텍스트 마이닝을 이용한 KEI 연구동향 분석

6th Bigdata Research Team Seminar : Progress Report

2017. 05. 24

빅데이터연구팀 김도연

목차

연구 개요

- **과제명** : 텍스트 마이닝을 이용한 KEI 연구동향 분석

- **참여 연구진** :

성명	소속	주요업무	참여율(%)
강성원	빅데이터연구팀	텍스트 마이닝 분석 플랫폼 개발 및 분석	60
김도연	빅데이터연구팀	문헌 분석, 데이터 수집 및 분석	40

- **연구 기간** : 2017.1.1 ~ 2017.12.31

자문의견 반영 내용

위원	제안내용	반영내용
이명진 (KEI 부연구위원)	한글 전처리 문제 (조사 등)	• 한글 전처리 R package를 활용해서 해결 가능함. – 형태소분석기 실행(KoNLP 등) – Low TF-IDF 값 제거 – 불용어 처리 (특정 단어 삭제, 특수문자 제거, 소문자로 변경 등) – Word Lengths는 2글자 이상 – 동의어 처리
	한글 라이브러리 문제	• 토픽클러스터링(LDA)기법을 활용한 토픽별 사전을 구축할 예정임. – 토픽별 전문가 의견을 반영하여 사전 구축
우석진 (명지대학교 교수)	환경정책 수혜자 DB 분석을 통해 정책적 시사점 도출 필요	• 온라인 뉴스기사와 뉴스댓글 데이터를 통해 환경정책 수혜자 오피니언 마이닝 분석 실시할 예정임. – 향후, 환경부에서 제공하는 페이스북, 트위터, 블로그, 유튜브 등을 통해 정책수혜자 오피니언 마이닝 분석 실시할 계획임.

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 뉴스기사와 뉴스댓글 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

구 분		연구목적	연구방법	주요 연구내용
주요 선행 연구	1	- 과제명: 텍스트 마이닝 기법을 활용한 한국의 경제연구 동향 분석 - 연구자(년도): 송민 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 소셜빅데이터 이용 분석	- 키워드 분석 - 연관성 분석	-소셜 미디어 통일관련 연관어 분석 -통일관련 연관어와 통일인식간의 관계 분석
	3	- 과제명: 항공산업 미래유망분야 선정을 위한 텍스트 마이닝 기반의 트렌드 분석 - 연구자(년도): 신경식 외(2015) - 연구목적: 텍스트마이닝 트렌드 분석활용 항공산업 미래유망분야 발굴	- 토픽 모델링 분석	- 텍스트 마이닝 기법을 적용한 항공산업 관련 논문 트렌드 분석 - 토픽 모델링 분석 활용 항공산업 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 소셜 빅데이터 분석 - 전체문서 및 환경문서의 행복도 비교 분석
본 연구		- 기존의 연구는 단일 매체의 추세 파악에 집중하여 분석결과 활용이 미진 - 환경분야 문헌 텍스트 마이닝 연구는 초기단계	- 분석 대상이 단일 매체에 국한되어 전체적인 동향 파악에 한계가 있음	- 환경분야 연구문헌 텍스트 마이닝 기법을 적용하는 선도적 연구가 필요 - 다양한 매체의 동향을 비교 분석하여 연구 동향의 시사점을 파악하는 연구가 필요

