텍스트 마이닝을 이용한 KEI 연구동향 분석

Bigdata Research Team Seminar : Progress Report

2017. 08. 31

빅데이터연구팀 김도연

목차

연구 개요

선행 연구

연구 내용

연구 추진방법

기대효과

차 례

1. 서론
가. 연구배경 및 목적
나. 연구내용 및 범위
다. 선행연구 동향
라. 본문의 구성

2. 텍스트 마이닝 기반 연구동향 분석 방법론
가. 텍스트 마이닝 분석 기법
1) Association Rule Mining
2) Keyword Network Analysis
3) Latent Dirichlet Allocation(LDA)
4) Word2Vec
나. 연구동향분석을 위한 텍스트 마이닝 적용 가능성
다. 연구 분석 절차

3. 분석 데이터 개요
가. KEI 연구보고서
나. NAVER 환경뉴스

4. LDA기반 토픽 클러스터링 분석
가. 매체별 LDA 분석
1) KEI 연구보고서 LDA 분석
2) NAVER 환경뉴스 LDA 분석
나. 매체별 LDA 비교분석 결과

5. 연관어 및 네트워크 분석
가. 매체별 환경분야 동향 분석
1) Period 1(2004-2007)
2) Period 2(2008-2012)
3) Period 3(2013-2016)
나. 매체별 환경분야 동향 비교분석 결과

6. Word2Vec 분석
가. 매체별 기후변화 세부현상 키워드 분석
1) 온난화
2) 홍수
3) 가뭄
나. 매체별 Word2Vec 비교분석 결과

7. 결론 및 제언

초 록

본 연구는 자연언어분석을 이용하여 한국환경정책·평가연구원(이하 KEI)의 연구동향을 파악하고, KEI 연구동향이 환경연구에 대한 사회적 연구 수요와 조응하는 지 여부를 분석하였다. 연구주제 선정 범위는 연구를 수행하는 개별 연구자의 성향 및 경험에 따라 제한되기 때문에, KEI의 연구 동향이 국민적 관심과 유리될 수 있다는 우려는 지속적으로 존재해 왔다. 이러한 우려를 확인하기 위해서 본 연구는 연구동향을 나타내는 KEI 연구보고서와 연구수요를 대변하는 환경관련 언론 기사의 장기적인 추이를 비교 분석한다. 이러한 연구는 대량의 텍스트 자료 특성 추출이 필수불가결하기 때문에 본 연구에서는 다양한 텍스트 마이닝 기법을 적용하여 이를 수행하였다.

구체적으로 본 연구에서는 KEI에서 발행되는 연구보고서와 NAVER에서 제공하는 환경기사 데이터를 수집한 후 LDA기반 토픽 클러스터링 분석, 연관어 및 네트워크 분석, Word2Vec 분석 등 다양한 텍스트 마이닝 기법을 이용하여 장기간의 KEI 연구동향과 사회적 연구수요 동향을 시기별로 각각 추출하여 비교 분석하였다. 연구에서 제안한 텍스트 마이닝 기법을 이용한 연구수요를 파악하는 방법은 기존의 개별 연구자의 직관에 의존하는 방식을 보완하는 방법으로 활용도가 높을 것으로 기대된다.

연구 배경 및 목적

■ **KEI 연구동향이 국민적 관심에 반응하고 있는 지 여부에 대한 회의 존재**

- 개별 연구자의 시간적 제약 및 개인적 연구 성향에 의해서 연구수요 정보 파악 범위가 제한
- 파악된 정보에 부여되는 우선순위가 개별 연구자의 선호에 영향을 받으므로 최신 정보 및 시의성 있는 연구수요 반영에 제약이 존재

■ **최근 트렌드 분석에 활발하게 사용되는 텍스트 마이닝을 통해 KEI 연구동향과 민간의 환경연구 수요 간의 관계 파악 가능**

- 텍스트 마이닝은 실시간으로 생산되는 다량의 비정형데이터 속에서 의미 있는 패턴을 발견하여 트렌드를 파악하는데 주로 활용
- 대용량 텍스트 자료 분석이 가능하므로 KEI 연구동향과 민간의 연구수요 동향을 시기별로 트렌드를 각각 추출하여 비교 분석 가능

▶ **본 연구는 텍스트 마이닝을 이용한 24년(1993~2016) KEI 연구동향 분석을 시도하여 시간적 추이 및 민간 연구수요와의 조응여부를 탐구**

- KEI 연구문헌 및 온라인 뉴스기사 분석을 병행하여 환경관련 연구공급 동향 및 연구수요 동향을 파악
- 텍스트 마이닝 기법을 이용하여 연구수요를 파악하는 방법의 예를 제공하여 기존의 개별 연구자의 직관에 의존하는 방식을 보완하는 방법을 제공

연구 개요
선행 연구
연구 내용
연구 추진방법
기대효과

선행 연구 현황

구분	연구 목적	연구 방법	주요 연구내용
1	과제명: 텍스트 마이닝 기법을 활용한 한국의 경제연구 동향 분석 연구자(년도): 송혜지 외(2013) 연구목적: 텍스트 마이닝 기법 활용 외국 학술지 한국 경제 분야 트렌드 분석	키워드 분석 네트워크 분석 토픽모델링 분석	외국 학술지의 한국경제 연구에 대한 연구 동향 및 지적 구조 파악
2	과제명: 소셜 빅데이터를 활용한 국민 통일인식 동향 분석 연구자(년도): 송태민 (2015) 연구목적: 2014년 '통일대박론' 대두 이후 통일인식 변화를 소셜 빅데이터 이용 분석	키워드 분석 연관성 분석	소셜 미디어 통일관련 연관어 분석 통일관련 연관어와 통일인식간의 관계 분석
3	과제명: 항공 산업 미래유망분야 선정을 위한 텍스트 마이닝 기반의 트렌드 분석 연구자(년도): 김현정 외(2015) 연구목적: 텍스트마이닝 트렌드 분석활용 항공 산업 미래유망분야 발굴	토픽모델링 분석	텍스트 마이닝 기법을 적용한 항공 산업 관련 논문 트렌드 분석 토픽 모델링 분석 활용 항공 산업 미래유망부분 추출
4	과제명: 빅데이터를 활용한 환경분야 정책수요 분석 연구자(년도): 이미숙 외(2014) 연구목적: 매체별(뉴스, 블로그, 트위터) 환경정책 수요 분석	감성 분석 연관성 분석 네트워크분석	세부 환경분야별 소셜빅데이터 분석 전체문서 및 환경문서의 행복도 비교 분석

주요 연구 내용

KEI 연구동향 파악

- 텍스트 마이닝 기법을 활용한 연구동향 분석
  - KEI가 설립된 1993년부터 2016년까지의 KEI DB에서 제공하는 연구보고서(제목, 목차, 요약, 날짜)를 분석에 활용
  - 24년 간(1993-2016) 연구보고서 1,697건

매체별 환경분야 이슈 비교 분석

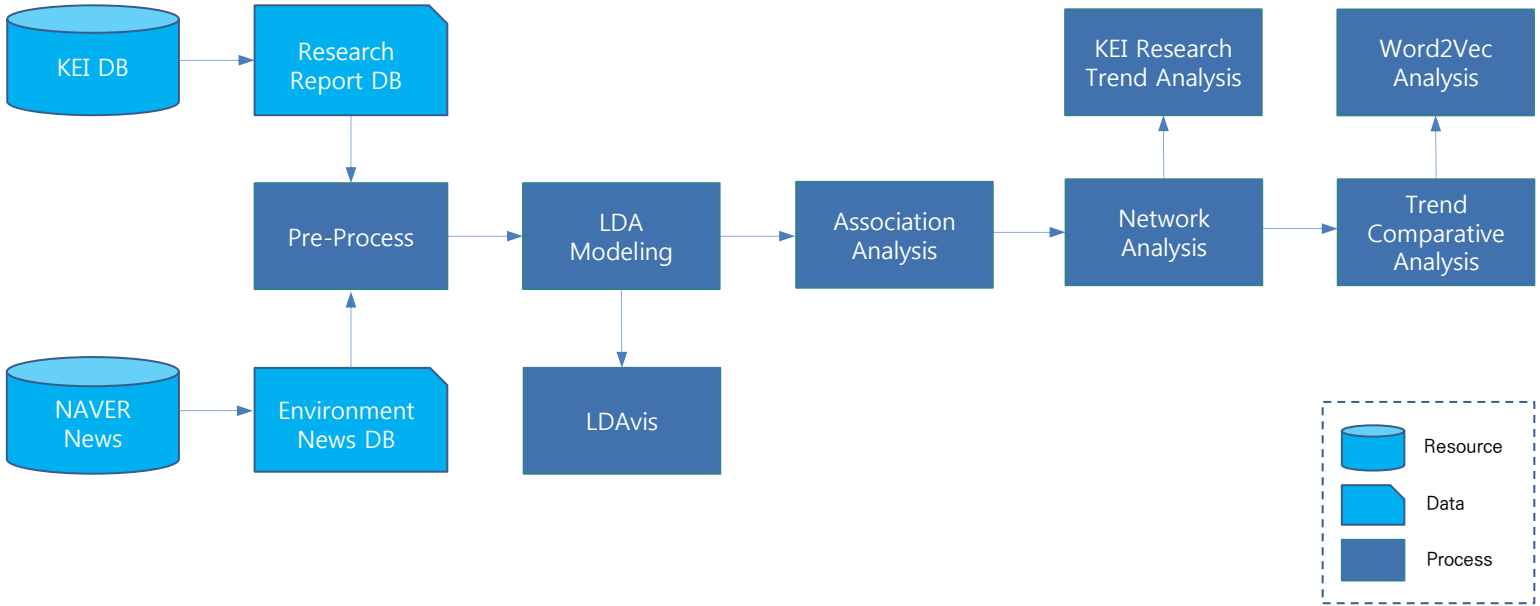
- 연구공급 동향과 연구수요 동향 비교 분석
  - 연구공급 동향 파악 : 2004년부터 2016년(13개년)까지의 KEI DB에서 제공하는 연구보고서(제목, 목차, 요약, 날짜) 분석
  - 연구수요 동향 파악 : 2004년부터 2016년(13개년)까지의 언론매체에서 제공하는 뉴스기사(제목, 날짜) 분석
- 매체별 추출한 키워드를 시계열로 파악하고 두 시계열을 비교 분석

매체별 '기후변화' 세부현상 키워드 분석

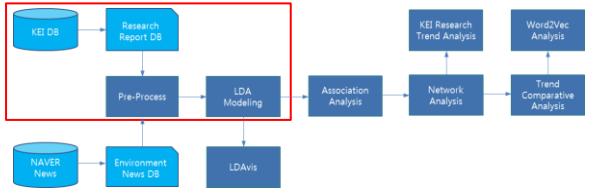
- 기후변화 세부 현상(온난화, 홍수, 가뭄) 별 연관어 분석

Plan '17		Process	Code	Description	Input	Output	Note
상반기	3월 상순	Pre-processing(1)	topic_clustering.R	- 형태소분석기 실행(KoNLP , tctStart) - 불용어처리 등 전처리 과정 (특정 단어, 특수문자 삭제) - Word Lengths 한글자 삭제 - Sparse Terms 삭제 - Low TF-IDF 삭제	kei.xlsx	out_kei.csv DocumentTermMatrix 부록1_제거 대상 키워드 목록.hwp	- 자문의견(이명진 박사님) : 한글 처리 문제 -> 다양한 한글 전처리 방법을 통해 해결 가능함.
	3월 하순	LDA Modeling	topic_clustering.R	- LDA기반 토픽 모델링 - 토픽별 핵심 단어 출력 - 문서별 토픽번호 및 확률값 출력 - 단어별 토픽번호 및 확률값 출력	Document TermMatrix	term_topic.csv doc_Prob_df.csv doc_prob_df_max.csv id_topic.csv lda_tm.csv	- 입력값 : SEED = 2017, K = 5
	4월	LDAvis	topic_clustering.R	- 토픽모델링 - 2차원 시각화 및 주요 키워드 확률분포 목록 시각화	lda_tm.csv	HTML 등 웹파일	- apache-tomcat-8.5.12 사용 - 산출물 서버업로드 필요
	5월 상순	LDA Result Analysis		- 토픽별 키워드 분석 - 토픽별 연구보고서 동향 분석	id_topic.csv	id_topic_Analysis.xlsx	-1993~2016년 연구보고서 동향 분석
	5월 하순	Association Analysis(1)	Association_Analysis.R	- 지지도, 신뢰도가 0.01 이상 값 출력 - 3가지측도(지지도, 신뢰도, 향상도) 분석	1993_2002.txt 2003_2007.txt 2008_2012.txt 2013_2016.txt	Association.xlsx	- 연구보고서 제목 데이터 활용 - 초록으로 분석시 메트릭스가 너무 커짐 - 4개 시기별 동향 분석
		Network Analysis(1)	Association_Analysis.R	- 원의 크기 : 언급량이 많을수록 크기가 큼 - 원의 색깔 : 매개중심성이 높을수록 색깔이 진함		93-02.png 03-07.png 08-12.png 13-16.png	
	6월 상순	Data Collection	naver_news1.java naver_news2.java naver_news3.java	- Java jsoup을 사용하여 Web crawling - 조건: 네이버 뉴스 -> 사회-> 환경 - 기간: 2004.1.1~2016.12.12 (13개년) - 영역: 제목, 날짜, 언론사 - 양: 193,636개		Naver_news.csv Naver_news_Analysis.xlsx 부록2_네이버 환경뉴스 언론사별 산출 양	- 2004년 이전 네이버 뉴스 기사 부실
하반기	6월 상순	Pre-processing(2)	topic_clustering.R	- 형태소분석기 실행(KoNLP , tctStart) - 불용어처리 등 전처리 과정 (특정 단어, 특수문자 삭제) - Word Lengths 한글자 삭제 - Sparse Terms 삭제 - Low TF-IDF 삭제	naver.xlsx	out_naver.csv DocumentTermMatrix 부록1_제거 대상 키워드 목록.hwp	- 자문의견(이명진 박사님) : 한글 처리 문제 -> 다양한 한글 전처리 방법을 통해 해결 가능함.
	6월 상순	LDA Modeling(2)	topic_clustering.R	- LDA기반 토픽 모델링 - 토픽별 핵심 단어 출력 - 문서별 토픽번호 및 확률값 출력 - 단어별 토픽번호 및 확률값 출력	Document TermMatrix	news_term_topic.csv news_doc_Prob_df.csv news_doc_prob_df_max.csv news_id_topic.csv news_lda_tm.csv	- 입력값 : SEED = 2000000, K = 10
	6월 상순	LDAvis(2)	topic_clustering.R	- 토픽모델링 - 2차원 시각화 및 주요 키워드 확률분포 목록 시각화	lda_tm.csv	HTML 등 웹파일	- apache-tomcat-8.5.12 사용 - 산출물 서버업로드 필요
	6월 하순	LDA Result Analysis(2)		- 토픽별 키워드 분석 - 토픽별 네이버뉴스 동향 분석	news_id_topic.csv	news_id_topic_Analysis.xlsx	- 2004~2016년 네이버 뉴스 기사 연도별 동향 분석 - 매체별 동향 비교 분석
	중간자문회의(2017.6.29)						
	7월	Association Analysis(2)	Association_Analysis.R	- 지지도 0.001, 신뢰도가 0.005 이상 값 출력 - 3가지측도(지지도, 신뢰도, 향상도) 분석	2004_2007.txt 2008_2012.txt 2013_2016.txt	news_Association.xlsx	- 네이버뉴스 제목 데이터 활용 - 본문으로 분석시 메트릭스가 너무 커짐 - 3개 시기별 동향 분석
		Network Analysis(2)	Association_Analysis.R	- 원의 크기 : 언급량이 많을수록 크기가 큼 - 원의 색깔 : 매개중심성이 높을수록 색깔이 진함		N_04-07.png N_08-12.png N_13-16.png	
	8월 상순	Trend Comparative Analysis		- 매체별 동향을 3개 시기로 나눠서 비교 - 3개시기(2004~2007, 2008~2012, 2013~2016)	Association.xlsx Naver_Association.xlsx		- 매체별 키워드 연관성 및 네트워크 분석 결과 활용 - 3개 시기 분류 기준: 대통령 재임기간 - 분석결과 '기후변화'가 중요 키워드로 나타남
	8월 하순	Word2Vec Analysis	Word2Vec.R	- Skip-Gram Model 사용함 - 거리측정법: 코사인거리 - 기후변화 세부 현상들 3가지(온난화, 홍수, 가뭄) 키워드 연관어 분석	out_kei.csv out_naver.csv	kei_w2v.txt kei_w2v_2.bin naver_w2v.txt naver_w2v_2.bin	- Window size: 10 - Worker threads: 3 - Word vector dimensionality: 100
	9월	결론 및 시사점 도출					
	10월	향후 계획 수립					

연구 프로세스



Text Mining Process List



Plan 2017	Process	Code	Description	Input	Output	Note
3월 상순	**Pre-processing(1)**	topic_clustering.R	- 형태소분석기 실행(KoNLP 등) - Low TF-IDF 값 제거 - 불용어처리 등 전처리 과정 (특정 단어 삭제, 특수문자 제거, 소문자로 변경 등) - Word Lengths는 2글자 이상 - 동의어 처리	kei.xlsx	out_kei.csv DocumentTermMatrix 부록1_제거 대상 키워드 목록.hwp	-자문의견(이명진 박사님) : 한글 처리 문제 -> 다양한 한글 전처리 방법을 통해 해결 가능함.
3월 하순	**LDA Modeling(1)**	topic_clustering.R	- LDA기반 토픽 모델링 - 토픽별 핵심 단어 출력 - 문서별 토픽번호 및 확률값 출력 - 단어별 토픽번호 및 확률값 출력	Document TermMatrix	term_topic.csv doc_Prob_df.csv doc_prob_df_max.csv id_topic.csv lda_tm.csv	- 입력값 : SEED = 2017, K = 5
4월	**LDAvis(1)**	topic_clustering.R	- 토픽모델링 - 2차원 시각화 및 주요 키워드 확률분포 목록 시각화	lda_tm.csv	HTML 등 웹파일	- apache-tomcat-8.5.12 사용 - 산출물 서버업로드 필요

Pre-processing(1)

1. Pre-processing with R

```
#형태소 분석기 실행
system("tctStart")

#Corpus 생성
corp<-VCorpus(VectorSource(parsedDataRe$pContent))
#특수문자 제거
corp <- tm_map(corp, removePunctuation)
#소문자로 변경
corp <- tm_map(corp, tolower)
#특정 단어 삭제
corp <- tm_map(corp, removeWords,
               c("전략", "연구", "평가", "마련", "조사", "관리", "보다", "분석", "구축"))
#TXET 문서 형식으로 변환
corp <- tm_map(corp, PlainTextDocument)
#Document Term Matrix 생성 (단어 Length는 2로 세팅)
dtm<-DocumentTermMatrix(corp, control=list(removeNumbers=FALSE, wordLengths=c(2,Inf)))
#한글자 단어 제외하기
#colnames(dtm) = trimws(colnames(dtm))
#dtm = dtm[,nchar(colnames(dtm)) > 1]

#Sparse Terms 삭제
dtm <- removeSparseTerms(dtm, as.numeric(0.997))
#Remove low tf-idf col and row
term_tfidf <- tapply(dtm$v/row_sums(dtm)[dtm$i], dtm$j, mean) * log2(nDocs(dtm)/col_sums(dtm > 0))
new_dtm <- dtm[,term_tfidf >= 0]
new_dtm <- new_dtm[row_sums(new_dtm)>0,]
```

