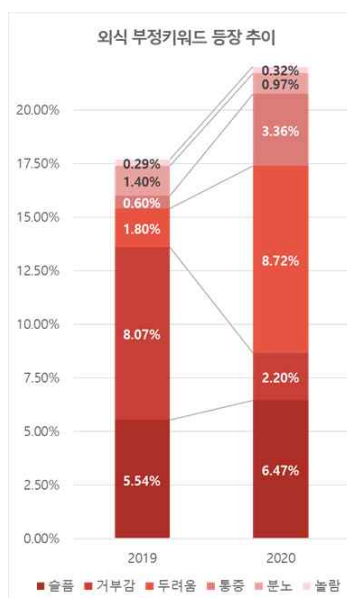
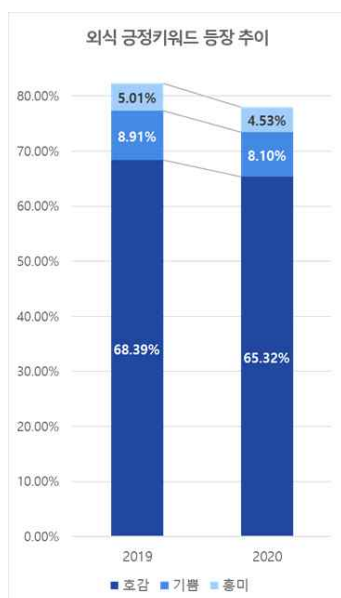
## ○ 감성 분석 (Sentiment Analysis)

- 감성분석은 텍스트에 나타난 사람들의 태도, 의견, 성향과 같은 주관적인 데이터를 분석하는 자연어 처리 기술을 말함. 텍스트 감성 분석은 크게 감성 어휘 사전 기반 분석과 문서의 감성 분류 기법(기계학습)이 활용됨
- 감성분석을 통해 언어 표현의 극성(polarity) 분석이 자동으로 빠른 시간에 처리될 수 있어, 언어학적 사고체계와 정보 기술 분야에서 각광받고 있음
- **(감성 어휘 사전)** 텍스트에서 자체 제작한 감성어휘사전을 이용하여, 감성 어휘를 분류하고 등장 빈도를 계산함. 감성어휘는 긍정과 부정이라는 카테고리 안에, 흥미, 호감, 기쁨, 통증, 슬픔, 분노, 두려움, 놀람, 거부감 등 9개 세부 감성으로 구성됨

< 사전 기반 감성 분석 시각화 예시 >



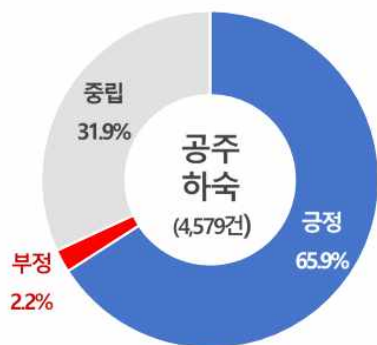
긍/부정	키워드	빈도수
긍정	혁신적	116
긍정	새롭다	107
긍정	최고다	96
긍정	기대하다	91
부정	충격	90
긍정	성장하다	90
긍정	바르다	89
부정	난감하다	78
부정	난해하다	73
긍정	현대적	67



구분	빈도(건)	비율(%)
긍정 합계	36,495	82.31
호감	30,324	68.39
기쁨	3,949	8.91
흥미	2,222	5.01
부정 합계	7,844	17.69
슬픔	2,456	5.54
거부감	3,579	8.07
두려움	796	1.80
통증	265	0.60
분노	620	1.40
놀람	128	0.29
총계	44,339	100.00

- (문서 감성 분류) 긍정/부정/중립의 감성으로 자동 분류하는 기법으로, 기계학습을 활용해 감성 패턴을 학습하고 전체 데이터에 적용함. 연구자가 직접 긍정/부정/중립으로 라벨링한 학습 데이터를 구축하여 적용하므로, 분석 주제의 제한 없이 다양한 분야에서 감성분석이 가능함

< 문서 감성 분류 분석 시각화 예시 >



구분	개수	백분율
긍정	3,018	65.9%
부정	101	2.2%
중립	1,460	31.9%
총합계	4,579	100.0%

※ 2005.1.~2024.9. 기간의 "하숙" 데이터 중 '공주' 키워드를 포함한 5,091건 중 감성 분류 분석 진행 후, 검수 과정에서 스팸 데이터 및 분석과 관련 없는 데이터를 제외함. 최종 감성 분류 분석 건수는 4,579건으로 산출됨

긍정 원문 (일부)	부정 원문 (일부)
제민천따라걷기 좋은 공주하숙마을 공주하숙마을겨울의 공주는 생각보다 훨씬 아름다웠으며 곳곳에 하얀 눈으로 뒤덮인 풍경을 감상할 수 있었다 눈 내리는 공주가 자주 생각나 조만간 대설주의보가 내리면 버스 표를 끊고 바로 내려갈 예정 그만큼 공주에 애정이 있는 나는 오늘 원도심을 따라 천천히 산책하는 코스를 선택했다 공주시 원도심에 위치한 복합 문화공간이자 숙박시설인 공주하숙마을을	2주 로컬여행 혼자서 공주 원도심 제대로 여행하는법이 동네가 가진 독특한 스토리를 활용해서 2014년 즈음에 도시재생사업의 일환으로 이 하숙마을을 만들고 게스트하우스이자 복합문화공간으로 운영한다고 합니다 사실 저도 하루 정도는 여기서 묵어볼까 생각했는데예약자가 아무도 없길래 무서워서 못했어요이 날 가보니까투숙객분들이 계시더라구요하숙마을은 사실 딱히 볼게없고그 뒤편으로
1박 2일 공주 여행 후기 공산성제민천맛집 추천 하천을 따라 걷다 보면 공주 하숙 마을 이 보이는데 고즈넉한 한옥의 아름다움을 느끼기 좋은 곳입니다공주성과는 다른 분위기의 제민천도 공주 여행에서 빼놓으면 안 될 여행지입니다 공주성과 제민천에대한자세한 후기는 아래 글을 참고해 주세요 공주 가볼만한곳공산성부터 제민천까지 공주 여행 코스 가을맞이 국내 여행으로 1박 2일 공주 여행을	한적하게 여행한 공주한옥펜션한적한가 정도 하숙 마을 별다른 특색 없으나 어차피 다른 곳들을 가기 위해 지나가게 됨꽃꽃문학과 시를 좋아하지 않는다면 필수 코스 아님 공주 중동 성당 시간 없으면 안 가도 됨 역사박물관 앞이라 그냥 가보자 메타세콰이어그냥 흔한 길 미르섬여길 왜 마무리 위낙 비를 좋아하지만 이번엔 불명을할 수 없어서