주요 연구 내용

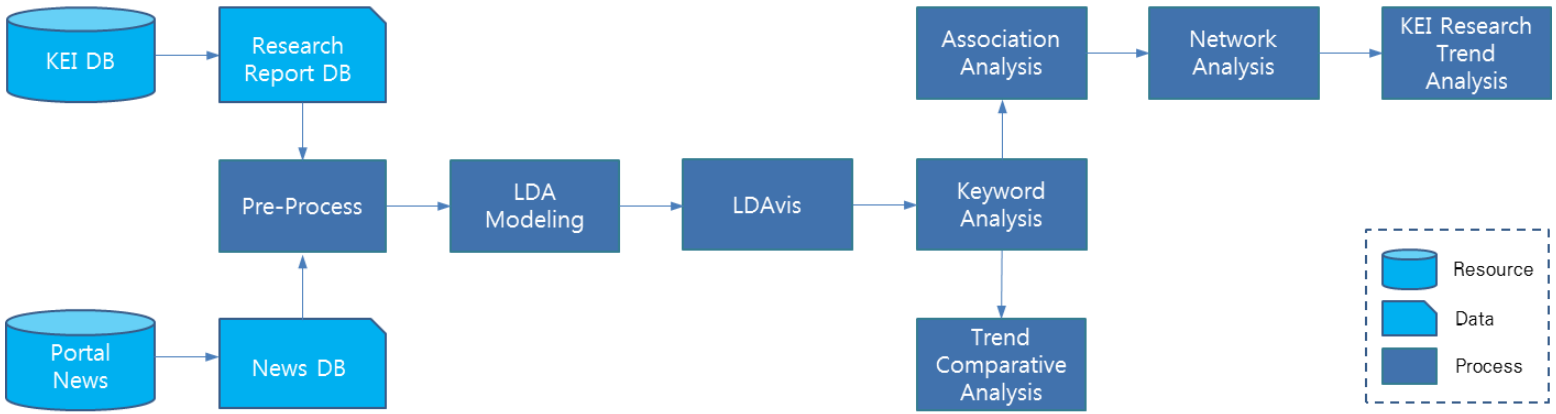
KEI 연구동향 파악

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

매체별 환경분야 이슈 비교 분석

- 연구공급 동향과 연구수요 동향 비교 분석
  - 연구공급 동향 파악: 2007년부터 2016년까지의 KEI DB에서 제공하는 연구보고서 데이터 분석
  - 연구수요 동향 파악: 2012년부터 2016년까지의 언론매체에서 제공하는 뉴스 기사 및 댓글 데이터 분석
- 매체별 추출한 키워드를 시계열로 파악하고 두 시계열을 비교 분석

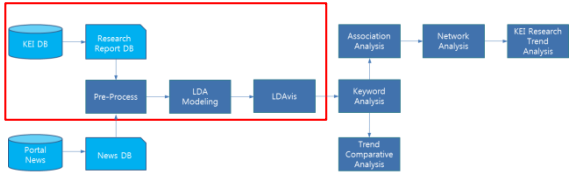
Text Mining Flowchart



Text Mining Process List

	Plan 2017	Process	File Name	Description	Input	Output	Note
상반기	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.csv	DocumentTermMatrix	- 자문의견(이명진 박사님) : 한글 처리 문제(조사 등) -) 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_1012.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_news.java	- Web crawling		Naver_news.csv	- 2012~2016년 네이버에서 제공하는 환경뉴스 데이터 수집 - Java-joup 사용
하반기	6월 하순	Pre-processing(2)		상동			
	7월 상순	LDA Modeling(2)					
	7월 상순	LDAvis(2)					
	7월 하순	Keyword Analysis(2)					
	8월	Trend Comparative Analysis					
	9월	시사점 도출 및 정책제언					
	10월	향후 계획 수립					

Text Mining Process List



Plan 2017	Process	File Name	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.csv	DocumentTermMatrix	- 자문의견(이명진 박사님) : 한글 처리 문제(조사 등) -) 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