연구자의 판단으로 …

2. Pre-processing data: id_topic.csv 생성

rown	id	doc_topic	pContent	maxProb	Title	author	class	year	month	day
1	a1	5	환경 개선 마스터플랜 수립 지원 국내 사후 설	0.840	환경분야 공적개발원조(O	조공장	수시연구	2016	06	30
2	a2	4	추진 국내 탄소 감축 정책 해외 탄소 감축 정책	0.450	제주 탄소제로섬 추진전략	이병국	수시연구	2016	06	24
3	a3	4	국내외 기술 사회 경제 시나리오 사회 경제 시	0.340	저탄소 기후변화 적응 사회	채여라	수탁연구	2016	05	31
4	a4	3	화학 사고 피해액 추정 제안 화학 사고 인적 상	0.805	화학사고의 경제적 손실 추	서양원,곽	수시연구	2016	04	30
5	a5	3	나노 폐기물 나노 물질 나노 폐기물 세계 나노	0.586	나노폐기물의 안전처리를	조지혜	수시연구	2016	04	30
6	a6	2	2015년 충남 서북 부 지역 가뭄 대응 2015년	0.760	가뭄 단계에 따른 적응형	김호정	수시연구	2016	03	31
7	a7	2	국내외 기술 국내 기술 국내 산지 정책 동향 국	0.352	국내 능선축 GIS기반 통합	이수재	수탁연구	2016	03	31
8	a8	3	국내외 기술 최종 최종 단계 추진 토의 건강 영	0.542	기후변화에 따른 건강영향	신용승	수탁연구	2016	02	28
9	a9	2	제네바 텍스트 중재 제네바 텍스트 제네바 텍스	0.502	Post-2020 신기후체계 협상	이승준	수탁연구	2015	12	31
10	a10	4	사물 인터넷 혁명 핵심 기술 부상 물 환경 사물	0.404	사물인터넷(IoT)을 활용한	한혜진	연구	2016	10	31
11	a11	4	나다 차별 중국 전략 아시아 인프라 투자 은행	0.998	중국의 '일대일로(一帶一路	추장민	연구	2016	10	31
12	a12	2	도시 기후 회복력 도시 기후 회복력 도시 기후	0.810	도시의 기후 회복력 확보를	김동현	연구	2016	10	31
13	a13	2	국내 지역 사회 환경 보건 문제 진단 국내 지역	0.458	지역기반 환경보건정책 지	신용승,배	연구	2016	10	31
14	a14	2	파리 협정 핵심 파리 협정 파리 협정 적응 손실	0.910	신기후체제의 기후변화 적	이승준	수시연구	2016	09	30
15	a15	4	차별친환경 차 보급 정책 동향 국내 정책 동향	0.898	대기환경비용을 고려한 친	한진석	수시연구	2016	09	30
16	a16	1	경유 차 실 도로 대기 오염 물질 초과 배출 원	0.821	실도로에서 경유차의 대기	강광규	수시연구	2016	09	22
17	a17	5	접근 국내 정화 곤란 부지 시사점 건설 기계 사	0.703	토양정화 곤란 부지의 최적	박용하	수시연구	2016	08	31
18	a18	2	과업 과업 실행 생물 다양 정책 여건 중간 점검	0.896	제3차 국가생물다양성전략	이현우	수탁연구	2016	08	30
19	a19	4	지속 가능 발전 87년 환경 관 세계 위원회 발표	0.733	국가 지속가능성 평가 등이	김종호	수탁연구	2016	07	30
20	a20	2	국외 지지다 공원 동향 유네스코 아시아 태평	0.772	유네스코 세계지질공원 운	이수재	수시연구	2016	11	22
21	a21	2	시스템 네트워크 언어미래 환경 정책 에너지	0.526	시스템과 네트워크 이론을	이승준	기초연구	2016	11	06
22	a22	5	국가 지역 미래 성장 동력 미래 성장 동력 미래	0.479	국가 및 지역 미래성장동력	방상원	연구	2016	10	31
23	a23	5	국가 지역 미래 성장 동력 미래 성장 동력 미래	0.479	지중환경 를 위한 제도 개	황상일	연구	2016	10	31
24	a24	3	정책 정부 패러다임 주민수 용성 정책 주민수	0.699	정부3.0 기반 지역기피시설	김태현	연구	2016	10	31
25	a25	3	차별국내외 재해 폐기물 제도 비교 국내 수해	0.596	공간정보를 활용한 재해폐	조지혜,김	연구	2016	10	31
26	a26	3	국내 폐기물 활용 활용 산업 국내 폐기물 발생	0.540	자원순환사회 전환 촉진을	이소라,신	연구	2016	10	31
27	a27	3	전기 전자 제품 활용 정책 활용 국내 일본 국내	0.844	폐자원흐름분석을 통한 전	이희선	연구	2016	10	31
28	a28	4	물 환경 인프라 사회 투자 수익 물 환경 인프라	0.605	사회적 투자수익률(SROI)을	류재나,강	연구	2016	10	31
29	a29	2	자연 자본 여건 전망 자연 자본 특성 국내 자연	0.393	생태계서비스 기반의 자연	이현우	연구	2016	10	31
30	a30	2	크리티컬 존 국내외 정책 동향 국외 크리티컬	0.527	근지표환경 임계영역(Criti	현윤정	기초연구	2016	12	06
31	a31	5	국내 외 환경 재난 사후 대응 정책 국내 환경	0.359	드론을 이용한 환경재난 사	손승우	기초연구	2016	12	06
32	a32	4	건물 지속 가능 고찰 건물 환경 주다 영향 측정	0.648	건물부문의 환경부하 평가	송지윤	기초연구	2016	12	06
33	a33	3	고찰 기후 변화 노동자 대 영향 노동자 위험 대	0.712	미래 고온환경 변화와 직장	김동현	기초연구	2016	12	06

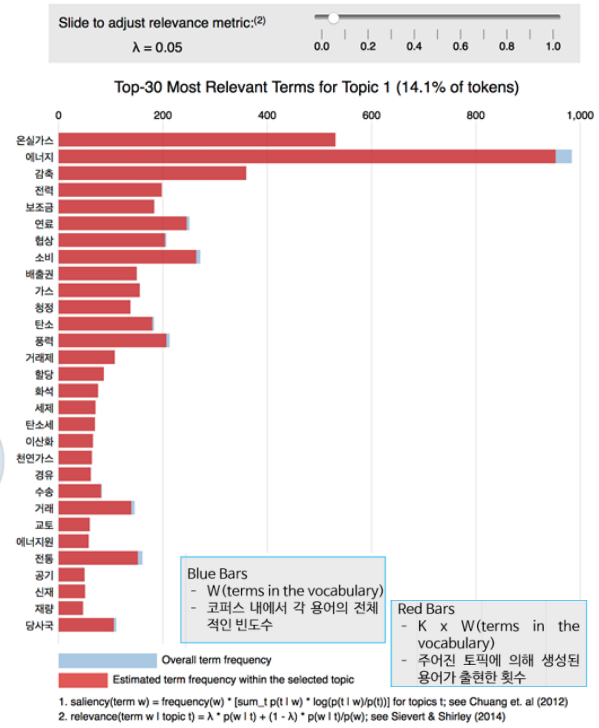
Sheet1 id_topic Sheet2

Pre-processing(1)

〈부록〉 제거 대상 키워드 목록

채널	제거 대상 키워드
KEI 연구 보고서	10년, 1990년, 1장, 1절, 2000년, 2001년, 2002년, 2003년, 2004년, 2005년, 2006년, 2007년, 2008년, 2009년, 2010년, 2011년, 2012년, 2013년, 2014년, 2015년, 2020년, 2장, 3장, 3절, 4장, 4절, 5개년, 5장, 가능, 가다, 같다, 개념, 개발, 개선, 개요, 결과, 결론, 결정, 경우, 계수, 계획, 고려, 과제, 관련, 관리, 관점, 관하다, 구조, 구축, 그리다, 기반, 기본, 기존, 기준, 기초, 기타, 내용, 다루다, 다양, 대책, 대하다, 도출, 되다, 따른, 마련, 말다, 모형, 목록, 목적, 목차, 문헌, 미치다, 발전, 방법, 방안, 방향, 배경, 범위, 보고서, 보급, 보다, 보이다, 본론, 부록, 부문, 분석, 비교, 사업, 사용, 사항, 산정, 서다, 서론, 설정, 수립, 수행, 시기, 시스템, 업무, 업종, 여건, 연구, 영향, 요소, 요약, 우리, 운영, 위하다, 유형, 의하다, 이러하다, 이루어지다, 이용, 인하다, 작성, 적용, 적절하다, 전략, 절차, 정보, 정의, 제공, 제기, 제도, 제시, 제안, 조사, 종합, 중심, 지속가능, 지점, 차례, 참고, 처리, 체계, 체제, 초록, 추진, 측면, 통하다, 통합, 특성, 특징, 평가, 평가모형, 필요, 하다, 허용, 현황, 협약, 활용, 회의, 효과 **(총 154개)**

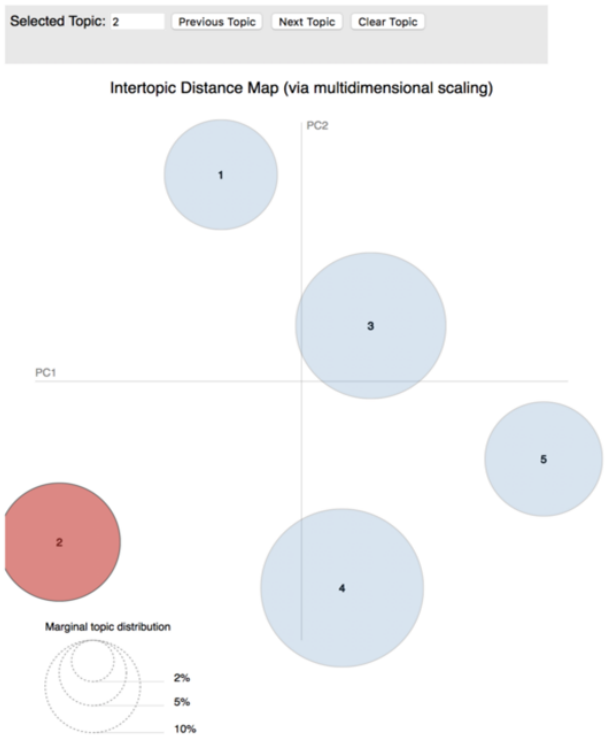
LDAvis (Topic 1)

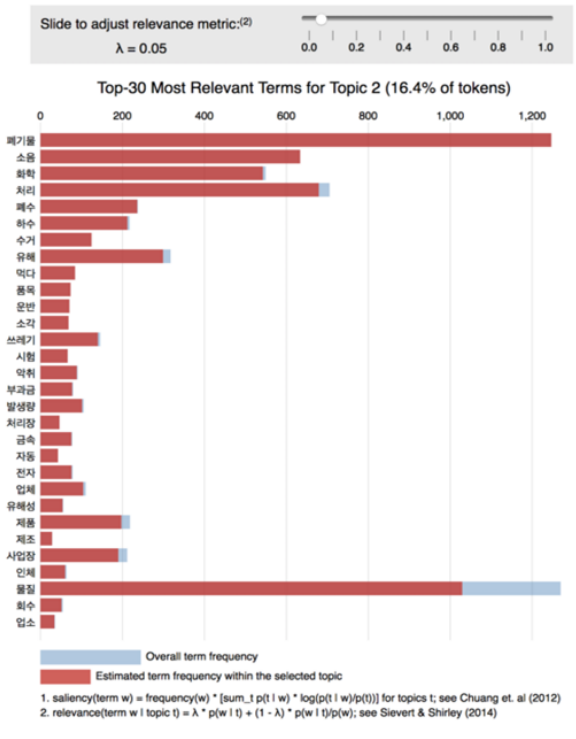


Term
온실가스
에너지
전력
연료
가스
청정
탄소
풍력
세제
탄소세
이산화탄소
천연가스
경유
공기
신재생에너지

“에너지 자원”

LDAvis (Topic 2)

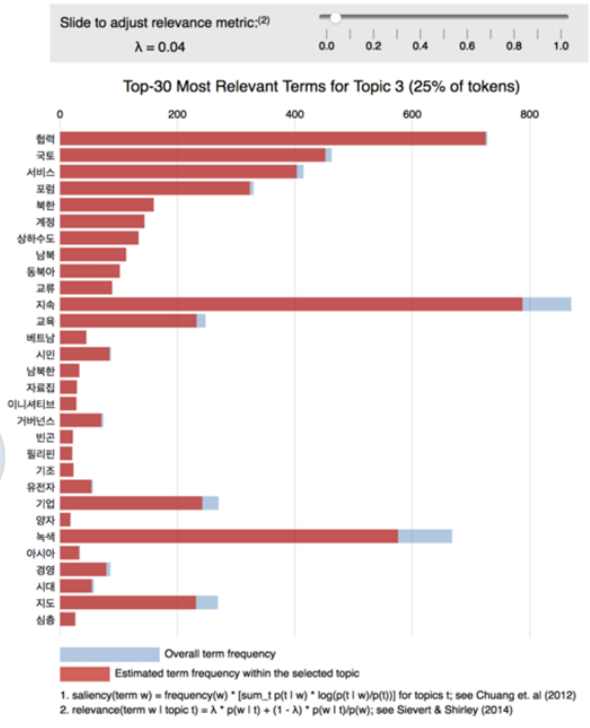




Term
폐기물
소음
화학
처리
폐수
하수
수거
유해
막다
소각
쓰레기
악취
부담금
처리장
유해성

“폐기물”

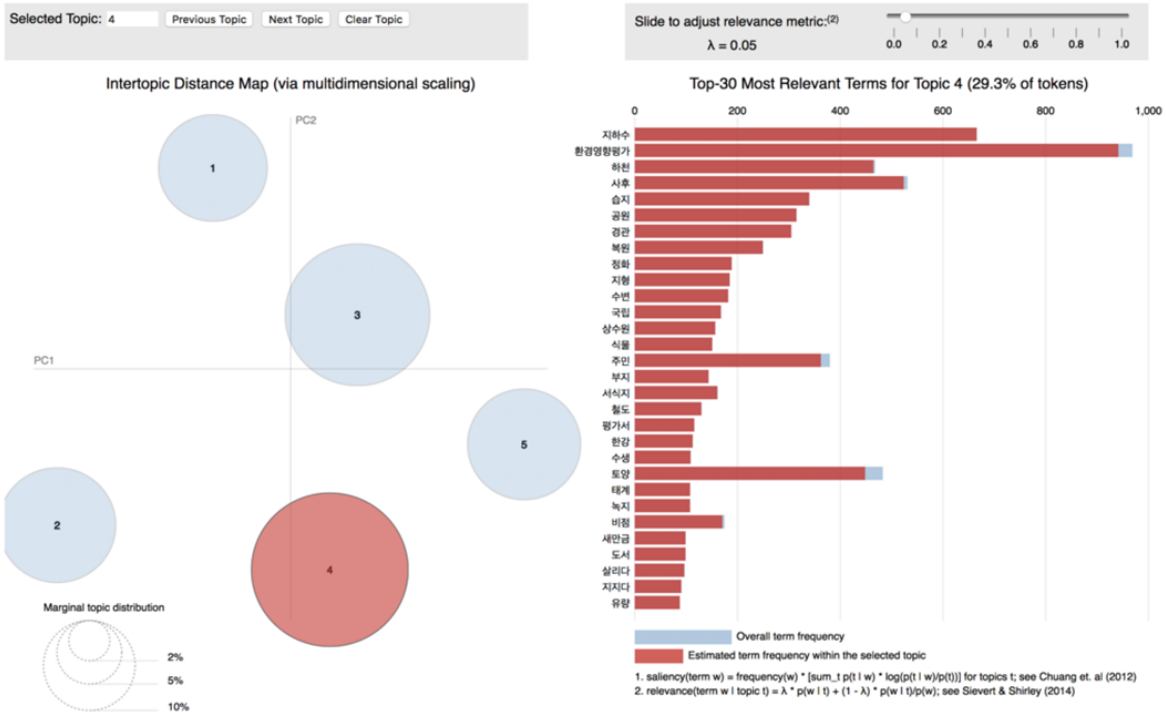
LDAvis (Topic 3)



Term
협력
포럼
북한
상하수도
남북
동북아
교류
지속
베트남
시민
남북한
이니셔티브
거버넌스
필리핀
아시아

“대외 협력”

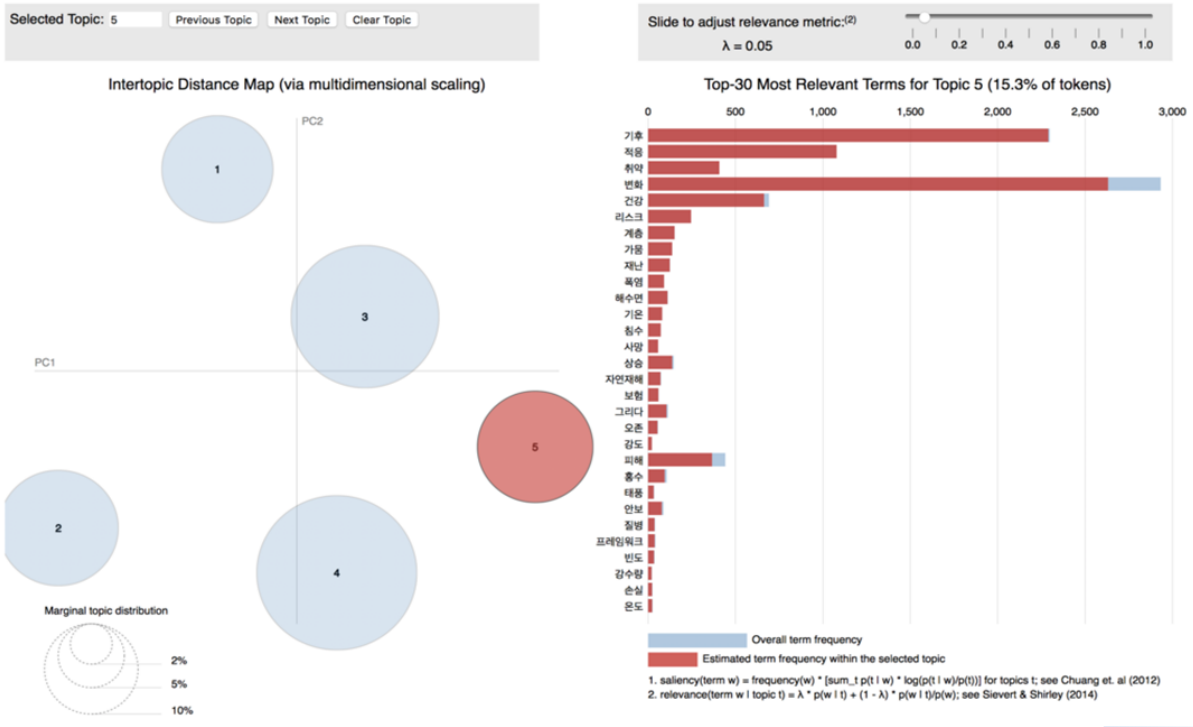
LDAvis (Topic 4)



Term
지하수
환경영향평가
하천
습지
정화
지형
수변
상수원
부지
서식지
한강
수생
토양
녹지
새만금

“물환경”
+
“환경영향평가”

LDAvis (Topic 5)



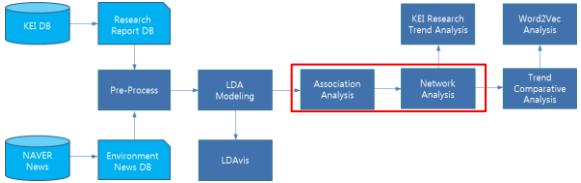
Term
기후
변화
가뭄
재난
폭염
해수면
기온
침수
사망
자연재해
오존
홍수
태풍
강수량
온도

“기후변화”

LDAvis (Topic 전체)

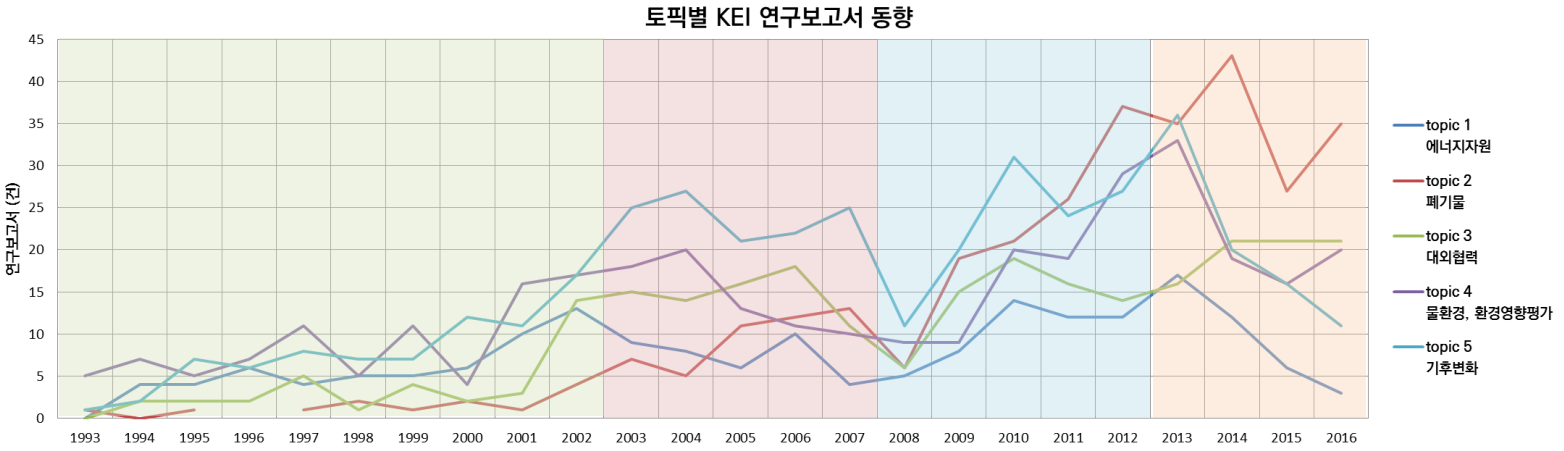
No.	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Title	에너지 자원	폐기물	대외협력	물 환경, 환경영향평가	기후변화
1	온실가스	폐기물	협력	지하수	기후
2	에너지	소음	포럼	환경영향평가	변화
3	전력	화학	북한	하천	가뭄
4	연료	처리	상하수도	습지	재난
5	가스	폐수	남북	정화	폭염
6	청정	하수	동북아	지형	해수면
7	탄소	수거	교류	수변	기온
8	풍력	유해	지속	상수원	침수
9	세제	막다	베트남	부지	사망
10	탄소세	소각	시민	서식지	자연재해
11	이산화탄소	쓰레기	남북한	한강	오존
12	천연가스	악취	이니셔티브	수생	홍수
13	경유	부담금	거버넌스	토양	태풍
14	공기	처리장	필리핀	녹지	강수량
15	신재생에너지	유해성	아시아	새만금	온도
...					

Text Mining Process List



Plan 2017	Process	Code	Description	Input	Output	Note
5월 상순	**LDA Result Analysis(1)**		- 토픽별 키워드 분석 - 토픽별 연구보고서 동향 분석	id_topic.csv	id_topic_Analysis .xlsx	-1993~2016년 연구보고서 연도별 동향 분석
5월 하순	**Association Analysis(1)**	Association_ Analysis.R	- 지지도, 신뢰도가 0.01 이상 값 출력 - 3가지측도(지지도, 신뢰도, 향상도) 분석	1993_2002.txt 2003_2007.txt 2008_2012.txt 2013_2016.txt	Association.xlsx	- 연구보고서 제목 데이터 활용 - 초록으로 분석시 메트릭스가 너무 커짐 - 4개 시기별 동향 분석
	Network Analysis(1)	Association_ Analysis.R	- 원의 크기 : 언급량이 많을수록 크기가 큼 - 원의 색깔 : 매개중심성이 높을수록 색깔이 진함		93-02.png 03-07.png 08-12.png 13-16.png	

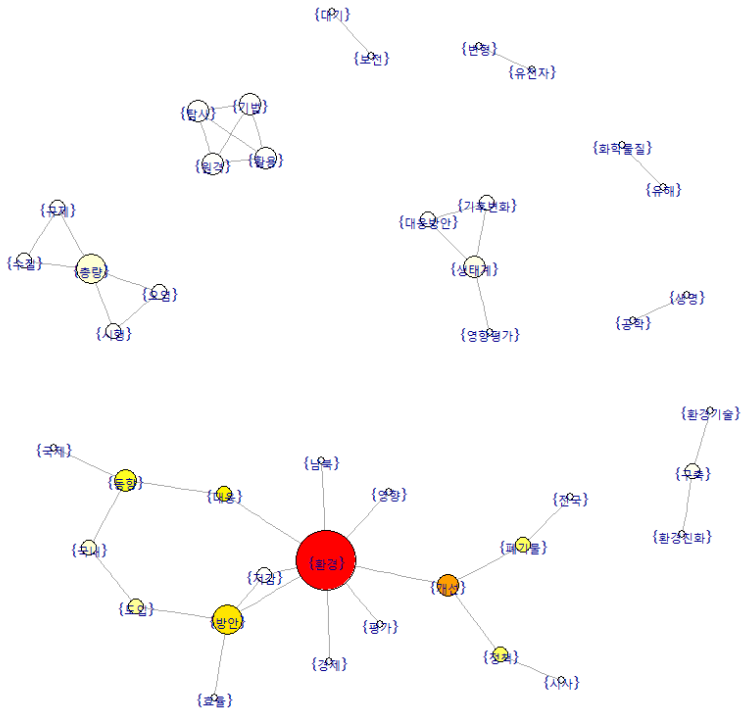
토픽별 KEI 연구보고서 동향 분석



- 1993~ 2002년도: 전반적으로 토픽별 연구추세가 비슷함.
- 2003~ 2007년도: 기후변화 관련 연구가 활발하게 진행
  물 환경/환경영향평가, 에너지자원 관련 연구는 감소하는 추세를 보임.
- 2008~ 2012년도: 폐기물, 물 환경/환경영향평가 연구가 급증함.
- 2013~ 2016년도: 폐기물 관련 연구가 활발하게 진행
  2015년을 기점으로 연구의 양이 적어짐.

Doc Topic	Title	1993~ 2002	2003~ 2007	2008~ 2012	2013~ 2016	총합
1	에너지자원	57	37	51	38	**183**
2	폐기물	13	48	109	140	**310**
3	대외협력	35	74	70	79	**258**
4	물 환경, 환경영향평가	88	72	86	88	**334**
5	기후변화	78	120	113	83	**394**
NA(영문, 한문)		5	27	11	0	**43**
총합		276	378	440	428	**1,522**

1. 키워드 연관성 및 네트워크 분석(1993-2002년)



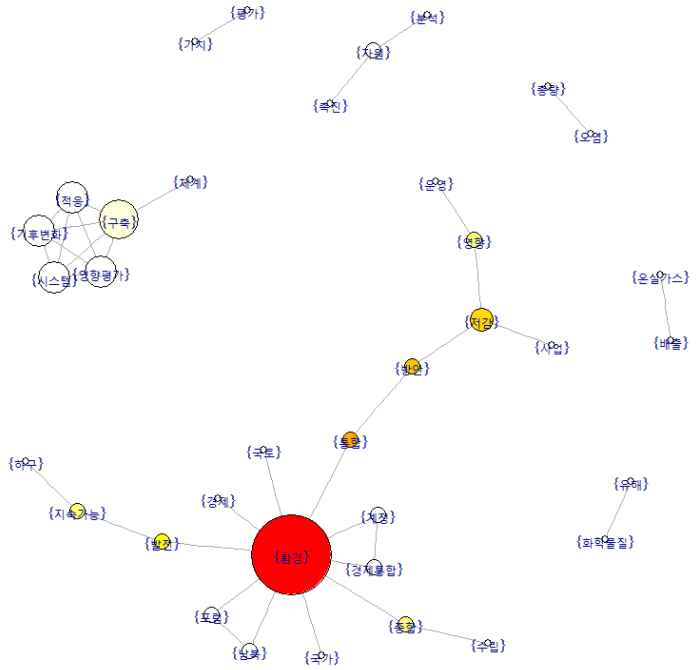
* 키워드 네트워크 분석 결과
1. 원의 크기 : 언급량이 높을수록 크다.
2. 원의 색깔 : 매개중심성이 높을수록 진하다.(하얀색〈노란색〈주황색〈빨간색)

no	lhs		rhs	support	confidence	lift
1	유전자	=>	변형	0.0123	1.0000	81.3333
2	변형	=>	유전자	0.0123	1.0000	81.3333
3	기후변화	=>	생태계	0.0123	0.7500	45.7500
4	생태계	=>	기후변화	0.0123	0.7500	45.7500
5	기후변화	=>	대응방안	0.0123	0.7500	36.6000
6	대응방안	=>	기후변화	0.0123	0.6000	36.6000
7	영향평가	=>	생태계	0.0123	1.0000	61.0000
8	생태계	=>	영향평가	0.0123	0.7500	61.0000
9	남북	=>	환경	0.0123	0.7500	4.8158
10	환경	=>	남북	0.0123	0.0789	4.8158
11	생태계	=>	대응방안	0.0123	0.7500	36.6000
12	대응방안	=>	생태계	0.0123	0.6000	36.6000
13	규제	=>	수질	0.0123	0.5000	20.3333
14	수질	=>	규제	0.0123	0.5000	20.3333
15	환경친화	=>	구축	0.0164	0.4000	7.5077
16	구축	=>	환경친화	0.0164	0.3077	7.5077

* 연광성 분석 평가지표
1. 지지도(supprot) = P(X∩Y)
2. 신뢰도(confidence) = P(X∩Y) / P(X)
3. 향상도(lift) = P(X∩Y) / P(X) * P(Y) => lift=1(독립), lift〈1(음의 연관성), lift〉1(양의 연관성)

- 수질오염총량제 시행, 원격탐사기법 활용, 환경친화 기술, 유전자 변형, 전국 폐기물 개선 등의 연구가 활발했음.

2. 키워드 연관성 및 네트워크 분석(2003-2007년)



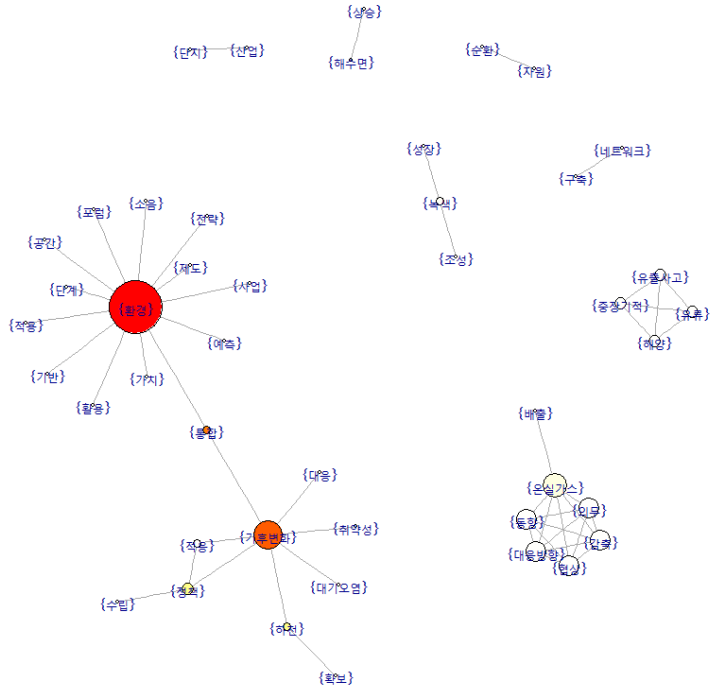
* 키워드 네트워크 분석 결과
1. 원의 크기 : 언급량이 높을수록 크다.
2. 원의 색깔 : 매개중심성이 높을수록 진하다.(하얀색〈노란색〈주황색〈빨간색)

no	lhs		rhs	support	confidence	lift
1	남북	=>	포럼	0.0117	1.0000	68.2000
2	포럼	=>	남북	0.0117	0.8000	68.2000
3	경제통합	=>	환경	0.0147	1.0000	4.5467
4	환경	=>	경제통합	0.0147	0.0667	4.5467
5	영향평가	=>	기후변화	0.0147	0.8333	23.6806
6	기후변화	=>	영향평가	0.0147	0.4167	23.6806
7	총량	=>	오염	0.0117	0.5714	27.8367
8	오염	=>	총량	0.0117	0.5714	27.8367
9	화학물질	=>	유해	0.0117	0.6667	37.8889
10	유해	=>	화학물질	0.0117	0.6667	37.8889
11	자원	=>	분석	0.0117	0.5000	12.1786
12	분석	=>	자원	0.0117	0.2857	12.1786
13	시스템	=>	기후변화	0.0117	0.4444	12.6296
14	기후변화	=>	시스템	0.0117	0.3333	12.6296
15	경제	=>	환경	0.0147	0.5556	2.5259
16	환경	=>	경제	0.0147	0.0667	2.5259

* 연광성 분석 평가지표
1. 지지도(supprot) = P(X∩Y)
2. 신뢰도(confidence) = P(X∩Y) / P(X)
3. 향상도(lift) = P(X∩Y) / P(X) * P(Y) => lift=1(독립), lift〈1(음의 연관성), lift〉1(양의 연관성)

- 기후변화 영향평가 및 적응시스템 구축, 온실가스 배출, 환경경제통합 계정 키워드가 새롭게 등장함.
- 전구간에 이어 유해화학물질, 남북 키워드는 계속 등장함.

3. 키워드 연관성 및 네트워크 분석(2008-2012년)



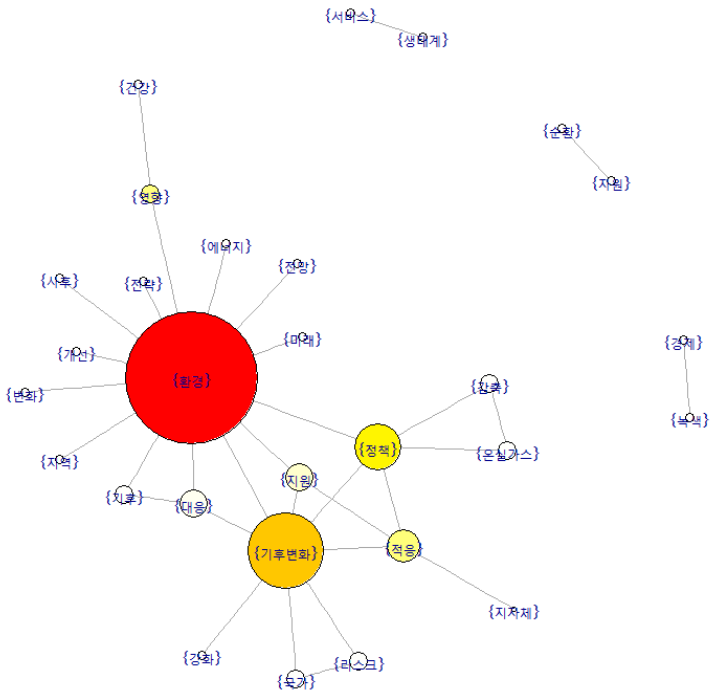
* 키워드 네트워크 분석 결과
1. 원의 크기 : 언급량이 높을수록 크다.
2. 원의 색깔 : 매개중심성이 높을수록 진하다.(하얀색〈노란색〈주황색〈빨간색)

no	lhs		rhs	support	confidence	lift
1	상승	=>	해수면	0.0101	1.0000	99.5000
2	해수면	=>	상승	0.0101	1.0000	99.5000
3	순환	=>	자원	0.0101	1.0000	66.3333
4	자원	=>	순환	0.0101	0.6667	66.3333
5	대응방향	=>	감축	0.0101	1.0000	39.8000
6	감축	=>	대응방향	0.0101	0.4000	39.8000
7	대응방향	=>	온실가스	0.0101	1.0000	22.1111
8	온실가스	=>	대응방향	0.0101	0.2222	22.1111
9	의무	=>	감축	0.0101	1.0000	39.8000
10	감축	=>	의무	0.0101	0.4000	39.8000
11	의무	=>	온실가스	0.0101	1.0000	22.1111
12	온실가스	=>	의무	0.0101	0.2222	22.1111
13	협상	=>	온실가스	0.0101	1.0000	22.1111
14	온실가스	=>	협상	0.0101	0.2222	22.1111
15	중장기적	=>	유출사고	0.0151	1.0000	56.8571
16	유출사고	=>	중장기적	0.0151	0.8571	56.8571

* 연광성 분석 평가지표
1. 지지도(supprot) = P(X∩Y)
2. 신뢰도(confidence) = P(X∩Y) / P(X)
3. 향상도(lift) = P(X∩Y) / P(X) * P(Y) => lift=1(독립), lift〈1(음의 연관성), lift〉1(양의 연관성)

- 기후변화, 온실가스 키워드의 매개중심성이 높아짐.
- 해양 유류 유출사고, 녹색성장 조성, 해수면 상승, 소음 키워드가 새롭게 등장함.

4. 키워드 연관성 및 네트워크 분석(2013-2016년)



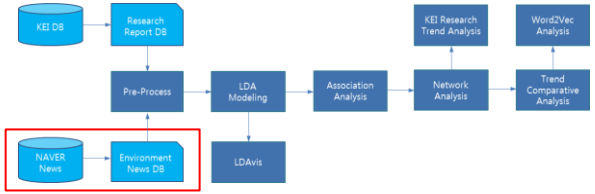
* 키워드 네트워크 분석 결과
1. 원의 크기 : 언급량이 높을수록 크다.
2. 원의 색깔 : 매개중심성이 높을수록 진하다.(하얀색〈노란색〈주황색〈빨간색)

no	lhs		rhs	support	confidence	lift
1	지자체	=>	적응	0.0125	0.7143	8.4244
2	적응	=>	지자체	0.0125	0.1471	8.4244
3	감축	=>	온실가스	0.0175	0.8750	43.8594
4	온실가스	=>	감축	0.0175	0.8750	43.8594
5	감축	=>	정책	0.0125	0.6250	5.8285
6	정책	=>	감축	0.0125	0.1163	5.8285
7	온실가스	=>	정책	0.0125	0.6250	5.8285
8	정책	=>	온실가스	0.0125	0.1163	5.8285
9	녹색	=>	경제	0.0175	0.7000	20.0500
10	경제	=>	녹색	0.0175	0.5000	20.0500
11	서비스	=>	생태계	0.0150	0.6000	16.0400
12	생태계	=>	서비스	0.0150	0.4000	16.0400
13	리스크	=>	국가	0.0125	0.5000	10.0250
14	국가	=>	리스크	0.0125	0.2500	10.0250
15	리스크	=>	기후변화	0.0125	0.5000	3.3417
16	기후변화	=>	리스크	0.0125	0.0833	3.3417

* 연광성 분석 평가지표
1. 지지도(supprot) = P(X∩Y)
2. 신뢰도(confidence) = P(X∩Y) / P(X)
3. 향상도(lift) = P(X∩Y) / P(X) * P(Y) => lift=1(독립), lift〈1(음의 연관성), lift〉1(양의 연관성)

- 전구간에 이어 기후변화 키워드의 매개중심성이 높아짐.
- 환경 키워드와 관련하여 건강, 미래, 전망, 에너지 키워드가 새롭게 등장함.

Text Mining Process List



Plan 2017	Process	Code	Description	Input	Output	Note
6월 상순	**Data Collection**	naver_news1.java naver_news2.java naver_news3.java	- Java jsoup을 사용하여 Web crawling - 조건: 네이버 뉴스 -> 사회-> 환경 - 기간: 2004.1.1~2016.12.12 (총 13개년) - 영역: 제목, 날짜, 언론사 - 양: 193,636개		Naver_news.csv Naver_news_Analysis.xlsx 부록2_네이버 환경 뉴스 언론사별 산출 양	- 2004년 이전 네이버 뉴스 기사 부실

Data Collection

- **네이버 뉴스 〉 사회 〉 환경 관련 기사 전체 수집**
  파싱 도구 : JAVA Jsoup

id	title	year	month	day	time	source	page
a1	[날씨]주말 대체로 흐리고 미세먼지 주의	2016	01	01	10:36:00 PM	아시아경제	0
a2	고창 동림저수지, 거대한 거북이 모양의 가창오리 군무	2016	01	01	8:19:00 PM	뉴시스	1
a3	고창 동림저수지, 하늘을 뒤덮는 가창오리 군무	2016	01	01	7:28:00 PM	뉴시스	2
a4	고창 동림저수지, 가창오리의 화려한 군무	2016	01	01	7:25:00 PM	뉴시스	3
a5	고창 동림저수지, 노을 속 가창오리 군무	2016	01	01	7:24:00 PM	뉴시스	4
a6	고창 동림저수지 가창오리 군무	2016	01	01	7:23:00 PM	뉴시스	5
a7	고창 동림저수지, 가창오리 군무	2016	01	01	7:23:00 PM	뉴시스	6
a8	파타고니아코리아 '쓰레기 없는 바다'	2016	01	01	4:15:00 PM	뉴스1	7
a9	양양 앞바다에 펼쳐진 환경 캠페인	2016	01	01	4:15:00 PM	뉴스1	8
a10	'쓰레기 없는 바다' 메시지 전하는 서퍼들	2016	01	01	4:15:00 PM	뉴스1	9
a11	고속도로, 차량 흐름 대체로 원활…정체 구간은?	2016	01	01	3:47:00 PM	한국경제	10
a12	군산 탁류길 해돋이 문화제	2016	01	01	2:08:00 PM	뉴시스	11
a13	'군산새만금 해맞이 행사'	2016	01	01	2:08:00 PM	뉴시스	12
a14	'행복한 한해가 되기를 바랍니다'	2016	01	01	2:08:00 PM	뉴시스	13
a15	미세먼지와 맞는 새해 첫 날, 수도권·중부내륙 콜록…오후께 개선	2016	01	01	11:54:00 AM	헤럴드경제	14
a16	2016년 새해맞이 인파	2016	01	01	11:09:00 AM	뉴스1	15
a17	소백산 제2연화봉 새해 첫 일출 장관	2016	01	01	10:26:00 AM	뉴시스	16
a18	2016 새해 해돋이 인파	2016	01	01	9:11:00 AM	뉴스1	17
a19	2016 새해 기분 좋은 출발	2016	01	01	9:06:00 AM	뉴스1	18
a20	2016 새해 희망찬 출발	2016	01	01	9:06:00 AM	뉴스1	19
a21	새해 소원 비는 해맞이 관광객들	2016	01	01	9:02:00 AM	뉴스1	0
a22	2016 해돋이 인파로 가득찬 영일대해수욕장	2016	01	01	8:54:00 AM	뉴스1	1
a23	2016 새해를 반기는 시민들	2016	01	01	8:53:00 AM	뉴스1	2
a24	'2016 힘차게 솟아라'	2016	01	01	8:53:00 AM	뉴스1	3
a25	전주 황방산서 멧돼지떼 출몰…1마리 사살	2016	01	02	10:02:00 PM	연합뉴스	0
a26	중부 지방, 극심한 가뭄에 물 확보 '안간힘'	2016	01	02	8:23:00 PM	MBC 뉴스	1
a27	<날씨> 더 포근한 일요일…낮 최고 7~16도(3일)	2016	01	02	8:00:00 PM	연합뉴스	2
a28	3일 가끔 구름 많아…미세먼지 주의	2016	01	02	5:43:00 PM	뉴시스	3
a29	'주말 날씨' 미세먼지 농도↑ 노약자 외출 자제	2016	01	02	3:28:00 PM	데일리안	4
a30	중국 헤이룽장성 진모 6.4 규모 지진, 한국에는 영향 없어	2016	01	02	3:13:00 PM	세계일보	5
a31	회색빛 도심	2016	01	02	10:56:00 AM	뉴시스	6
a32	'다가오는 미세먼지'	2016	01	02	10:56:00 AM	뉴시스	7
a33	'하늘은 흐려도 우리는 즐겁게!'	2016	01	02	10:56:00 AM	뉴시스	8
a34	'미세 먼지도 같이 닦아 볼까'	2016	01	02	10:56:00 AM	뉴시스	9
a35	'미세먼지를 닦자!'	2016	01	02	10:56:00 AM	뉴시스	10
a36	'먼지를 닦자'	2016	01	02	10:56:00 AM	뉴시스	11
a37	'아무 것도 안보이네'	2016	01	02	10:55:00 AM	뉴시스	12
a38	'아무 것도 볼 수 없네'	2016	01	02	10:55:00 AM	뉴시스	13
a39	'새해에도 찾아온 미세먼지'	2016	01	02	10:55:00 AM	뉴시스	14
a40	해수담수화 돌파구 찾을까…6일 첫 대화협의체 모임	2016	01	02	8:41:00 AM	연합뉴스	15
a41	빗나간 원숭이 사랑, 사랑의 끝은 유기	2016	01	02	8:00:00 AM	연합뉴스	16
a42	2일 날씨 전국 구름, 미세먼지 일부 지역 제외 '보통' 예상	2016	01	02	7:16:00 AM	세계일보	17
a43	전북 주요 하천 8곳 생태독성 '이상 무'	2016	01	02	7:00:00 AM	연합뉴스	18
a44	새해 첫 토요일, 전국 흐리고 포근…'안개 주의'	2016	01	02	6:06:00 AM	뉴시스	19
a45	[금주 뉴시스 포토(⑤)]2015 혜조선 굿바이	2016	01	02	6:00:00 AM	뉴시스	0
a46	애완견 '기내·호텔·식당 반입' 놓고 숨바꼭질 성행	2016	01	02	2:13:00 AM	연합뉴스	1
a47	새해 첫출근길 구름 많아…미세먼지 농도 '나쁨'	2016	01	03	8:28:00 PM	헤럴드경제	0