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,]
```

+

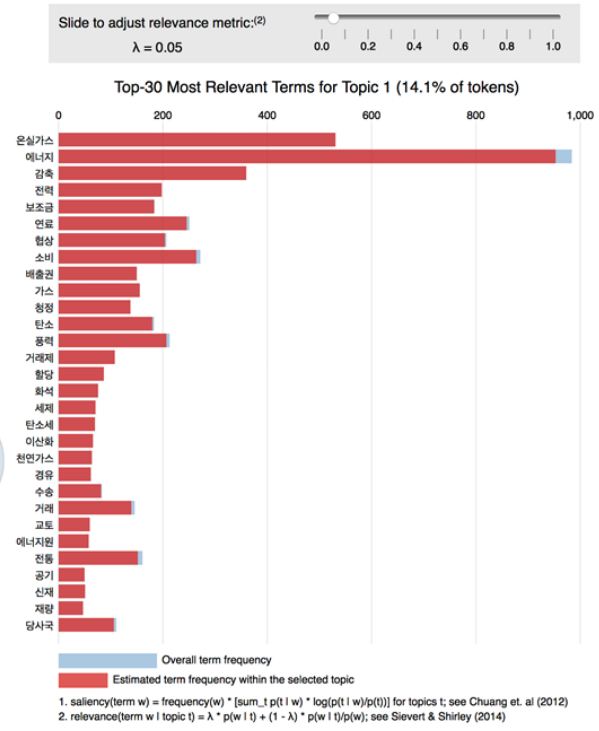
연구자의 판단 -> 불용어 처리!

2. Pre-processing data: id_topic.csv 생성

A	B	C	D	E	F	G	H	I	J	K
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 사호	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 중남 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

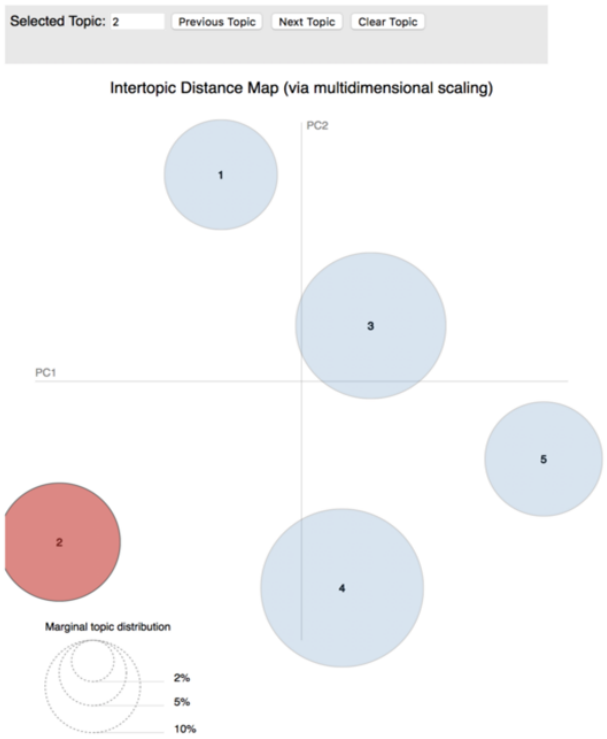
LDAvis (Topic 1)

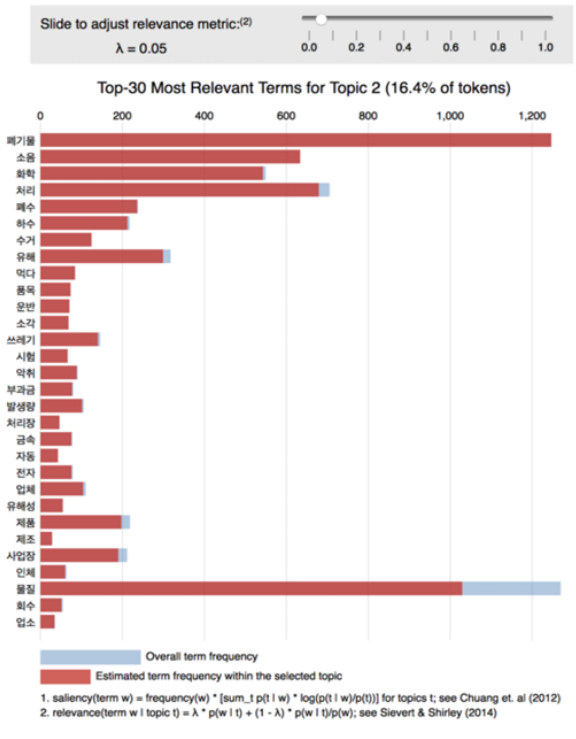


Term
온실가스
에너지
전력
연료
가스
청정
탄소
풍력
세제
탄소세
이산화탄소
천연가스
경유
공기
신재생에너지

“에너지 자원”

LDAvis (Topic 2)

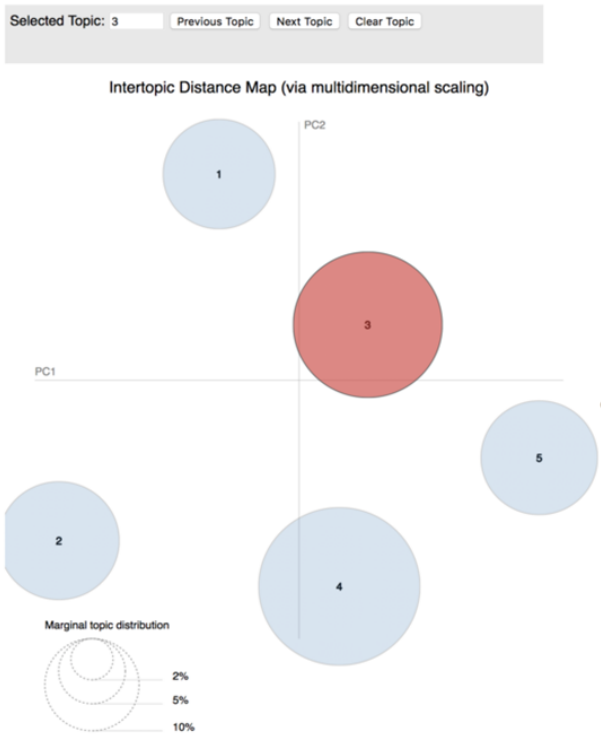


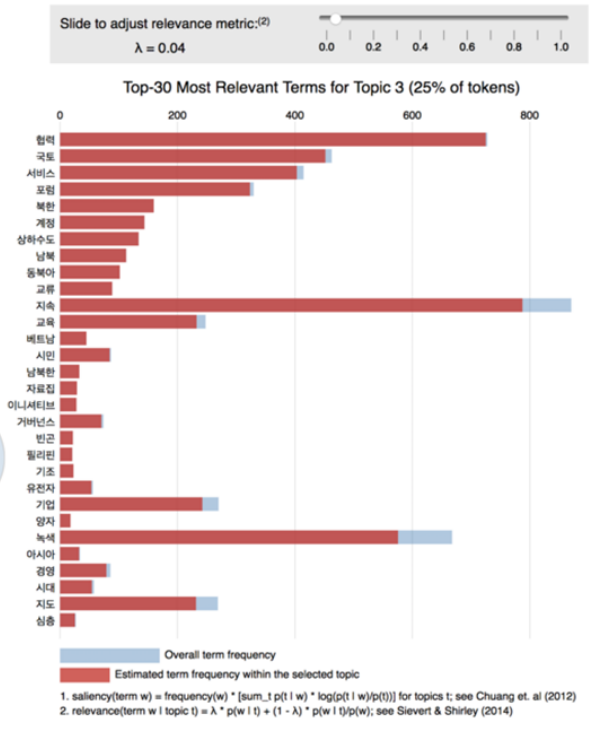


Term
폐기물
소음
화학
처리
폐수
하수
수거
유해