Data Collection

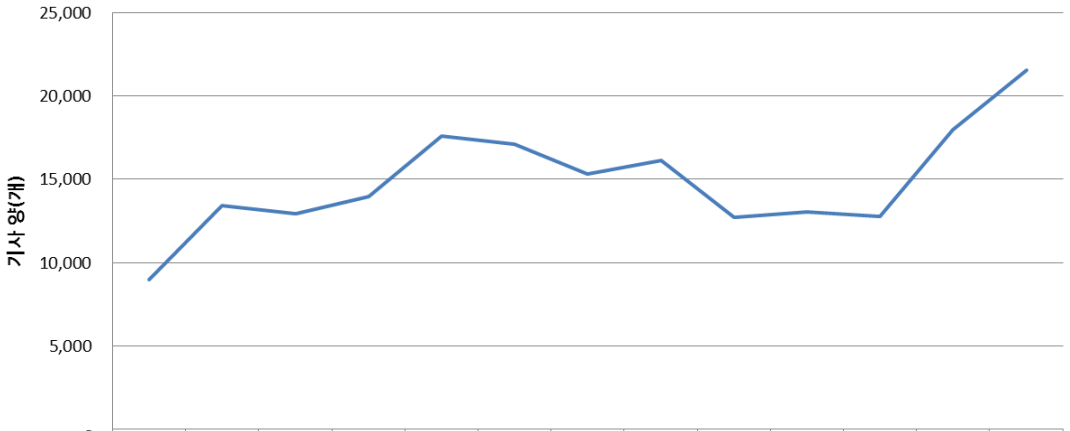
- **네이버 뉴스 기사 데이터 산출 범위**

구분	내용
채널	네이버 뉴스
산출 조건	네이버 뉴스 -> 사회 분야 -> 환경 분야
산출 기간	2004-01-01 00:00:00 ~ 2016-12-12 23:59:59 (총 13개년)
산출 영역	제목, 날짜(년, 월, 일, 시간), 언론사
산출 유형	지면기사, 보도자료
언론사	EBN, EPA연합뉴스, JTBC, KBS 뉴스, MBC IMTV, MBC 뉴스, MBN, OSEN, SBS, SBS CNBC, SBS funE, SBS 뉴스, TV리포트, TV조선, Y-STAR, YTN, YTN 현장생중계, ZDNet Korea, 강원일보, 경향신문, 광주드림, 국민일보, 국정브리핑,내일신문, 노컷뉴스, 뉴스1, 뉴시스, 대전일보, 데일리 서프라이즈, 데일리안, 동아일보, 디지털데일리, 디지털타임스, 라디오코리아, 레이디경향, 마이데일리, 매경이코노미, 매일경제, 매일신문, 머니S, 머니투데이, 문화일보, 미디어오늘, 부산일보, 블로터, 서울경제, 서울신문, 세계일보, 소년한국일보, 스타뉴스, 스포츠경향, 스포츠동아, 스포츠서울, 스포츠서울닷컴, 스포츠조선, 스포츠한국, 시사IN, 시사저널, 신동아 ,아시아경제, 아이뉴스24, 업코리아, 엑스포츠뉴스, 연합뉴스, 연합뉴스 TV, 오마이TV, 오마이뉴스, 이데일리, 이코노미21, 이코노믹리뷰, 인터뷰365, 일간스포츠(OLD), 일다, 전자신문, 제주일보, 조선비즈, 조선일보, 조세일보, 주간경향, 주간동아, 주간한국, 중앙SUNDAY, 중앙일보, 참세상, 참세상 vod, 컬처뉴스, 코메디닷컴 , 쿠키뉴스, 파이낸셜뉴스, 팝뉴스, 프라임경제, 프레시안, 프로메테우스, 한겨레, 한겨레21, 한국경제, 한국경제TV, 한국일보, 헤럴드POP, 헤럴드경제, 헬스조선 (총 101개)
산출 양	193,636개

Data Collection

- 환경분야 네이버 뉴스 기사 데이터 기초 분석

〈 연도별 환경분야 네이버뉴스 산출 양 추이〉



	2004년	2005년	2006년	2007년	2008년	2009년	2010년	2011년	2012년	2013년	2014년	2015년	2016년
네이버뉴스	9,013	13,452	12,915	13,971	17,595	17,114	15,342	16,161	12,724	13,021	12,759	17,998	21,571

연구 개요
선행 연구
연구 내용
연구 추진방법
기대효과

Data Collection

- 환경분야 네이버 뉴스 기사 데이터 기초 분석

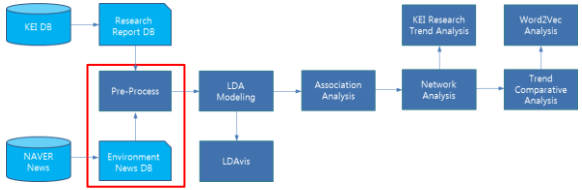
〈 101개 언론사별 환경분야 네이버뉴스 산출 양〉

연도 / 언론사	'04	'05	'06	'07	'08	'09	'10	'11	'12	'13	'14	'15	'16	합계
EBN	-	7	2	4	5	-	-	-	-	-	-	-	-	18
EPA연합뉴스	-	1	-	-	-	-	-	-	-	-	-	-	-	1
JTBC	-	-	-	-	-	-	-	-	-	16	-	-	3	19
KBS 뉴스	-	-	-	-	-	-	-	-	-	34	7	3	5	49
MBC iMTV	-	-	-	4	-	-	-	-	-	-	-	-	-	4
MBC 뉴스	-	-	130	297	189	56	-	-	2	89	1	29	10	803
MBN	3	83	105	75	38	1	-	-	-	26	3	-	-	334
OSEN	-	-	1	-	-	-	-	-	-	1	6	-	-	8
SBS	30	74	36	57	50	3		1	4	4	3	1	-	263
SBS CNBC	-	-	-	-	-	-	-	-	-	1	-	-	-	1
SBS funE	-	-	-	-	-	-	-	-	-	-	1	-	-	1
SBS 뉴스	89	165	348	501	307	77	4	1	21	59	-	2	3	1.577
TV리포트	-	-	18	3	-	-	-	-	1	-	-	-	-	22
TV조선	-	-	-	-	-	-	-	-	-	-	4	-	1	5
Y-STAR	-	1	-	-	-	-	-	-	-	-	-	-	-	1
YTN	629	996	624	628	382	46	4	5	100	44	3	15	-	3.476
YTN 현장생중계	-	1	-	-	-	-	-	-	-	-	-	-	-	1
ZDNet Korea	-	-	-	-	3	1	-	-	-	-	-	-	-	4
강원일보	21	8	-	1	1	-	-	-	3	1	-	-	-	35
경향신문	312	509	404	313	508	589	785	1.205	1.053	1.103	519	309	157	7.746
광주드림	-	-	-	-	46	20	-	-	-	-	-	-	-	66
국민일보	80	263	217	170	227	32	-	3	3	77	-	6	83	1.161
국정브리핑	5	5	-	1	-	-	-	-	-	-	-	-	-	11
내일신문	82	373	546	622	309	508	318	296	208	20	-	-	-	3.282
노컷뉴스	151	711	480	892	476	390	274	378	291	352	90	4	2	4.491
뉴스1	-	-	-	-	-	-	-	1	623	1.509	1.855	2.546	3.731	10.265
뉴시스	233	-	985	4.314	7.129	8.472	7.334	7.718	3.959	2.473	2.680	3.731	7.970	56.998
대전일보	1	5	4	1	1	-	-	-	-	-	-	-	-	12
데일리 서프라이즈	-	-	10	1	4	-	-	-	-	-	-	-	-	15
데일리안	-	8	17	18	28	-	-	-	-	10	1	15	38	135
동아일보	145	281	162	100	187	232	121	240	135	71	21	492	335	2.522
디지털데일리	-	4	-	-	-	1	-	-	-	-	-	-	-	5
디지털타임스	2	14	2	1	2	1	-	-	3	2	2	-	-	29

연도 / 언론사	'04	'05	'06	'07	'08	'09	'10	'11	'12	'13	'14	'15	'16	합계
라디오코리아	-	-	-	1	-	-	-	-	-	-	-	-	-	1
레이디경향	-	-	-	-	1	-	-	-	-	-	-	-	-	1
마이데일리	-	9	-	-	26	-	-	-	-	-	-	-	-	35
매경이코노미	-	-	4	-	-	-	-	-	-	-	-	-	-	4
매일경제	47	184	76	89	50	12	2	4	7	14	5	1	1	492
매일신문	31	27	15	10	-	2	-	-	-	-	3	-	-	88
머니S	-	-	-	-	-	-	-	-	-	-	4	-	-	4
머니투데이	56	181	126	168	94	12	-	2	10	23	-	-	1	673
문화일보	336	522	201	143	107	38	3	1	-	23	1	2	77	1.454
미디어오늘	4	-	-	-	6	-	-	-	-	-	-	-	-	10
부산일보	79	106	108	38	3	4	-	1	7	28	2	-	-	376
불교타	-	-	-	1	1	1	-	-	-	-	-	-	-	3
서울경제	78	171	172	142	76	6	-	-	1	15	-	17	43	721
서울신문	48	169	239	131	153	46	1	-	6	20	-	1	-	814
세계일보	151	451	387	294	259	291	406	449	621	678	940	1.027	571	6.525
소년한국일보	1	1	-	-	1	-	-	-	-	-	-	-	-	3
스타뉴스	-	7	1	-	-	-	-	-	-	-	-	-	-	8
스포츠경향	-	3	7	2	2	-	2	-	-	-	-	-	-	16
스포츠동아	-	-	-	-	-	-	-	-	-	1	4	-	-	5
스포츠서울	-	-	-	-	-	-	-	-	-	-	-	36	48	84
스포츠서울닷컴	13	-	4	9	5	-	1	-	2	7	-	-	-	41
스포츠조선	5	2	13	10	1	-	-	-	-	-	-	-	-	31
스포츠한국	-	-	-	-	-	-	-	-	-	1	-	-	-	1
시사IN	-	-	-	-	-	3	-	-	-	-	-	-	-	3
시사저널	-	6	12	5	2	-	-	-	-	-	-	-	-	25
신동아	-	-	-	7	3	-	-	-	-	-	-	-	-	10
아시아경제	-	-	-	38	91	15	525	846	721	928	676	532	244	4.616
아이뉴스24	1	-	2	13	12	-	1	9	-	9	3	3	46	99
업코리아	4	22	8	-	-	-	-	-	-	-	-	-	-	34
엑스포츠뉴스	-	-	-	-	-	-	-	-	2	1	1	-	-	4
연합뉴스	4.330	4.716	2.760	2.205	4.529	4.445	4.101	3.258	2.616	3.563	3.345	7.431	6.543	53.842
연합뉴스 TV	-	22	83	30	1	-	-	-	-	72	-	119	60	387
오마이TV	-	13	4	-	-	-	-	-	-	-	-	-	-	17
오마이뉴스	186	414	317	135	183	25	3	2	9	68		1	1	1.344
이데일리	24	80	26	12	10	1	-	-	12	33	464	342	277	1.281
이코노미21	-	-	-	3	-	-	-	-	-	-	-	-	-	3
이코노미리뷰	-	-	5	-	-	-	-	-	-	-	-	-	-	5
인터뷰365	-	-	-	-	4	-	-	-	-	-	-	-	-	4

연도 / 언론사	'04	'05	'06	'07	'08	'09	'10	'11	'12	'13	'14	'15	'16	합계
일간스포츠(OLD)	6	-	-	-	-	-	-	-	-	-	-	-	-	6
일다	23	9	20	12	6	49	17	38	38	21	23	51	7	294
전자신문	3	6	-	4	9	3	-	-	-	-	143	7	1	176
제주일보	2	-	-	2	-	-	-	-	-	-	-	-	-	4
조선비즈	-	-	-	-	-	-	-	-	-	4	4	-	-	8
조선일보	-	-	-	-	-	-	-	-	-	-	-	-	11	11
조세일보	-	3	-	-	-	-	-	-	-	-	-	-	-	3
주간경향	7	46	94	25	55	-	-	2	-	-	-	1	-	230
주간동아	4	-	-	12	4	2	1	-	-	1	-	-	2	26
주간한국	6	15	9	9	3	-	2	1	-	-	-	-	-	45
중앙SUNDAY	-	-	-	-	-	-	-	-	-	-	-	-	1	1
중앙일보	-	-	-	-	-	-	-	-	-	-	-	-	21	21
참세상	-	-	30	4	1	9	16	1	3	2	-	-	6	72
참세상 vod	-	-	11	1	-	-	-	-	-	-	-	-	-	12
컬쳐뉴스	-	-	-	-	1	-	-	-	-	-	-	-	-	1
코메디닷컴	-	-	-	-	1	2	-	-	-	-	-	-	-	3
쿠키뉴스	-	121	537	285	252	34	-	1	1	13	-	-	-	1.244
파이낸셜뉴스	149	164	80	128	103	46	27	90	4	8	6	30	214	1.049
팜뉴스	-	1	2	3	2	1	-	-	-	-	-	-	-	9
프라임경제	-	7	18	5	-	-	-	-	-	-	-	-	-	30
프레시안	88	197	180	20	32	2	-	1	2	3	-	-	-	525
프로메테우스	-	144	246	85	-	-	-	-	-	-	-	-	-	475
한겨레	808	956	1.954	885	548	329	544	725	1.214	739	754	630	352	10.418
한겨레21	44	35	42	22	51	14	7	6	5	2	2	1	-	231
한국경제	524	642	649	497	433	618	343	497	783	477	509	456	355	6.783
한국경제TV	-	8	6	4	5	-	-	-	10	3	7	-	1	44
한국일보	154	323	297	380	501	636	499	378	239	318	335	54	54	4.168
헤럴드POP	-	84	50	72	45	32	-	1	5	42	129	-	-	460
헤럴드경제	18	96	29	27	31	7	1	-	-	12	202	123	316	862
헬스조선	-	-	-	-	-	-	-	-	-	-	1	-	-	1
총합계	9.013	13.452	12.915	13.971	17.595	17.114	15.342	16.161	12.724	13.021	12.759	17.998	21.571	193.636

Text Mining Process List



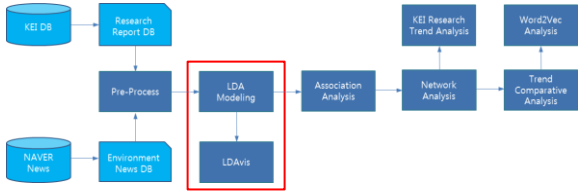
Plan 2017	Process	Code	Description	Input	Output	Note
6월 상순	**Pre-processing(2)**	topic_clustering.R	- 형태소분석기 실행(KoNLP 등) - Low TF-IDF 값 제거 - 불용어처리 등 전처리 과정 (특정 단어 삭제, 특수문자 제거, 소문자로 변경 등) - Word Lengths는 2글자 이상 - 동의어 처리	naver.xlsx	out_naver.csv DocumentTermMatrix 부록1_제거 대상 키워드 목록.hwp	-자문의견(이명진 박사님) : 한글 처리 문제 -> 다양한 한글 전처리 방법을 통해 해결 가능함.

Pre-processing(2)

〈부록〉 제거 대상 키워드 목록

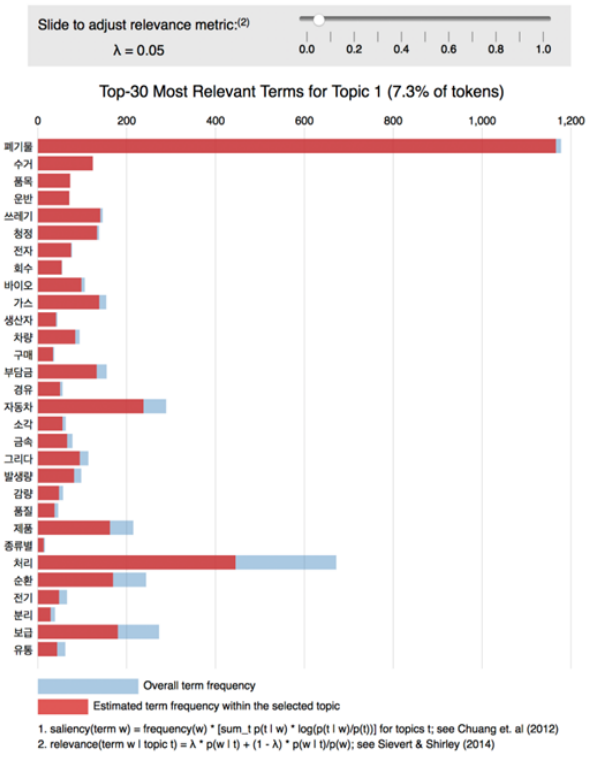
채널	제거 대상 키워드
네이버 환경 뉴스	날씨, 전국, 환경, 오후, 내일, 뉴스, 오늘, 주말, 기상, 관리, 곳곳, 아침, 지역, 영향, 사업, 올해, 국내, 일부, 규모, 종합, 개발, 주변, 조사, 우리, 대상, 활동, 연구, 회의, 생활, 마련, 처리, 가능, 첫날, 최고, 전면, 확인, 수준, 작업, 발령, 경향, 필요, 도입, 활용, 발표, 준비, 효과, 협력, 때문, 잇따, 시스템, 제품, 사회, 하게, 주요, 현상, 마지막, 이후, 대표, 시대, 개월, 단계, 계획, 위해, 가지, 구간, 언제, 통합, 운영, 개체, 차례, 아래, 프로그램, 구역, 기록, 등록, 보고, 연속, 이전, 하기, 재차, 이름, 반기, 들이, 양식, 부분, 누구, 목표, 구조, 기관, 이야기, 중심, 재개, 가득, 설명, 평년기온, 건립, 다양, 가운데, 업무, 다음, 모습, 공간, 하나, 기간, 완화, 초록, 행위, 구경, 공식, 주춤, 구상, 시행, 유의, 일반, 동안, 사전, 시내, 저녁, 낮, 오전, 과정, 최종, 진입, 작전, 자동, 연간, 제도, 특집, 현실, 구름많고, 구축, 방식, 본부, 생각, 선언, 중요, 포함, 사례, 일보, 중순, 노력, 개화, 표지, 쌀쌀, 월일, 유명, 기획, 광역, 그림, 기대, 구성, 관찰, 가로, 수립, 이젠, 전략, 사건, 제외, 추정, 하루, 이틀, 삼일, 월요일, 화요일, 수요일, 목요일, 금요일, 토요일, 일요일, 1월, 2월, 3월, 4월, 5월, 6월, 7월, 8월, 9월, 10월, 11월, 12월, 나흘째, 삼일째, 이틀째, 각종, 이곳, 저곳, 사흘째, 사흘, 나흘, 진짜, 모두, 분야, 표시, 특유, 정기, 단기, 구석, 근본, 기본, 기초, 파악, 수도, 모양, 특정, 데이터, 질문, 채택, 정도, 일교차, 일교, 소식, 전달, 전원, 정신, 직접, 발달, 제기, 선택, 개념, 내부, 봄기운, 이하, 업종, 역할, 어제, 다음주, 이번주, 이번, 금주, 내년, 작년, 세계 **(총 233개)**