막다
소각
쓰레기
악취
부담금
처리장
유해성

“폐기물”

LDAvis (Topic 3)

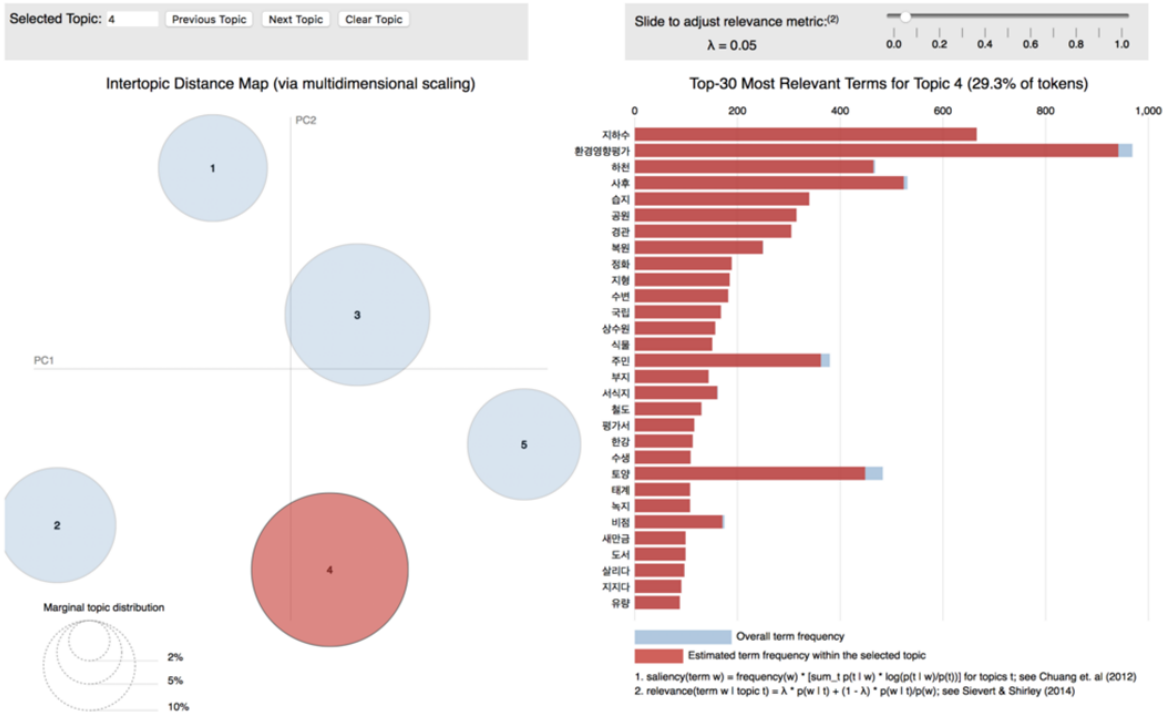




Term
협력
포럼
북한
상하수도
남북
동북아
교류
지속
베트남
시민
남북한
이니셔티브
거버넌스
필리핀
아시아

“대외 협력”

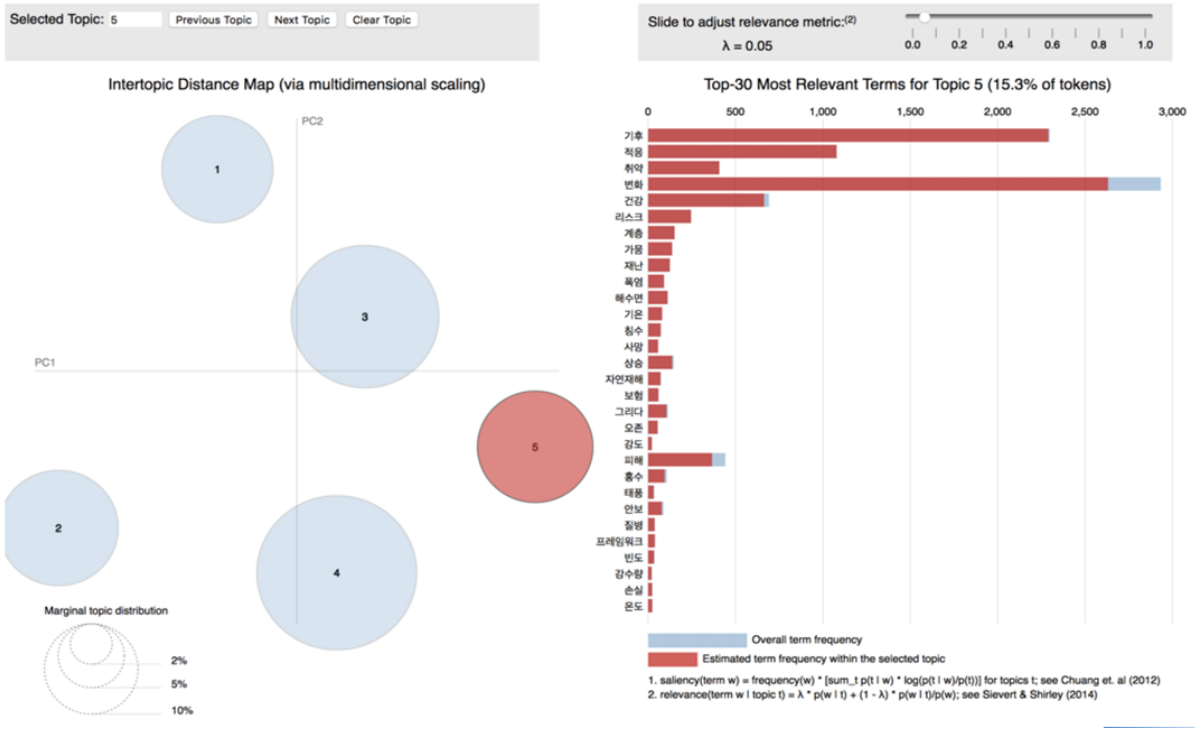
LDAvis (Topic 4)



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

“물환경”
+
“환경영향평가”

LDAvis (Topic 5)



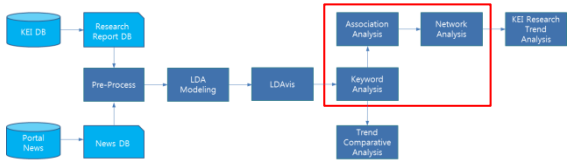
Term
기후
변화
가뭄
재난
폭염
해수면
기온
침수
사망
자연재해
오존
홍수
태풍
강수량
온도

“기후변화”

LDAvis (Topic 전체)

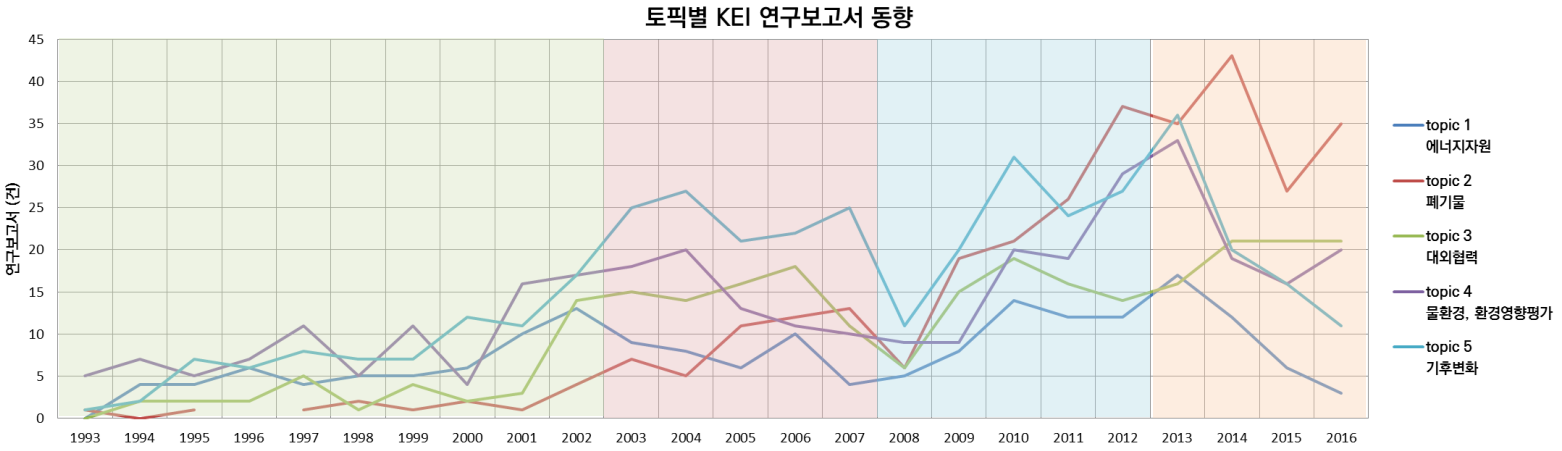
No.	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Title	에너지 자원	폐기물	대외협력	물 환경, 환경영향평가	기후변화
1	온실가스	폐기물	협력	지하수	기후
2	에너지	소음	포럼	환경영향평가	변화
3	전력	화학	북한	하천	가뭄
4	연료	처리	상하수도	습지	재난