Text Mining Process List



Plan 2017	Process	Code	Description	Input	Output	Note
6월 상순	**LDA Modeling(2)**	topic_clustering.R	- LDA기반 토픽 모델링 - 토픽별 핵심 단어 출력 - 문서별 토픽번호 및 확률값 출력 - 단어별 토픽번호 및 확률값 출력	Document TermMatrix	news_term_topic.csv news_doc_Prob_df.csv news_doc_prob_df_max.csv news_id_topic.csv news_lda_tm.csv	- 입력값 : SEED = 2000000 K = 10
6월 상순	**LDAvis(2)**	topic_clustering.R	- 토픽모델링 - 2차원 시각화 및 주요 키워드 확률분포 목록 시각화	lda_tm.csv	HTML 등 웹파일	- apache-tomcat-8.5.12 사용 - 산출물 서버업로드 필요

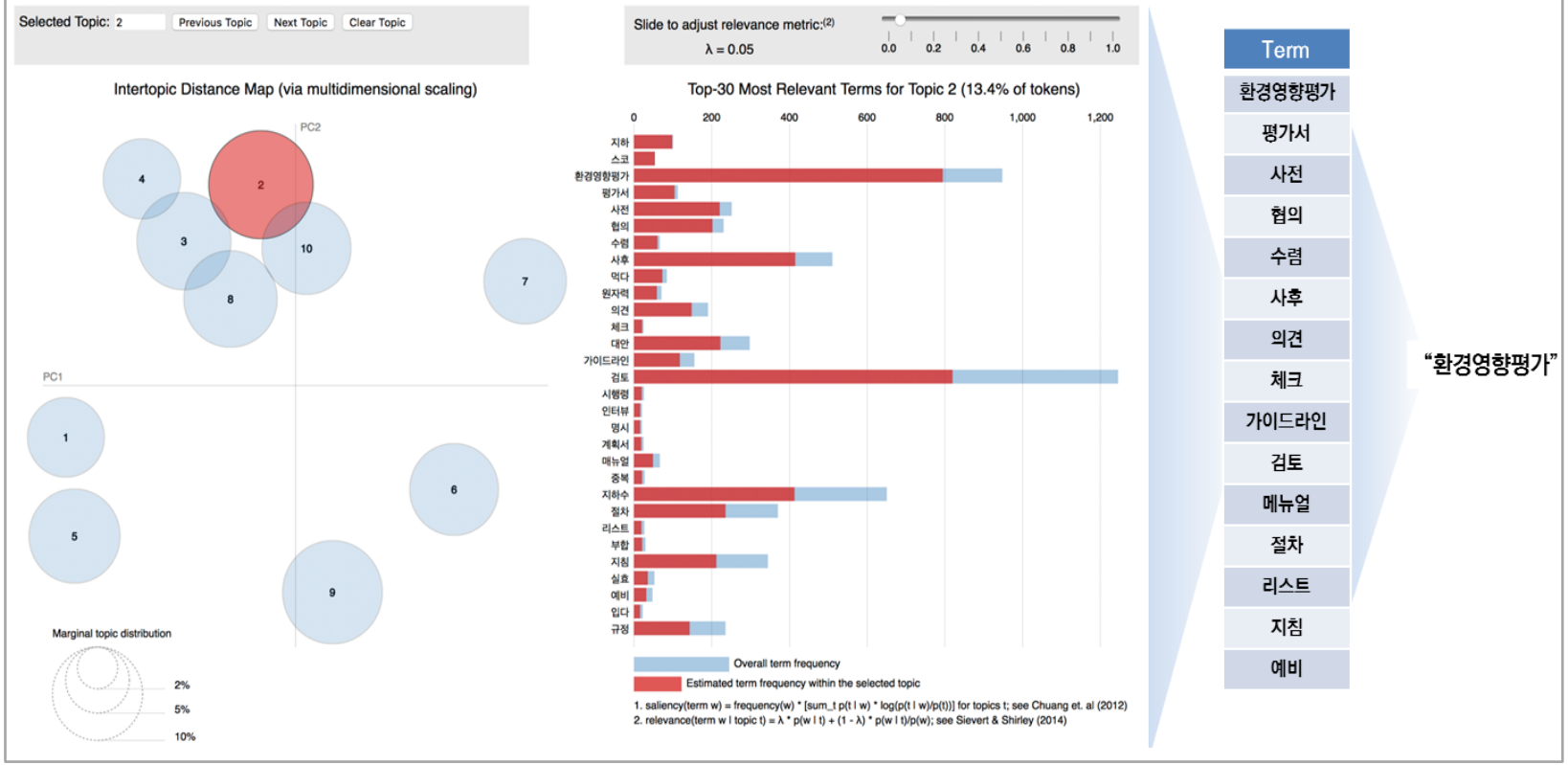
LDAvis (Topic 1)



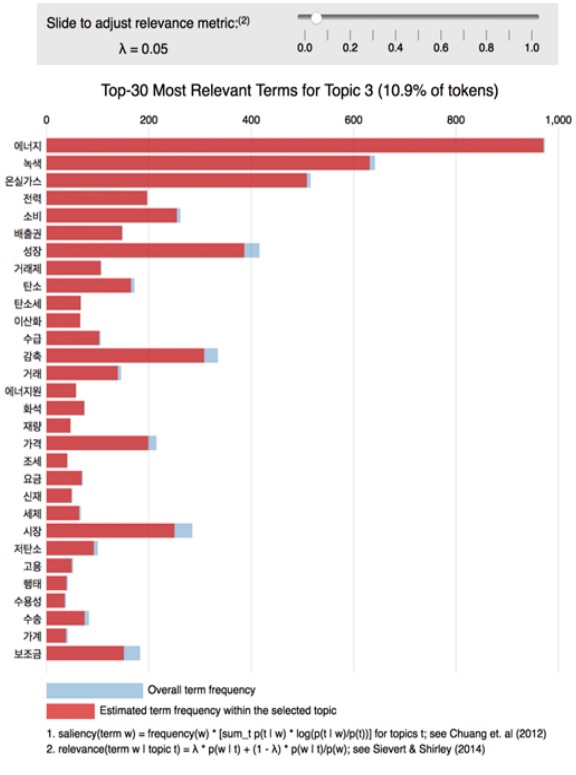
Term
폐기물
수거
운반
쓰레기
청정
가스
차량
부담금
경유
자동차
소각
발생량
처리
순환
전기

“폐기물”

LDAvis (Topic 2)



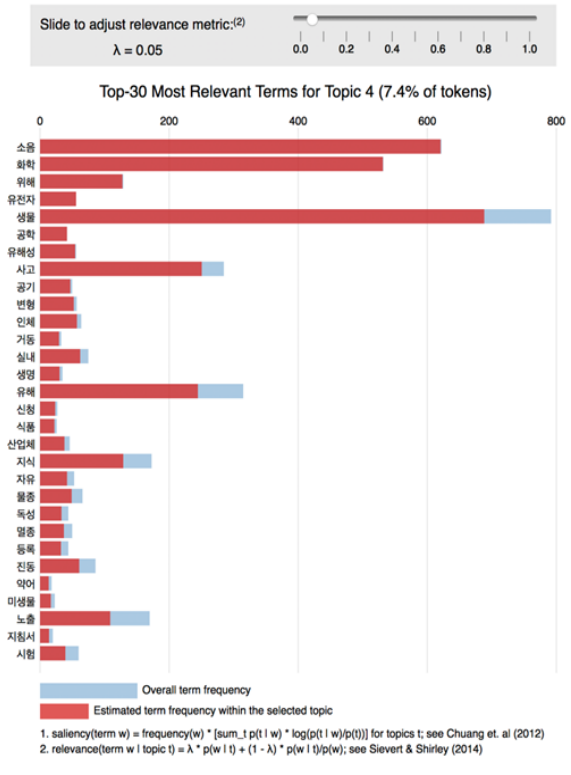
LDAvis (Topic 3)



Term
에너지
녹색
온실가스
전력
배출권
거래제
이산화탄소
탄소세
감축
에너지원
화석
신재생에너지
세제
저탄소
보조금

“에너지 자원”

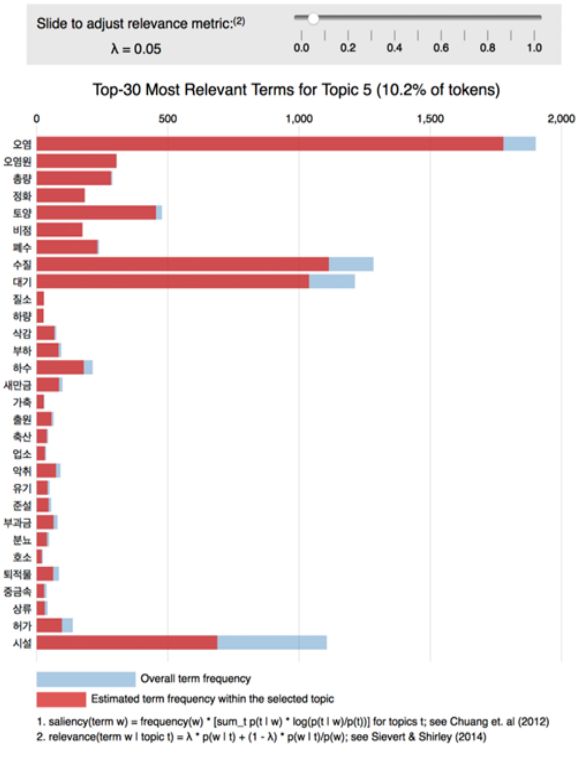
LDAvis (Topic 4)



Term
소음
화학
유전자
생물
공학
유해성
변형
인체
생명
식품
독성
멸종
진동
미생물
시험

“유전자 변형”
+
“소음”

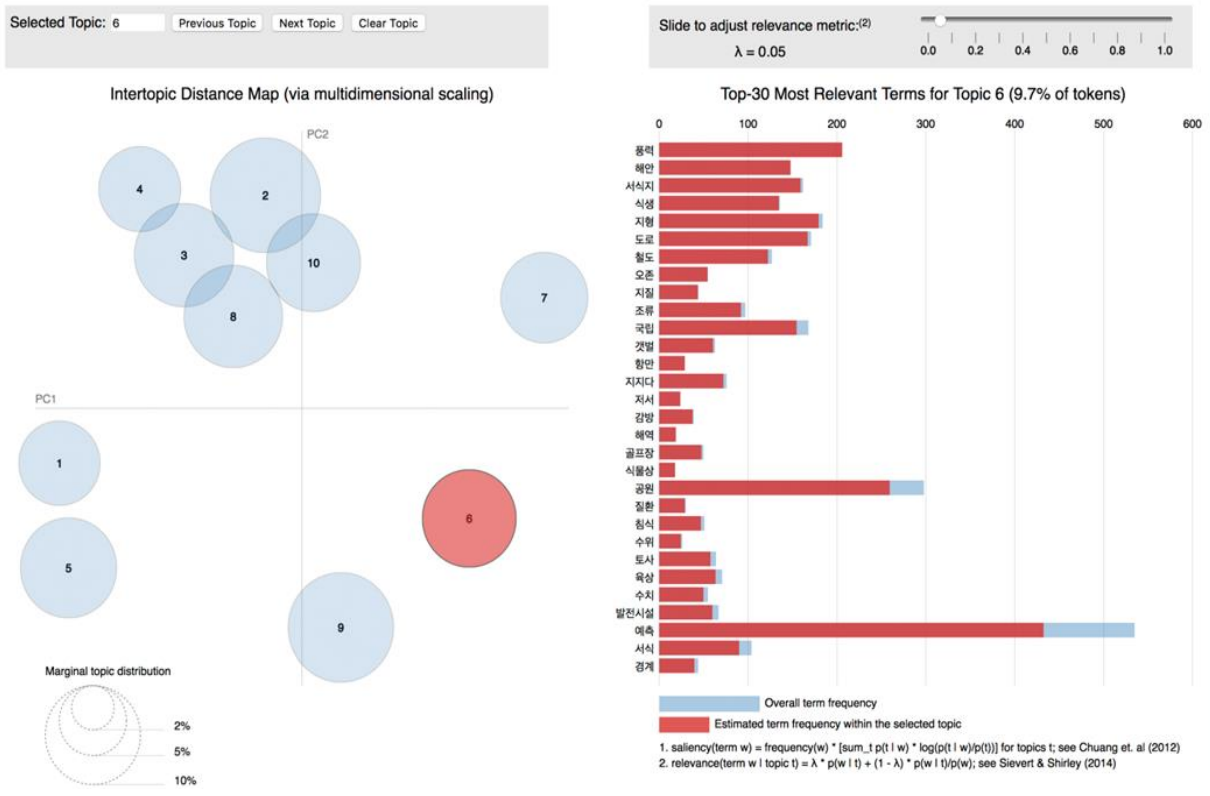
LDAvis (Topic 5)



Term
오염
오염원
정화
폐수
수질
대기
하수
새만금
가축
축산
악취
분뇨
퇴적물
중금속
상류

“수질오염”

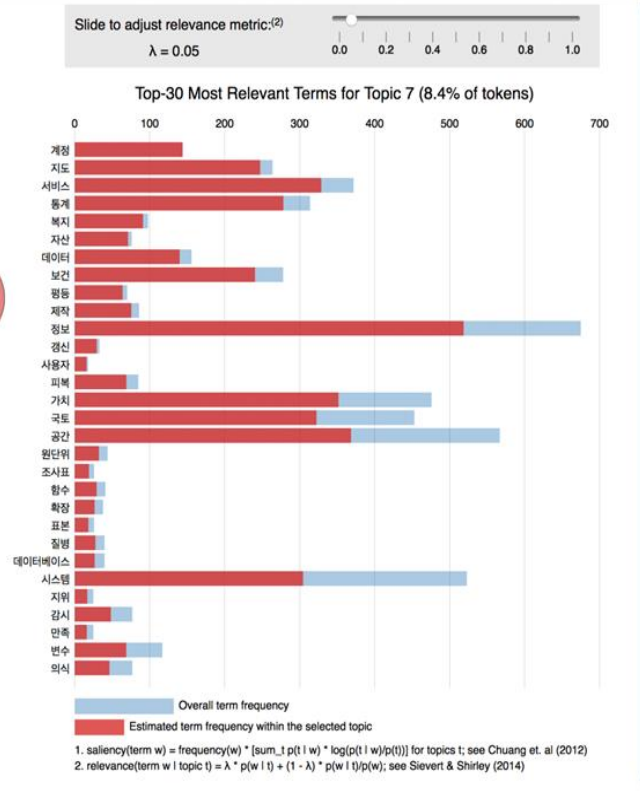
LDAvis (Topic 6)



Term
풍력
해안
서식지
식생
지형
지질
조류
갯벌
항만
해역
침식
수위
토사
육상
발전시설

“해양” + “풍력”

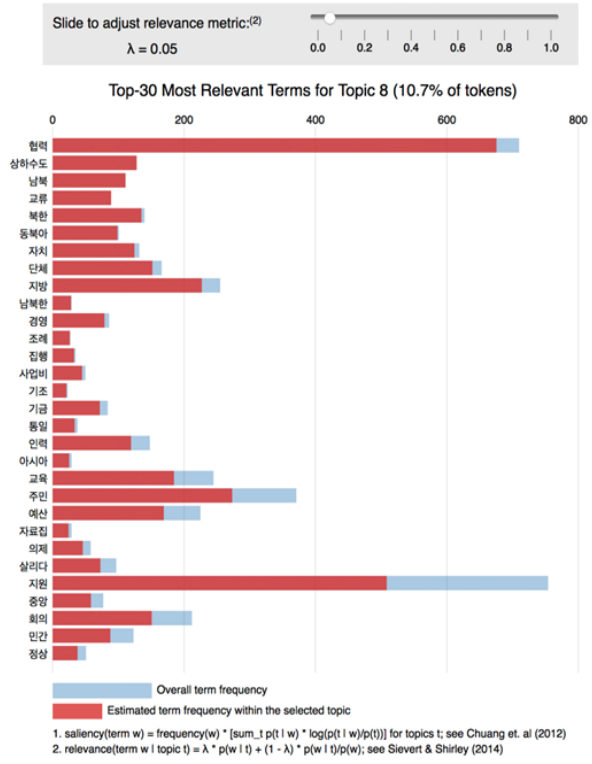
LDAvis (Topic 7)



Term
서비스
통계
복지
데이터
보건
정보
피복지도
국토
공간
조사표
함수
표본
질병
데이터베이스
변수

“보건”
+
“데이터”

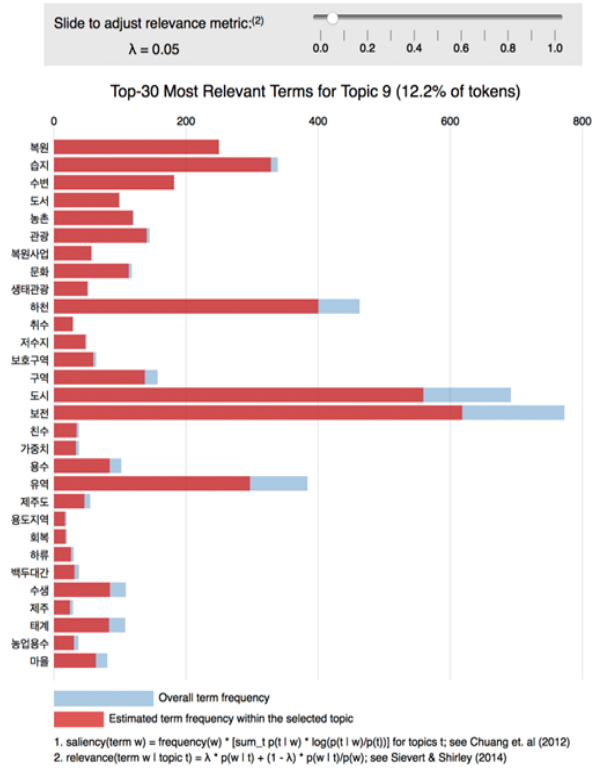
LDAvis (Topic 8)



Term
협력
상하수도
남북
교류
북한
동북아
남북한
경영
통일
아시아
의제
중앙
회의
민간
정상

"대외협력"

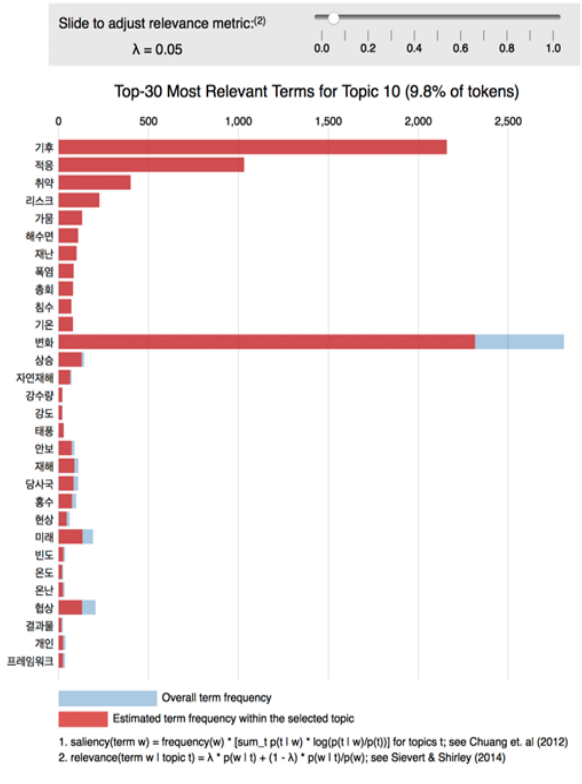
LDAvis (Topic 9)



Term
복원
습지
수변
복원사업
하천
취수
저수지
친수
용수
유역
하류
백두대간
수생
제주
농업용수

“물환경”

LDAvis (Topic 10)



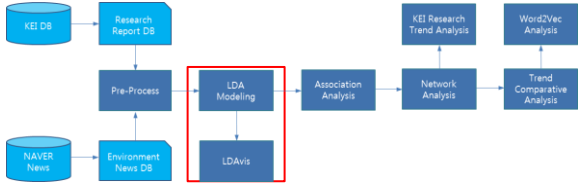
Term
기후
적응
취약
리스크
가뭄
재난
폭염
침수
기온
변화
상승
자연재해
태풍
홍수
온난

“기후변화”

LDAvis (Topic 전체)

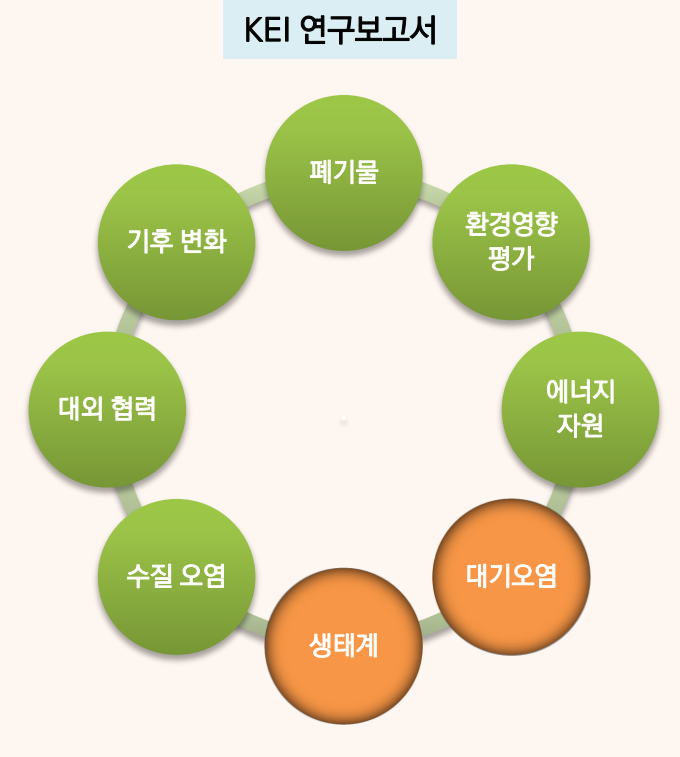
No.	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
Title	폐기물	환경영향 평가	에너지 자원	유전자 변형, 소음	수질오염	해양, 풍력	보건, 데이터	대외협력	물 환경	기후 변화
1	폐기물	환경영향 평가	에너지	소음	오염	풍력	서비스	협력	복원	기후
2	수거	평가서	녹색	화학	오염원	해안	통계	상하수도	습지	적응
3	운반	사전	온실가스	유전자	총량	서식지	복지	남북	수변	취약
4	쓰레기	협의	전력	생물	정화	식생	데이터	교류	복원사업	리스크
5	청정	수렴	배출권	공학	토양	지형	보건	북한	하천	가뭄
6	가스	사후	거래제	유해성	폐수	지질	정보	동북아	취수	재난
7	차량	의견	이산화탄소	변형	대기	조류	피복지도	남북한	저수지	폭염
8	부담금	체크	탄소세	인체	새만금	갯벌	국토	경영	친수	침수
9	경유	가이드라인	감축	생명	축산	항만	공간	통일	용수	기온
10	자동차	검토	에너지원	식품	악취	해역	조사표	아시아	유역	변화
11	소각	매뉴얼	화석	독성	부과금	침식	함수	의제	하류	상승
12	발생량	절차	신재생에너지	멸종	분뇨	수위	표본	중앙	백두대간	자연재해
13	처리	리스트	세제	진동	호소	토사	질병	회의	수생	태풍
14	순환	지침	저탄소	미생물	퇴적물	육상	데이터베이스	민간	제주	홍수
15	전기	예비	보조금	시험	중금속	발전시설	변수	정상	농업용수	온난
...										