5	가스	폐수	남북	정화	폭염
6	청정	하수	동북아	지형	해수면
7	탄소	수거	교류	수변	기온
8	풍력	유해	지속	상수원	침수
9	세제	막다	베트남	부지	사망
10	탄소세	소각	시민	서식지	자연재해
11	이산화탄소	쓰레기	남북한	한강	오존
12	천연가스	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	File Name	Description	Input	Output	Note
5월 상순	LDA Result Analysis		- 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_1012.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	

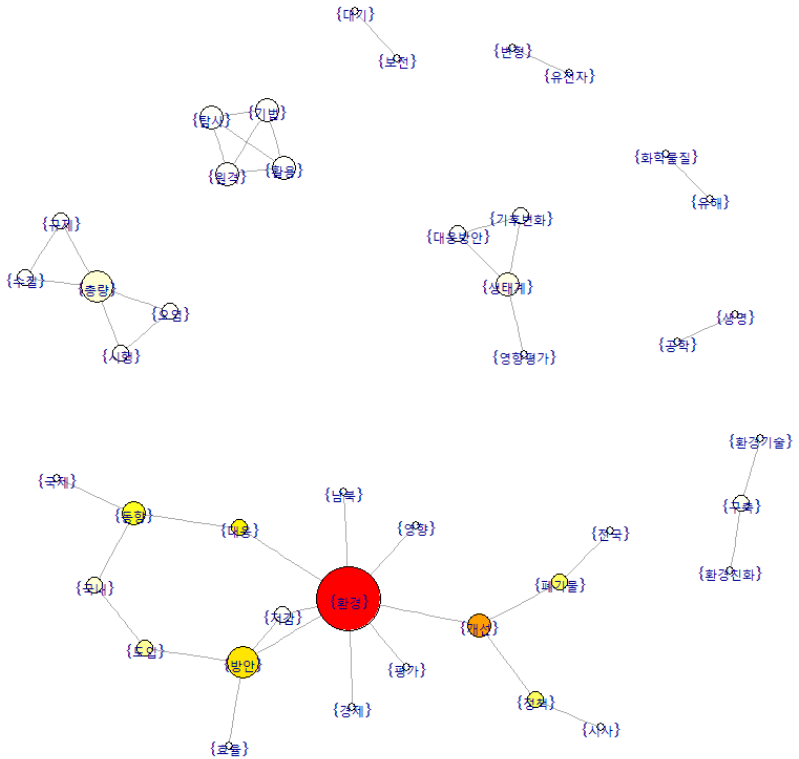
토픽별 KEI 연구보고서 동향 분석



- 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)

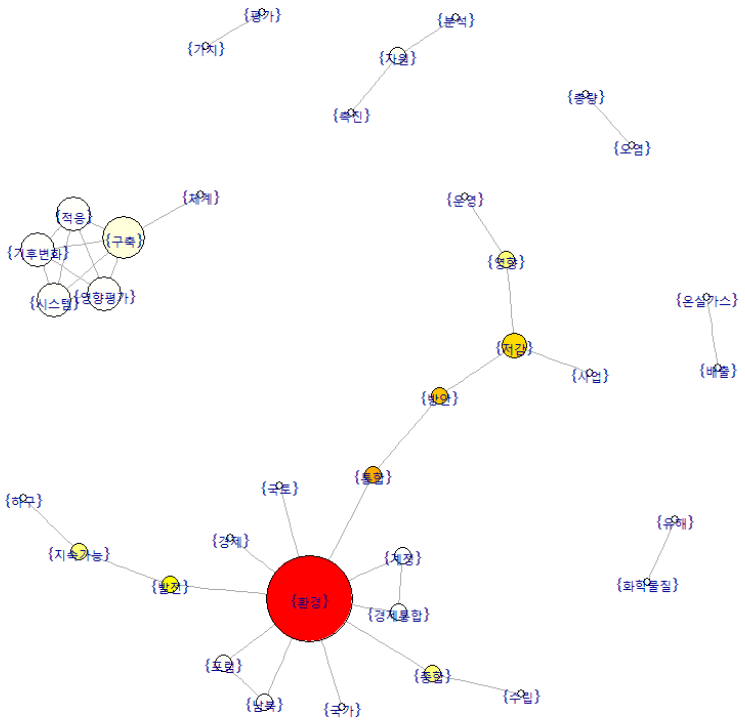


no	lhs		rhs	support	confidence	lift
1	유전자	=>	변형	0.0123	1.0000	81.3333
2	변형	=>	유전자	0.0123	1.0000	81.3333
3	기후변화	=>	생태계	0.0123	0.7500	45.7500