Text Mining Process List



Plan 2017	Process	Code	Description	Input	Output	Note
6월 하순	**LDA Result Analysis(2)**		– 토픽별 키워드 분석 – 토픽별 네이버 환경뉴스 동향 분석	news_id_topic.csv	news_id_topic_Analysis.xlsx	–2004~2016년 네이버 뉴스 기사 연도별 동향 분석 – 2004~2016년 네이버 뉴스와 KEI 연구보고서 비교 분석
중간자문회의(2017.06.29)						

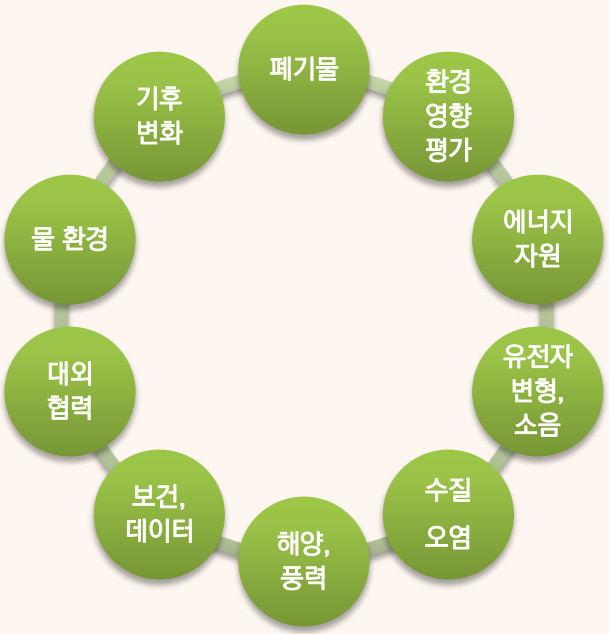
KEI 연구보고서 LDAvis (2004~2016)



No.	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
Title	폐기물	환경영향 평가	에너지 자원	대기오염	생태계	수질오염	대외협력	기후변화
1	폐기물	영향	에너지	대기	생태	지하수	협력	적응
2	처리	제도	전력	화학	생물	수질	협상	폭염
3	시설	사후	온실가스	먼지	습지	비점	포럼	해수면
4	배출	검토	원료	오염	서식지	오염원	아세안	침수
5	하수	개선	석탄	초미세	자연환경	새만금	동북아	범람
6	쓰레기	정책	화석	천식	식물	용담댐	의정서	태풍
7	발생량	주민	재생	황사	서식	녹조	교토	기상이변
8	폐수	지역	천연가스	방사능	외래	취수	베트남	자외선
9	종량제	갈등	신재생	미세먼지	야생동물	농업용수	국제	진동
10	악취	설문	화력	방사성	멸종	하량	아시아	자연재해

매체별 LDAvis 결과 비교(2004-2016)

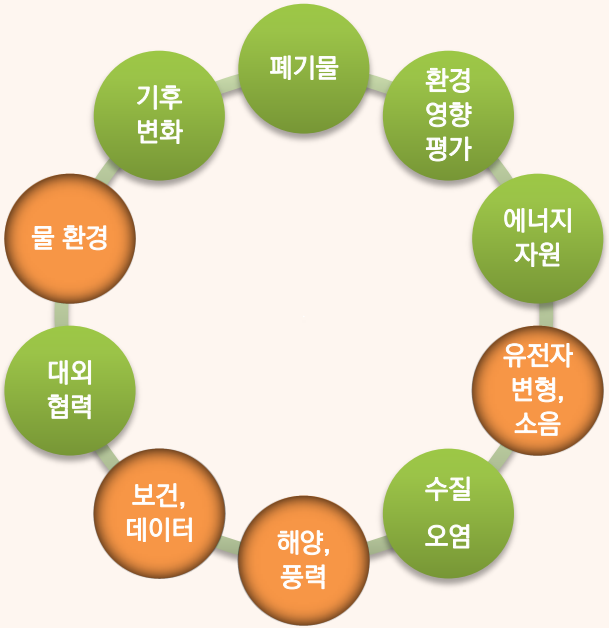
NAVER 환경뉴스



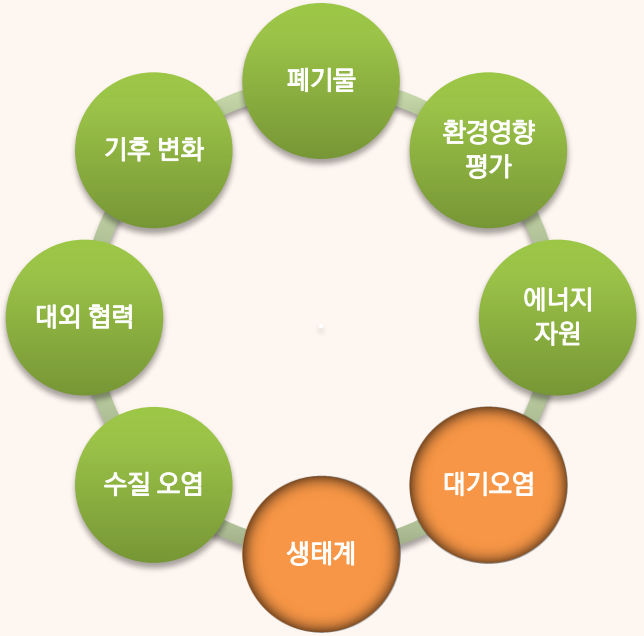
KEI 연구보고서

매체별 LDAvis 결과 비교(2004-2016)

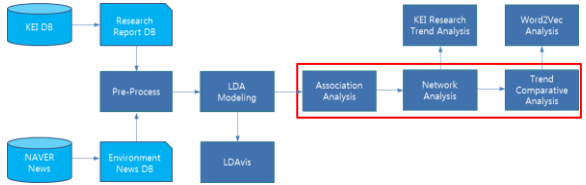
NAVER 환경뉴스



KEI 연구보고서

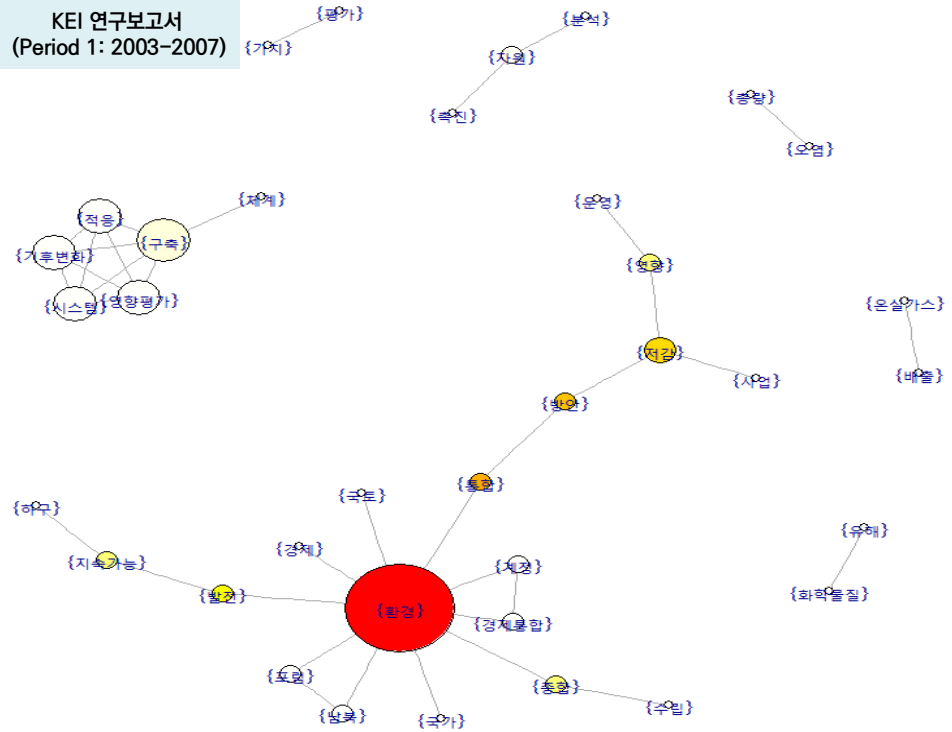


Text Mining Process List



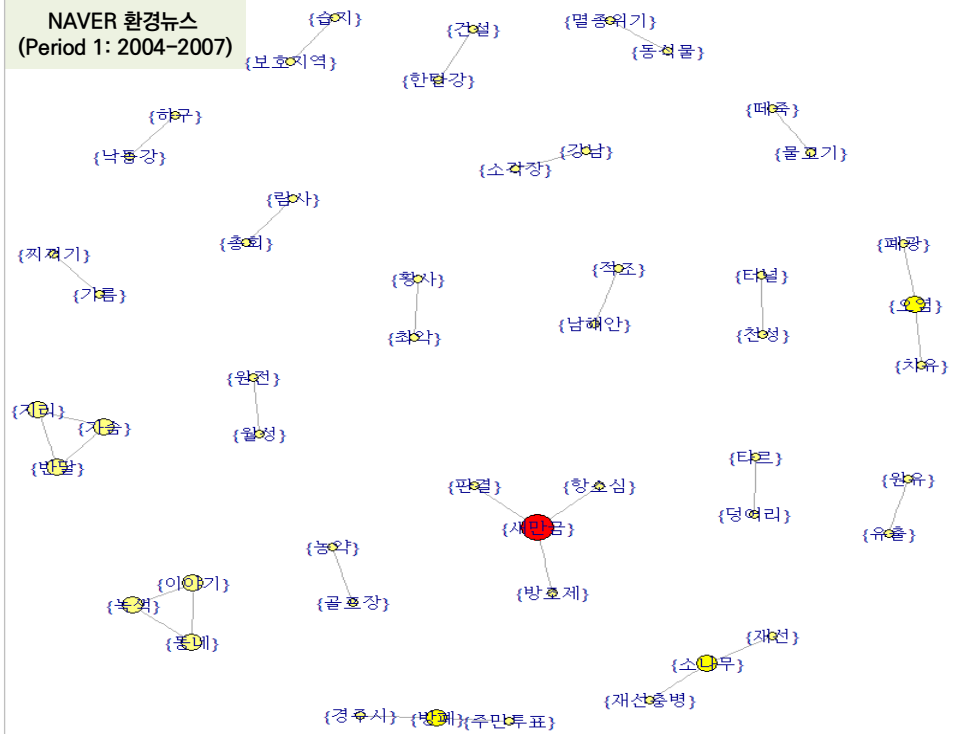
Plan 2017	Process	Code	Description	Input	Output	Note
7월	**Association Analysis(2)**	Association_Analysis.R	– 지지도, 신뢰도가 0.01 이상 값 출력 – 3가지측도(지지도, 신뢰도, 향상도) 분석	2004_2007.txt 2008_2012.txt 2013_2016.txt	Association.xlsx	– 네이버뉴스 제목 데이터 활용 – 본문으로 분석시 메트릭스가 너무 커짐 – 3개 시기별 동향 분석
	Network Analysis(2)	Association_Analysis.R	– 원의 크기 : 언급량이 많을수록 크기가 큼 – 원의 색깔 : 매개중심성이 높을수록 색깔이 진함		04-07.png 08-12.png 13-16.png	
8월 상순	**Trend Comparative Analysis**		– 매체별 동향을 3개 시기로 나눠서 비교 – 3개시기(2004~2007, 2008~2012, 2013~2016)	Association.xlsx Naver_Association.xlsx		– 매체별 키워드 연관성 및 네트워크 분석 결과 활용 – 3개 시기 분류 기준: 대통령 재임기간 – 분석결과 '기후변화'가 중요 키워드로 나타남

KEI 연구보고서 (Period 1: 2003-2007)



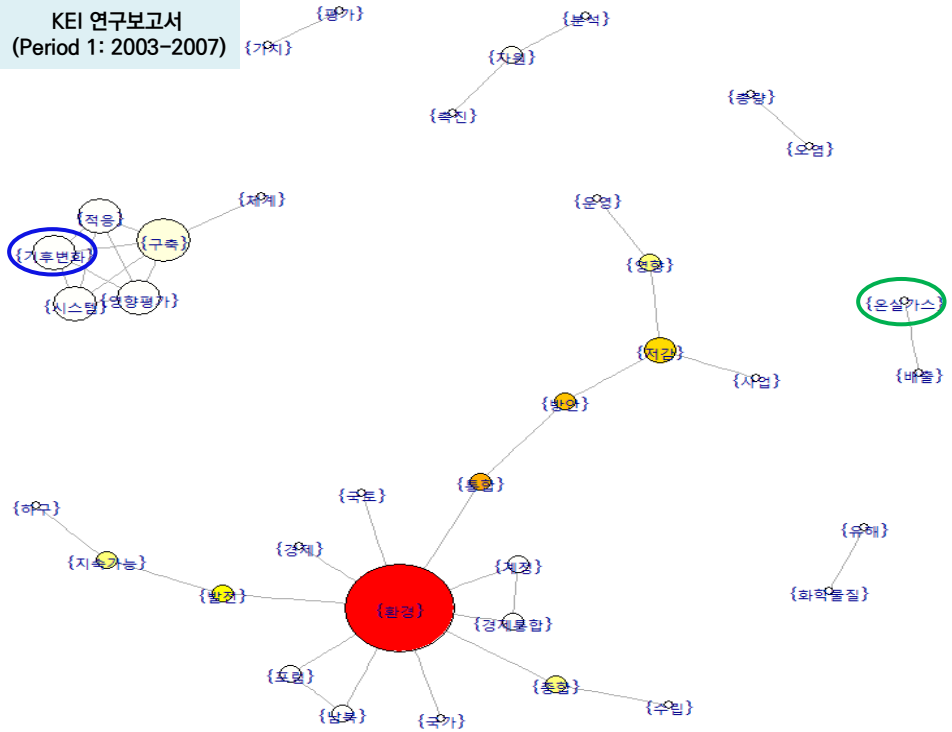
	A		B	신뢰도	지지도	향상도
1	하구	=>	지속가능	0.015	0.833	21.859
2	지속가능	=>	하구	0.015	0.385	21.859
3	경제통합	=>	환경	0.015	1.000	4.547
4	환경	=>	경제통합	0.015	0.067	4.547
5	경제	=>	환경	0.015	0.556	2.526
6	환경	=>	경제	0.015	0.067	2.526
7	남북	=>	포럼	0.012	1.000	68.200
8	포럼	=>	남북	0.012	0.800	68.200
9	자원	=>	분석	0.012	0.500	12.179
10	분석	=>	자원	0.012	0.286	12.179
11	시스템	=>	기후변화	0.012	0.444	12.630
12	기후변화	=>	시스템	0.012	0.333	12.630

NAVER 환경뉴스 (Period 1: 2004-2007)



	A		B	신뢰도	지지도	향상도
1	동식물	=>	멸종위기	0.002	0.620	63.038
2	멸종위기	=>	동식물	0.002	0.235	63.038
3	주민투표	=>	방폐	0.002	0.671	32.029
4	방폐	=>	주민투표	0.002	0.098	32.029
5	방조제	=>	새만금	0.002	0.870	49.356
6	새만금	=>	방조제	0.002	0.108	49.356
7	남해안	=>	적조	0.002	0.489	60.799
8	적조	=>	남해안	0.002	0.231	60.799
9	월성	=>	원전	0.002	0.884	80.033
10	원전	=>	월성	0.002	0.139	80.033
11	경주시	=>	방폐	0.001	0.821	39.163
12	방폐	=>	경주시	0.001	0.062	39.163

KEI 연구보고서
(Period 1: 2003-2007)



1. 기후변화 2. 온실가스 3. 태안 기름 유출 사고 4. 녹색성장 5. 환경오염 대책

A. 기후변화 영향평가 및 적응시스템 구축, 온실가스 배출, 환경경제통합계정, 유해화학물질, 남북포럼 키워드가 등장함.

B. **기후변화** 키워드를 중심으로 영향평가, 적응시스템 구축의 키워드가 나타남.

NAVER 환경뉴스
(Period 1: 2004-2007)

A. 동식물-멸종위기, 지리산-반달가슴, 습지-보호지역, 람사르-총회

- 2004년 멸종위기 동식물 보호강화를 위해 멸종위기동식물의 범위를 대폭 확대함.
- 2006년 환경부에서 멸종위기 동식물 54종 증식복원사업 추진함.

B. **새만금**-판결-항소심-방조제

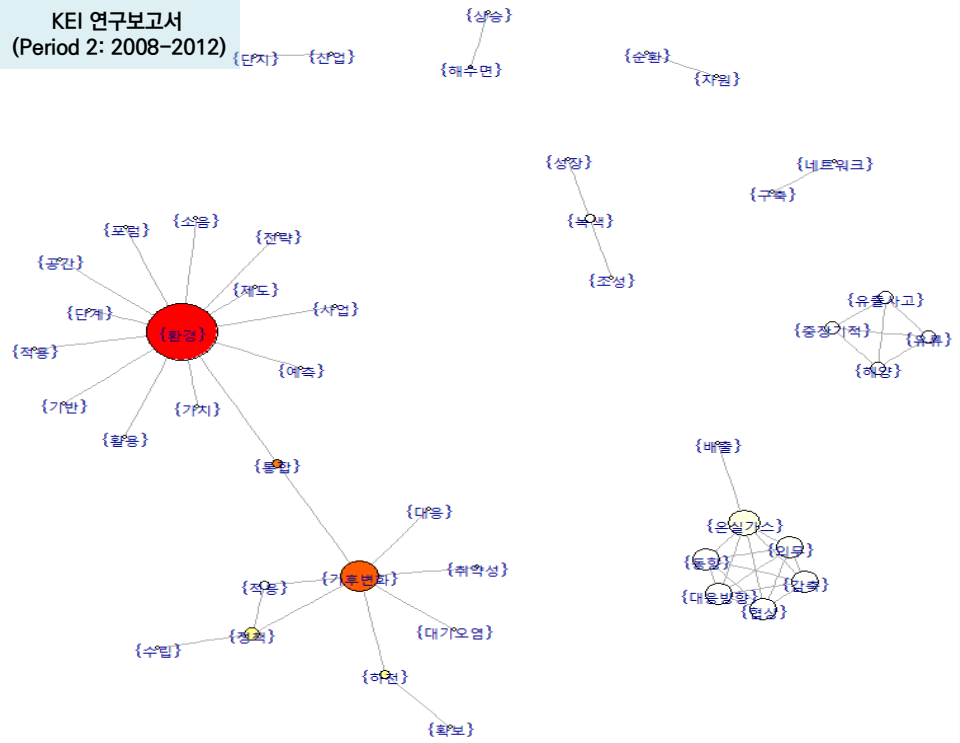
- 2006년 3월 새만금 사업은 긴 법정다툼 끝에 대법원 확정 판결이 나고 방조제 공사를 추진함.

C. **원유-유출, 타르-덩어리, 기름-찌꺼기, 물고기-떼죽음**

- 2007년 12월 태안 기름 유출 사고 발생함.

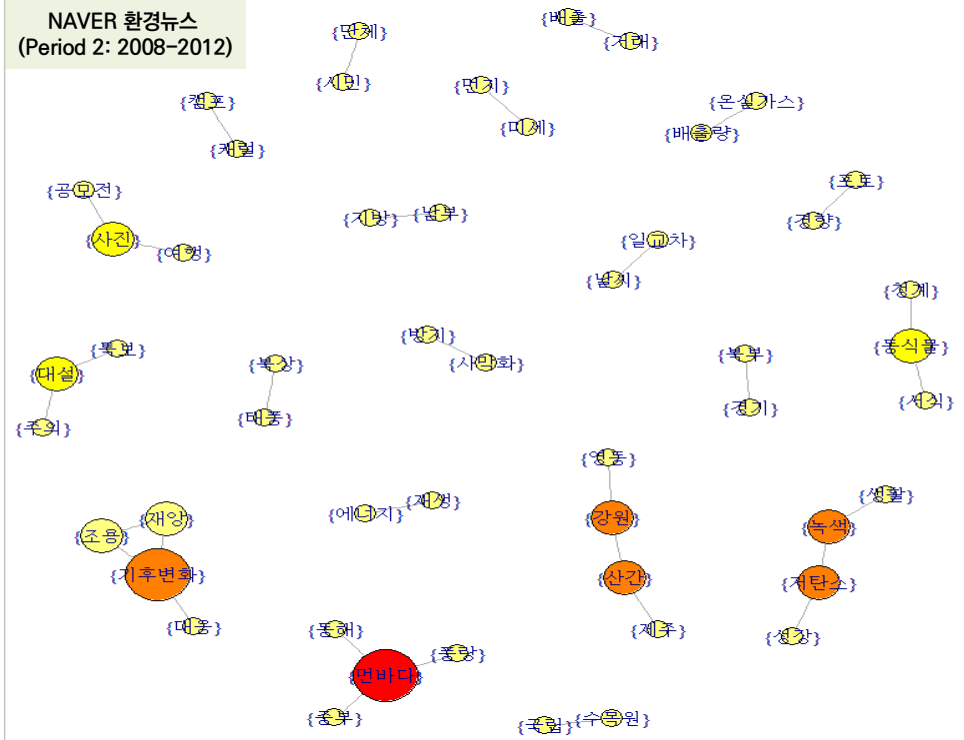
D. 남해안-적조, 경주지-방폐장-주민투표, 월성-원전, 황사-최악, 낙동강-하구 등의 키워드가 등장함.

KEI 연구보고서
(Period 2: 2008-2012)



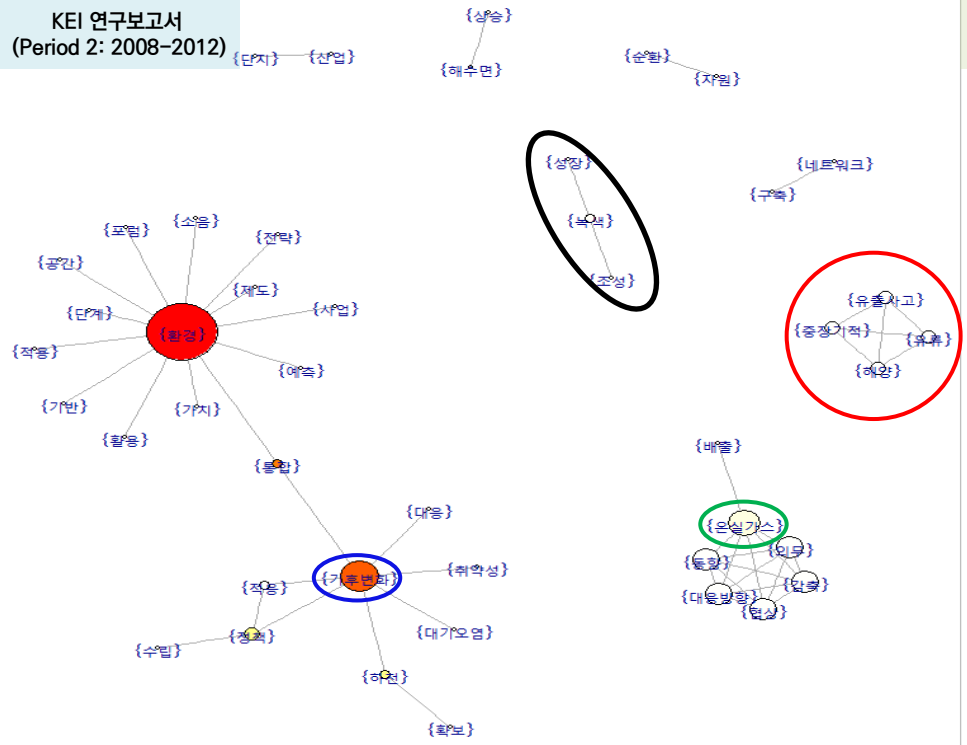
	A		B	신뢰도	지지도	향상도
1	중장기적	=>	유출사고	0.015	1.000	56.857
2	유출사고	=>	중장기적	0.015	0.857	56.857
3	중장기적	=>	유류	0.015	1.000	56.857
4	유류	=>	중장기적	0.015	0.857	56.857
5	대응방향	=>	동향	0.010	1.000	56.857
6	동향	=>	대응방향	0.010	0.571	56.857
7	대응방향	=>	감축	0.010	1.000	39.800
8	감축	=>	대응방향	0.010	0.400	39.800
9	대응방향	=>	온실가스	0.010	1.000	22.111
10	온실가스	=>	대응방향	0.010	0.222	22.111
11	협상	=>	온실가스	0.010	1.000	22.111
12	온실가스	=>	협상	0.010	0.222	22.111

NAVER 환경뉴스
(Period 2: 2008-2012)

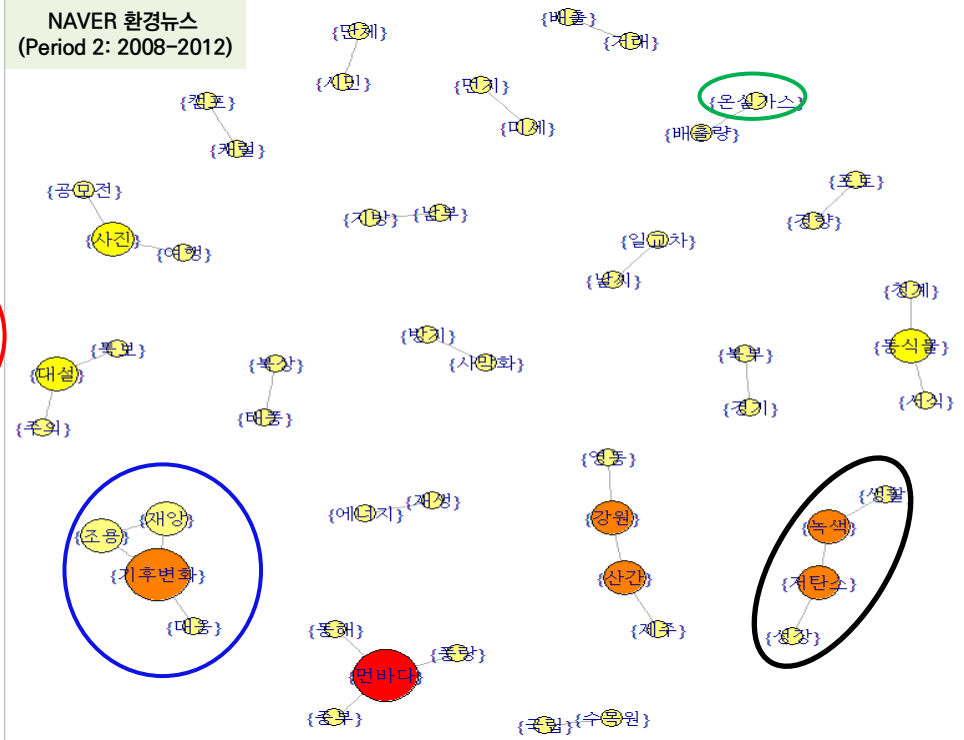


	A		B	신뢰도	지지도	향상도
1	대응	=>	기후변화	0.002	0.584	43.980
2	기후변화	=>	대응	0.002	0.143	43.980
3	대설	=>	주의	0.002	0.409	27.919
4	주의	=>	대설	0.002	0.126	27.919
5	재앙	=>	기후변화	0.002	0.566	42.564
6	기후변화	=>	재앙	0.002	0.121	42.564
7	저탄소	=>	녹색	0.002	0.612	39.608
8	녹색	=>	저탄소	0.002	0.100	39.608
9	북상	=>	태풍	0.001	0.673	62.289
10	태풍	=>	북상	0.001	0.138	62.289
11	대설	=>	특보	0.001	0.317	39.795
12	특보	=>	대설	0.001	0.179	39.795

KEI 연구보고서
(Period 2: 2008-2012)



NAVER 환경뉴스
(Period 2: 2008-2012)



1. 기후변화 2. 온실가스 3. 태안 기름 유출 사고 4. 녹색성장 5. 환경오염 대책

A. 전 구간 대비 **기후변화, 온실가스** 키워드의 매개중심성이 높아짐.

B. **해양 유류 유출사고, 녹생성장 조성**, 해수면 상승, 소음 키워드가 새롭게 등장함.

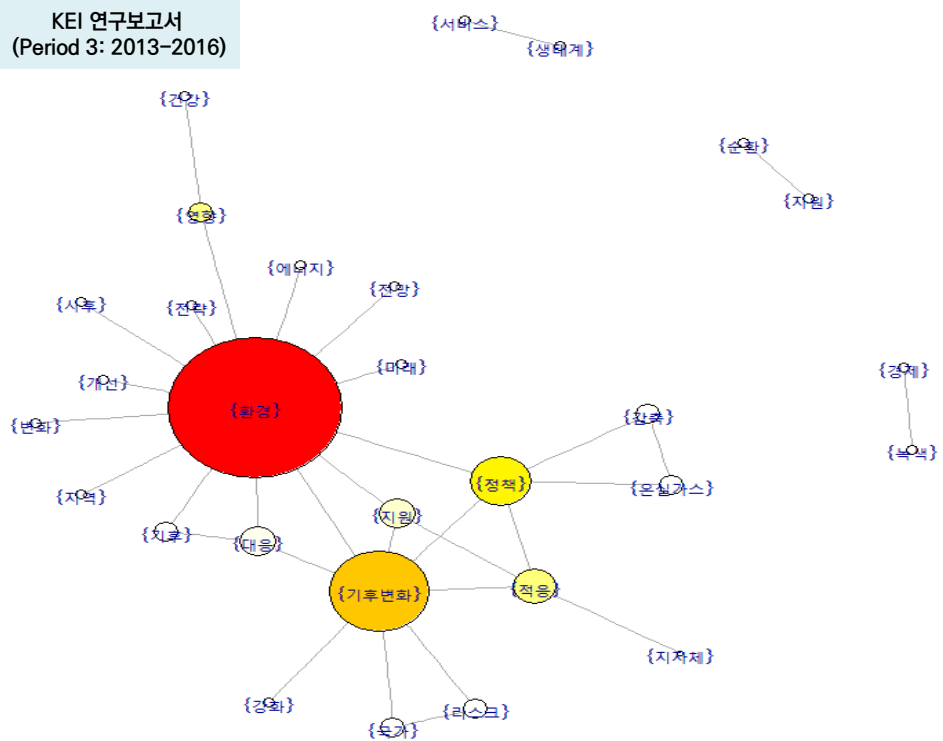
C. period1 NAVER 환경뉴스에서 등장한 원유-유출, 타르-덩어리, 기름-찌꺼기, 물고기-떼죽음 키워드가 의미하는 **'태안 기름 유출 사고'의 중장기적 영향분석 및 제도개선 방안 마련**과 관련한 연구가 진행되었음을 확인할 수 있음.

A. **기후변화, 온실가스, 녹색성장** 키워드가 새롭게 등장하고, KEI 연구보고서와 마찬가지로 중요하게 다뤄짐.

B. 미세먼지, 재생에너지, 대설-주의-특보, 사막화-방지, 태풍-북상, 날씨-일교차 등의 키워드가 새롭게 등장함.

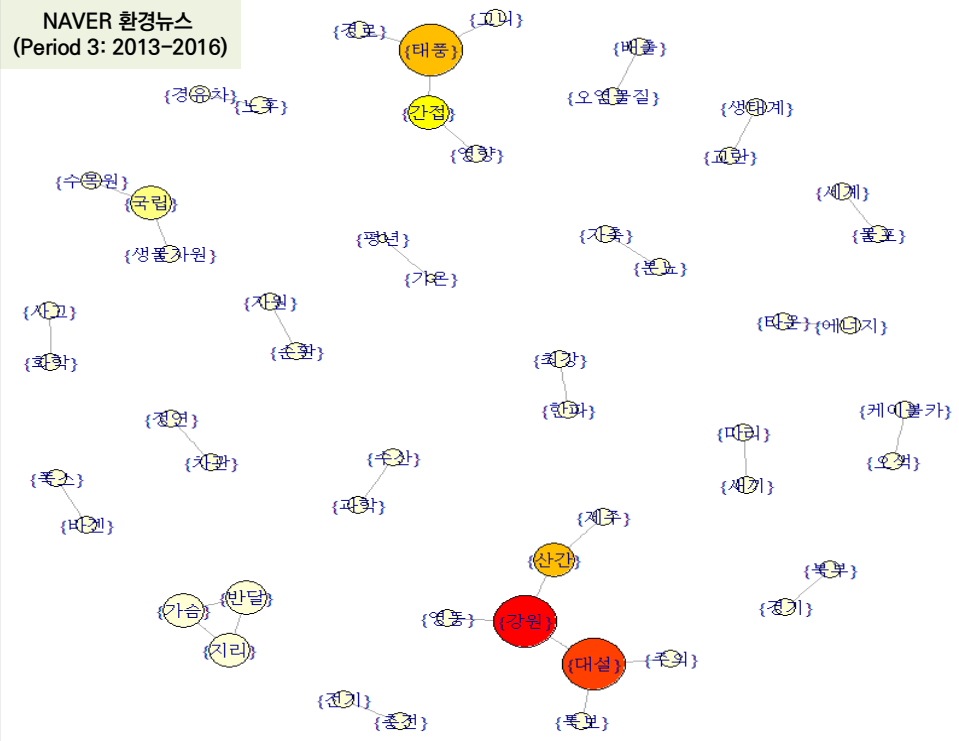
C. period1에 이어 **'동식물 보호'**에 대한 기사가 많았음을 동식물-청계-서식 키워드를 통해 알 수 있음.

KEI 연구보고서
(Period 3: 2013-2016)



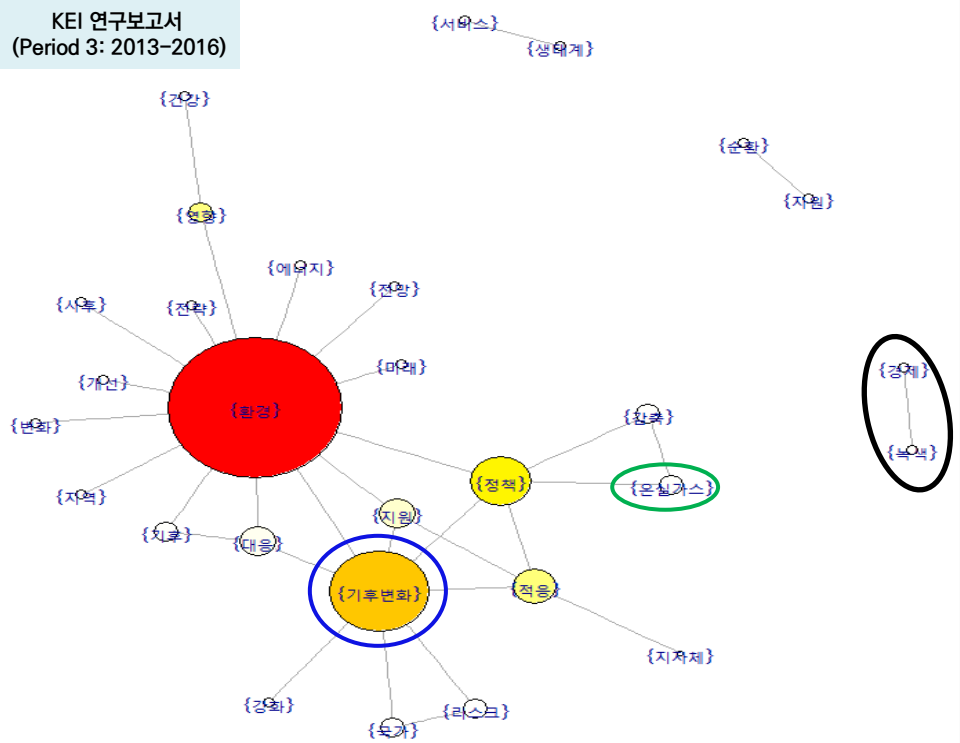
	A		B	신뢰도	지지도	향상도
1	정책	=>	기후변화	0.030	0.279	1.865
2	기후변화	=>	정책	0.030	0.200	1.865
3	대응	=>	기후변화	0.025	0.357	2.387
4	기후변화	=>	대응	0.025	0.167	2.387
5	전략	=>	환경	0.020	0.500	1.728
6	환경	=>	전략	0.020	0.069	1.728
7	기후변화	=>	환경	0.015	0.100	0.346
8	환경	=>	기후변화	0.015	0.052	0.346
9	리스크	=>	기후변화	0.013	0.500	3.342
10	기후변화	=>	리스크	0.013	0.083	3.342
11	미래	=>	환경	0.013	0.455	1.571
12	환경	=>	미래	0.013	0.043	1.571

NAVER 환경뉴스
(Period 3: 2013-2016)



	A		B	신뢰도	지지도	향상도
1	간접	=>	태풍	0.002	0.960	51.279
2	태풍	=>	간접	0.002	0.081	51.279
3	대설	=>	특보	0.001	0.380	40.323
4	특보	=>	대설	0.001	0.128	40.323
5	최강	=>	한파	0.001	0.688	62.212
6	한파	=>	최강	0.001	0.105	62.212
7	고니	=>	태풍	0.001	0.675	36.055
8	태풍	=>	고니	0.001	0.061	36.055
9	반달	=>	지리	0.001	0.685	153.626
10	지리	=>	반달	0.001	0.237	153.626
11	가슴	=>	지리	0.001	0.667	149.529
12	지리	=>	가슴	0.001	0.237	149.529

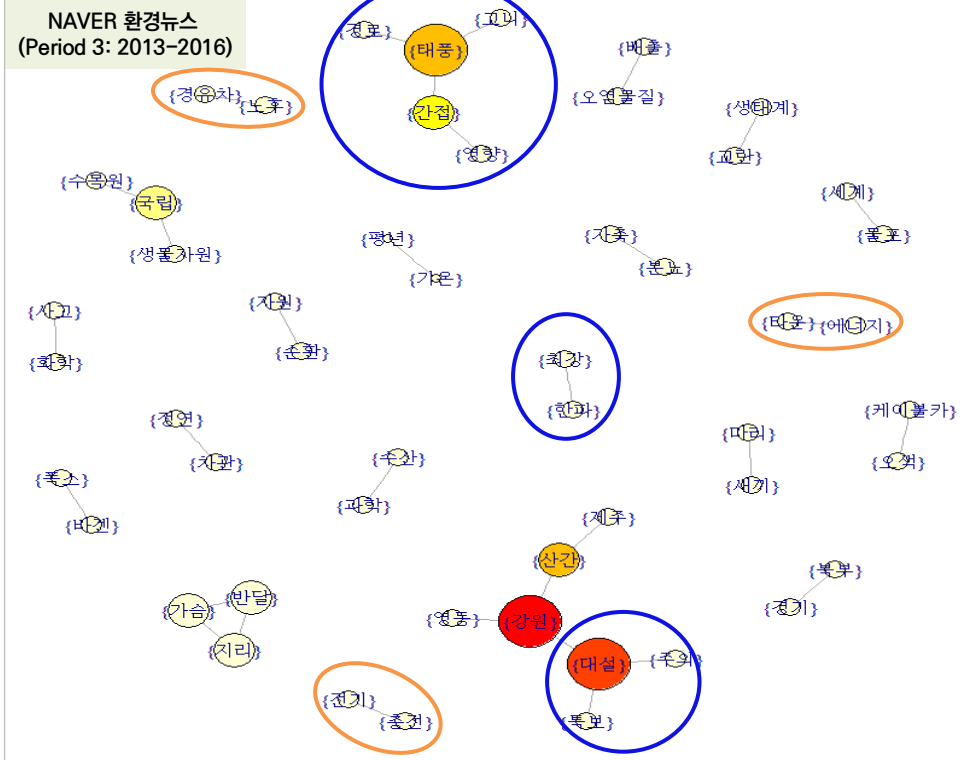
KEI 연구보고서
(Period 3: 2013-2016)



1. 기후변화 2. 온실가스 3. 태안 기름 유출 사고 4. 녹색성장 5. 환경오염 대책

A. period2에 이어 **기후변화** 키워드의 매개중심성이 높아짐
B. 환경 키워드를 중심으로 **건강, 미래, 전망, 에너지** 키워드가 새롭게 등장함.
C. period2에 '녹색' 키워드의 연관어였던 '성장' 키워드는 사라지고, '**경제**'가 연관어로 나타남.
D. period2에 분산되어 있던 키워드 네트워크들이 **환경, 기후변화, 정책** 키워드를 중심으로 합쳐짐.