4	생태계	=>	기후변화	0.0123	0.7500	45.7500
5	기후변화	=>	대응방안	0.0123	0.7500	36.6000
6	대응방안	=>	기후변화	0.0123	0.6000	36.6000
7	영향평가	=>	생태계	0.0123	1.0000	61.0000
8	생태계	=>	영향평가	0.0123	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

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

연구 개요
선행 연구
연구 내용
연구 추진방법
기대효과

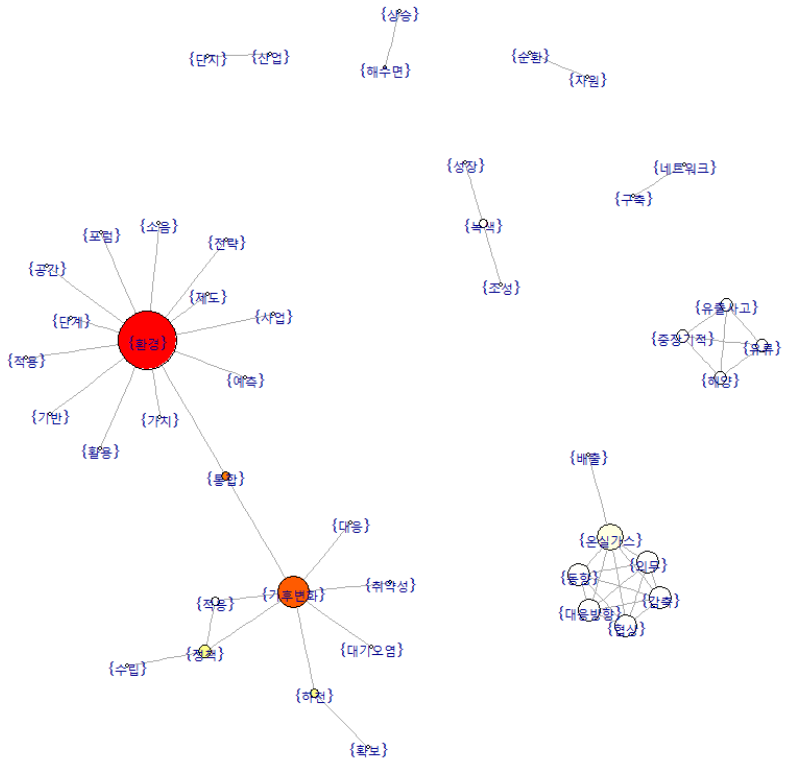
2. 키워드 연관성 및 네트워크 분석(2003-2007년)



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	0.0117	0.8333	23.6806
6	기후변화	=>	영향평가	0.0117	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

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

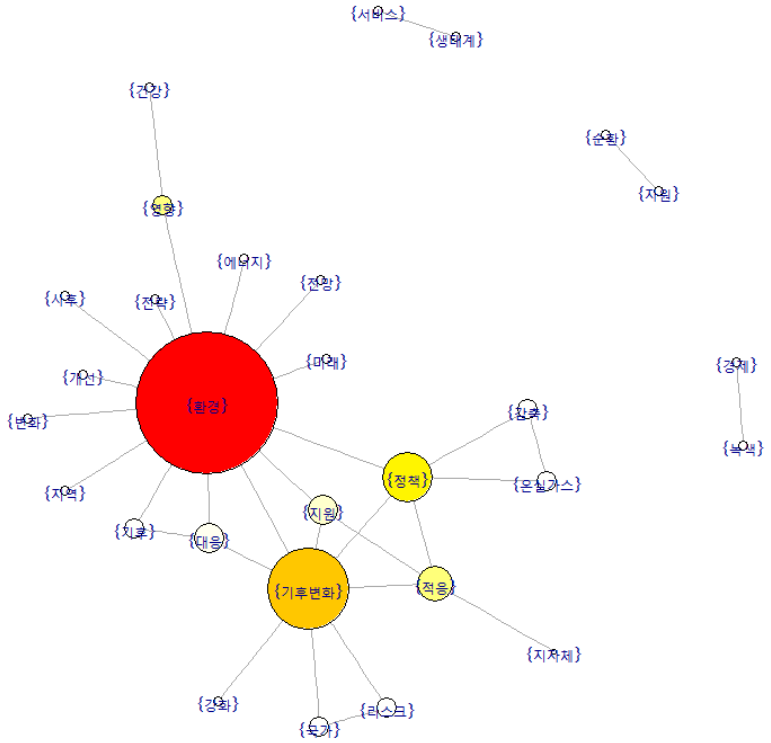
3. 키워드 연관성 및 네트워크 분석(2008-2012년)