NAVER 환경뉴스
(Period 3: 2013-2016)



A. 기후변화 키워드는 사라지고 **태풍, 대설** 키워드의 매개중심성이 높아짐.
B. 환경오염 대책을 위한 키워드가 등장함.
   - **노후 경유차**: 노후 경유차 폐차 후 두 달 안에 새차 구입 시 개별소비세를 줄여주는 제도
   - **전기 충전**: 환경 친화적 전기자동차의 개발 및 보급촉진을 위한 다양한 지원제도
   - **에너지 타운**: 환경부는 친환경 에너지타운을 전국으로 확산을 위해 노력
C. period2에 이어 **'동식물 보호'**에 대한 기사가 많았음을 국립-수목원-생물자원, 지리-반달-가슴 키워드를 통해 알 수 있음.

Trend Comparative Analysis

1. 기후변화

KEI 연구보고서에서 '기후변화' 키워드는 **시간이 지날수록 매개중심성**(Betweenness Centrality, Cb)**이 높아**지고, 언급량이 많아지는 주요 키워드로 판단된다. 초기에는 기후변화 **영향평가 및 적응시스템 구축**관련 연구가 진행되었고, 이후 **기후변화 대응을 위한** 연구들(하천공간 확보방안 연구, 기후변화 적응정책 수립 연구)이 진행되었다. 또한, 기후변화와 **대기오염을 함께** 연구하는 경향을 보였다. 최근에는 이러한 기후변화 적응정책을 **지원**하기 위한 연구가 진행됐으며, 기후변화에 따른 국가 리스크 연구도 진행되었다.

NAVER 환경뉴스에서 '기후변화' 키워드는 Period 2(2008-2012)에 '재앙' 키워드와 함께 많이 등장하기 시작하였다. 이후 기후변화 키워드보다는 '태풍', '최강 한파', '대설 주의' 등의 기후변화 관련 키워드들이 많이 언급되었다. 이를 통해 '기후변화'의 이슈는 앞으로 계속 중요하게 다뤄질 문제이기 때문에 향후 KEI는 **기후변화의 세부적인 현상들(온난화, 홍수, 가뭄 등)에 대한 연구가 개별적으로 진행**되야 한다고 판단된다.

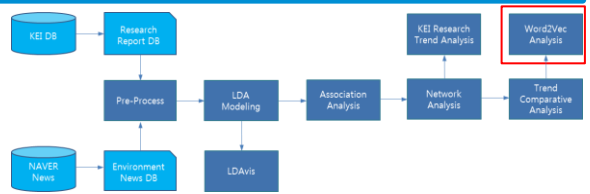
따라서 본 연구에서는 가장 중요하다고 판단되는 '기후변화' 키워드를 word2vec(skip-Gram)분석을 통해 구체적으로 살펴보고자 한다.

Trend Comparative Analysis

2. 온실가스
3. 태안 기름 유출 사고
4. 녹색성장
5. 환경오염 대책

-> 향후 결과 작성 예정

Text Mining Process List

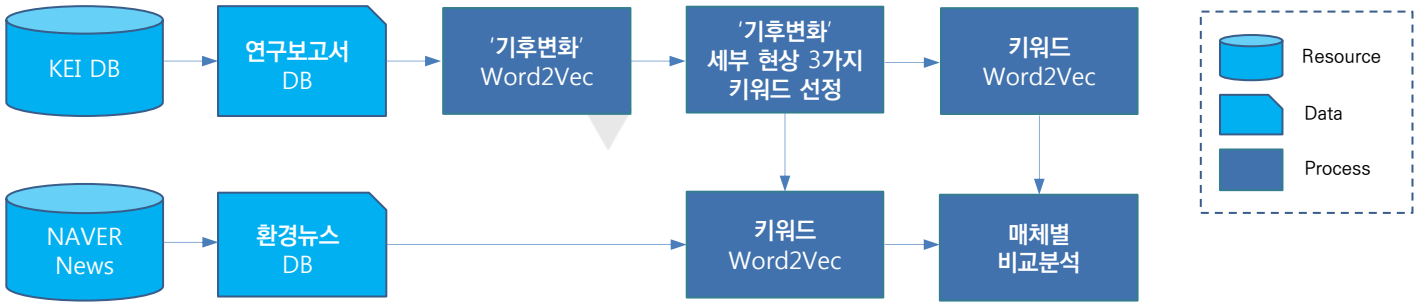


Plan 2017	Process	Code	Description	Input	Output	Note
8월 상순	**Word2Vec Analysis**	Word2Vec.R	- Skip-Gram Model 사용함 - 거리측정법: 코사인거리 - 기후변화 세부 현상들 3가지 (온난화, 홍수, 가뭄) 키워드 연관어 분석	out_kei.csv out_naver.csv	kei_w2v.txt kei_w2v_2.bin naver_w2v.txt naver_w2v_2.bin	- Window size: 10 - Worker threads: 3 - Word vector dimensionality: 100

기후변화 세부 현상 키워드 Word2Vec(Skip-gram) 과정

- '기후변화' 세부 현상 키워드 Word2Vec(Skip-gram) 과정

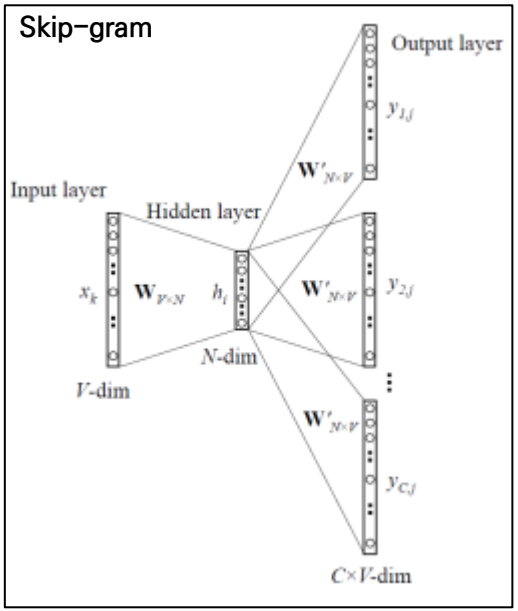
1. KEI 연구보고서(1993~2016) 전체 데이터를 바탕으로 **'기후변화' 키워드에 대한 Word2Vec(Skip-gram)** 분석을 실시
2. 분석결과 홍수위, 호우, 홍수, 태풍, 한파, 가뭄, 폭염, 수온, 온난화 등 기후변화 세부현상 키워드가 많이 출현
3. 분석결과와 전문가 의견을 바탕으로 **기후변화 세부 현상들 3가지(온난화, 홍수, 가뭄)** 선정
4. 3가지 기후변화 세부 현상 별 Word2Vec(Skip-gram) 분석을 매체별 실시하고 **비교 분석**



기후변화 관련 키워드 Word2Vec(Skip-gram)

❖ Word2Vec model 생성

1. Tool : R
   Package: devtools, wordVectors, tsne
2. Model : Skip-Gram
   -〉 모델종류 : CBOW(Continuous Bag-Of-Words), Skip-gram
3. Window size : 10
   -〉 키워드 양옆에 존재하는 총 10개 단어 고려
4. Worker threads : 3
   -〉 처리 속도 설정
5. Word vector dimensionality : 100
   -〉 100개의 특성 사용



기후변화 키워드 Word2Vec(Skip-gram)

- KEI 연구보고서(1993~2016) 전체 데이터를 바탕으로 **'기후변화'** 키워드에 대한 Word2Vec(Skip-gram) 분석을 실시
  - '기후변화'와 근접한 순으로 100개 단어 출력 (거리측정법: 코사인거리)
  - 분석결과 홍수위, 호우, 홍수, 태풍, 한파, 가뭄, 폭염, 수온, 온난화 등 기후변화 세부현상 키워드가 많이 출현

취약	적응	외력	리스크	현상	회의록	해수면	기상이변	기상청	하절기
0.2692499	0.2906538	0.3955267	0.4021677	0.4049839	0.4141995	0.4274032	0.4298225	0.434107	0.4385216
메	스케일	부탄	극한	영향력	사상	홍수위	호우	계층	전역
0.4542906	0.4544078	0.4561009	0.4565966	0.4577734	0.4596295	0.461989	0.4680592	0.4686221	0.4688754
양상	삼림	인구학	의심	피니언	만성	강수량	온도	응법	포지션
0.4701063	0.472633	0.4807527	0.4846653	0.4855634	0.488519	0.4886752	0.4904773	0.4935732	0.4944174
미래	강수	간과	극하다	근로	온난	직면	홍수	태풍	대응
0.499352	0.500258	0.5007341	0.5007484	0.5021937	0.502265	0.5028752	0.5103362	0.5110601	0.5130805
영향	자외선	한파	가속	긴급	침입	극단	열섬	응대	다이내믹스
0.5138669	0.514845	0.5157218	0.5177694	0.5181254	0.5196883	0.5198814	0.5247445	0.5259444	0.5271059
가뭄	함	폭염	에티오피아	상이하다	심포지엄	개인	수온	추다	돕다
0.5276781	0.528426	0.5288129	0.529674	0.5300491	0.5310185	0.5314472	0.5314991	0.5319397	0.536184
저소득	내수	마산만	체수	유엔	우선순위	쓰다	방어	능력	시공간
0.536995	0.5391981	0.5403516	0.540638	0.5413343	0.541469	0.5438729	0.5460488	0.5460887	0.5467132
조위	8개	건강	주류	틀다	떠오르다	물수지	염분	살피다	식량
0.5467424	0.5476009	0.5492733	0.5505579	0.5524308	0.5570488	0.5577965	0.5579844	0.559733	0.5605145
시나리오	남한	빈도	강우	신경	증발산량	극대	악영향	지상	리더
0.5605915	0.5629628	0.5642913	0.5653338	0.5677786	0.5681688	0.5682759	0.5684116	0.5692906	0.5701775
기상	민감	연구소	습도	심혈	온난화	침	주거지	역량	상승
0.570817	0.5709726	0.5714237	0.5716294	0.5717762	0.573174	0.5732041	0.5738196	0.574247	0.5749484

기후변화 키워드 Word2Vec(Skip-gram)

- 분석결과와 전문가 의견을 바탕으로 **기후변화 세부 현상들 3가지(온난화, 홍수, 가뭄)** 선정
  - 네이버 뉴스에서 '태풍'은 연관 키워드로 다양한 태풍 이름(다나스, 곤파스, 도라지 등)이 많이 출현해 형태소 분석에 어려움이 있으므로 배제함
  - 네이버 뉴스에서 '한파', '폭설', '폭염', '폭우' 키워드들은 대부분 기상뉴스에 출현해 연관 키워드로 불용어(35도, 곳곳, 오후 등)가 많이 출현해 분석에 어려움이 있으므로 배제함

- 3가지 기후변화 세부 현상 별 Word2Vec(Skip-gram) 분석을 매체별 실시하고 **비교 분석**
  - '기후변화' 세부 현상들 3가지 키워드 별 근접한 순으로 50개 단어 출력 (거리측정법: 코사인거리)

1. '온난화' 연관 키워드

KEI 연구보고서

극단	악영향	직면	가속	하절기
0.261	0.348	0.354	0.363	0.373
온난	인하다	최근	호우	국지
0.373	0.379	0.384	0.387	0.389
집중호우	현상	여름철	심화	일평균
0.391	0.396	0.402	0.402	0.405
강수량	양상	생존	태풍	피하다
0.413	0.417	0.419	0.419	0.422
지구	일으키다	미만	거주	살피다
0.424	0.433	0.434	0.435	0.437
증발산량	연평균	한파	강수	극한
0.438	0.441	0.442	0.443	0.444
성패	줄이다	빈도	관찰	막대
0.444	0.449	0.451	0.451	0.451
인류	4년	작용	일어나다	인명
0.452	0.459	0.461	0.461	0.463
길다	밝히다	호흡기계	급속	강도
0.463	0.468	0.469	0.471	0.476
이상	심혈	2050년	처하다	상승
0.476	0.477	0.48	0.48	0.48

A. **인류, 인명, 호흡기계, 심혈 등 인간 중심 키워드가 등장함**
   -) 지구 온난화로 인한 피해 중 인간에 미치는 영향에 대한 연구가 많음

B. **직면, 최근, 4년, 처하다 등 단기적 키워드가 등장함**
   -) KEI는 주로 단기적 시점의 지구 온난화를 연구하는 것으로 보임

NAVER 환경뉴스

온난	2080년	표면	지중해	2100년
0.158	0.252	0.277	0.291	0.294
툰드라	해빙	아열대	2040년	난대림
0.297	0.298	0.304	0.305	0.321
세기말	기후대	지표면	급변	팽창
0.322	0.324	0.328	0.333	0.334
방귀	북반구	생장	플랑크톤	엽
0.335	0.341	0.342	0.342	0.343
해수면	금세기	심상찮다	지구	상승
0.343	0.35	0.351	0.352	0.352
산호초	그린란드	임팩트	영구동토	빙하
0.353	0.354	0.36	0.364	0.366
마그마	말매미	빠르다	이변	경고
0.37	0.375	0.376	0.379	0.38
급상승	빙상	북극	2000년	난민
0.385	0.39	0.391	0.391	0.391
장강	칭장	2050년	징후	온기
0.392	0.393	0.394	0.396	0.397
대박	트림	티베트	식량	난류성
0.402	0.404	0.404	0.404	0.405

A. **플랑크톤, 산호초, 말매미 등 생물 키워드와 식량 키워드가 등장함**
   -) 지구 온난화로 인한 피해 중 생물과 식량에 미치는 영향에 대한 뉴스가 많음

B. **2080년, 2100년, 2040년, 2050년 등 중장기적 키워드가 등장함**
   -) 뉴스는 지구 온난화로 인한 미래의 모습에 관심이 많음

2. '홍수' 연관 키워드

KEI 연구보고서

치수	극한	방어	홍수위	빈도
0.286	0.291	0.316	0.329	0.332
극하다	집중호우	사상	태풍	강도
0.332	0.36	0.361	0.361	0.364
호우	제방	침수	자연재해	가뭄
0.381	0.388	0.397	0.4	0.4
재해	범람	열섬	홍수피해	강우
0.421	0.422	0.426	0.427	0.432
침	강수	건천	외력	미기후
0.432	0.433	0.436	0.441	0.451
긴급	처하다	극대	국지	노후
0.457	0.461	0.462	0.464	0.466
마산만	온도	리스크	해수면	인위
0.47	0.472	0.477	0.477	0.479
강수량	피해	횡	인명	기후
0.483	0.488	0.49	0.49	0.491
조위	막대	가뭄지수	위협	폭염
0.495	0.496	0.497	0.5	0.501
넘다	사빈	체수	저류지	대처
0.501	0.502	0.504	0.508	0.509

A. **건천, 마산만 등 대한민국 지역 키워드가 등장함**
   -) KEI는 주로 국내 홍수 피해를 연구하는 것으로 보임

B. **인명, 노후 등 인간 중심 키워드가 등장함**
   -) 홍수로 인한 인적 피해를 대처하기 위한 연구가 많음

NAVER 환경뉴스

가물막이	사방댐	후난	보	제방
0.281	0.314	0.331	0.339	0.374
홍수조절	무너지다	탁수	극하다	천보
0.382	0.387	0.391	0.394	0.395
싼샤댐	소방방	유속	홍수피해	대홍수
0.397	0.399	0.404	0.408	0.41
붕괴	댐	광동댐	임하	흙탕물
0.411	0.413	0.416	0.419	0.421
황하	조절	저수량	미보	두만강
0.421	0.422	0.424	0.425	0.425
다목	침하	물그릇	산사태	이변
0.428	0.429	0.429	0.43	0.431
대강	치수	군남	범람	팔당댐
0.438	0.438	0.44	0.441	0.446
가뭄	비만	임진강	저수조	결여
0.447	0.451	0.451	0.451	0.452
항공업	쓰촨	기근	부작용	재해
0.453	0.454	0.455	0.455	0.455
침수	합	필승	준설	급하다
0.456	0.458	0.458	0.459	0.461

A. **후난, 황하, 쓰촨 등 중국 지역 키워드가 등장함**
   -) 뉴스는 홍수 피해가 큰 국외 사건에 관심이 많음

B. **사방댐, 싼샤댐, 광동댐, 팔당댐 등 댐 키워드와 보, 천보, 미보 등 보 키워드가 등장함**
   -) 홍수 대비를 위한 구조물 키워드가 많이 나타남

3. '가뭄' 연관 키워드

KEI 연구보고서

가뭄지수	홍수	서북	체수	집중호우
0.304	0.4	0.462	0.468	0.468
강수	긴급	강수량	호우	자연재해
0.47	0.471	0.489	0.494	0.504
사상	빈도	극한	태풍	강도
0.507	0.508	0.513	0.513	0.519
급수	겪다	커지다	홍수위	재해
0.527	0.534	0.536	0.537	0.537
위기관리	극대	극하다	대응	리스크
0.539	0.543	0.544	0.544	0.544
변화	생활용수	수량	넘다	중단
0.545	0.546	0.548	0.561	0.561
재난	기후	증발산량	다각	국지
0.563	0.563	0.565	0.567	0.569
하절기	수자원	강우	침	대처
0.572	0.573	0.576	0.576	0.576
세우다	현상	치수	공급	직면
0.577	0.577	0.579	0.579	0.582
폭염	동일	인명	제방	대책
0.582	0.583	0.584	0.59	0.593

A. **생활용수, 인명 등 인간 중심 키워드가 등장함**
   -) 가뭄으로 인한 인적 피해를 대처하기 위한 연구가 많음

B. **긴급, 극한, 극대, 극하다 등 부정적 키워드가 등장함**
   -) KEI는 가뭄현상에 대한 심각성을 강조하는 것으로 보임

NAVER 환경뉴스

역부족	식수난	밭작물	끌어오다	목마르다
0.23	0.253	0.273	0.284	0.289
극심한	저수율	타들다	해갈	일조량
0.291	0.308	0.318	0.334	0.357
부족	생육	저수량	기우제	타들어가다
0.359	0.37	0.372	0.375	0.382
물그릇	급수	트라우마	단비	장강
0.382	0.383	0.385	0.391	0.412
해소	안희정	지독	용수	농업용수
0.417	0.42	0.422	0.422	0.424
이번	마른장마	최악	한시름	강수량
0.427	0.43	0.432	0.436	0.443
전전긍긍	황하	홍수	폭등	소양강댐
0.443	0.445	0.452	0.456	0.458
4분	엘니뇨	광동	마르다	다목
0.459	0.461	0.462	0.47	0.471
모자라다	벼농사	기근	대지	강우량
0.473	0.477	0.479	0.481	0.483
도움	갈아엎다	흙탕물	극복	말라가
0.485	0.486	0.487	0.487	0.489

A. **밭작물, 기우제, 농업용수, 벼농사 등 농업 관련 키워드가 등장함**
   -) 가뭄으로 인한 피해 중 농업에 미치는 영향에 대한 뉴스가 많음

B. **황하, 광동 등 중국 지역 키워드가 등장함**
   -) 뉴스는 가뭄 피해가 큰 국외 사건에 관심이 많음

	KEI 연구보고서	NAVER 환경뉴스
피해 대상	인간	생물, 식량, 농업 등
피해 지역	국내	국외
피해 기간	단기	중장기

Text Mining Process List

Plan 2017	Process	Code	Description	Input	Output	Note
9월	**결론 및 시사점 도출**					
10월	**향후 계획 수립**					

학술적 기대효과

- **장기간에 걸친 KEI 연구 동향을 정리하여 추후 환경연구 기획에 필요한 정보를 원내외 연구진에게 제공**

- **환경분야 텍스트 마이닝 분석기반 플랫폼 개발의 기초 구성**
  - 환경관련 연관어 분석, 네트워크 분석, 토픽 클러스터링, Word2Vec 등 다양한 텍스트 마이닝 분석 기법 집적 가능
  - 추후 이들 기법을 자동으로 처리하는 플랫폼을 구축하는 기초로 활용 가능

후속 연구

- **매체별 환경문제 인식 성향 분석을 소셜미디어, 전통미디어, 전문사이트(학술논문), 공공기관 발간문건 등으로 확대하여 연구동향과 사회적 인식간의 관계파악 범위를 확대**

- **KEI 제공 발간물 데이터 시각화 서비스를 구축하여 사용자의 이용 편이 증진**
  - 대량의 KEI 발간물 데이터에 대한 정보를 사용자가 효율적으로 파악할 수 있도록 정보 전달력을 제고

- **환경연구 트렌드 분석을 활용하여 미래 환경연구 수요 예측에 반영**
  - 기존의 정량적 전망을 활용한 미래 환경문제 예측을 반영하는 연구수요 예측과 매체 분석을 통해 수요자 선호를 반영하는 연구수요 예측을 병행

정책 개발

- **환경정책 수요자의 선호를 정책개발에 활용하여 “환경서비스 품질수준 제고[1)]” 도모 가능**
  - 매체별 환경분야 연구동향과 사회적 요구를 비교분석한 결과를 근거로 수요자의 선호를 파악하여 정책 개발 기초 자료로 활용

1) 국정과제 95. 생활환경 취약지역 개선 및 환경질 개선의 과제개요

Thank you.