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

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

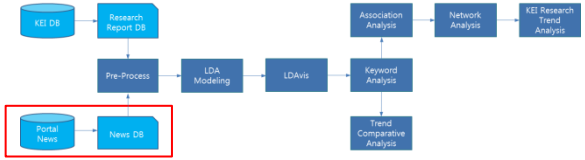
4. 키워드 연관성 및 네트워크 분석(2013-2016년)



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

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

Text Mining Process List



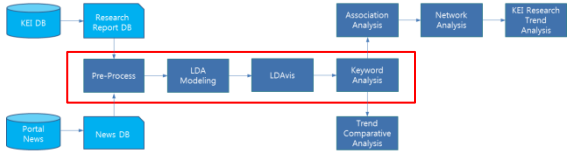
Plan 2017	Process	File Name	Description	Input	Output	Note
6월 상순	Data Collection	Naver_news.java	- Web crawling		Naver_news.csv	- 2012~2016년 네이버에서 제공하는 환경뉴스 데이터 수집 - Java-joup 사용

Data Collection_NaverNews

- 데이터수집범위에 대한 고민…

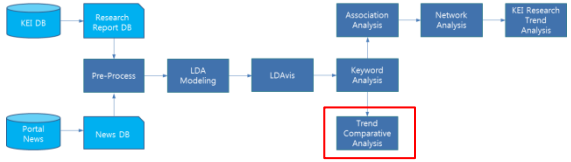
1. NAVER 뉴스 〉사회〉환경 관련 기사 전체 수집 or 환경사전 구축 〉환경 키워드 검색 〉키워드 관련 기사 수집
2. 2012년 ~ 2016년 (5개년) 데이터 수집 ?

Text Mining Process List



Plan 2017	Process	File Name	Description	Input	Output	Note
6월 하순	Pre-processing(2)		상동			
7월 상순	LDA Modeling(2)					
7월 상순	LDAvis(2)					
7월 하순	Keyword Analysis(2)					

Text Mining Process List



Plan 2017	Process	File Name	Description	Input	Output	Note
8월	Trend Comparative Analysis					
9월	시사점 도출 및 정책제언					
10월	향후 계획 수립		- 뉴스기사 댓글, 환경부 관련 페이스북, 트위터, 블로그, 유튜브 등을 통해 정책 수혜자 오피니언 마이닝 분석 실시			

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 키워드 빈도 분석, 연관성 분석, 토픽 클러스터링 등 다양한 텍스트 마이닝 분석 기법 집적 가능
  - 추후 이들 기법을 자동으로 처리하는 플랫폼을 구축하는 기초로 활용 가능

- **환경 키워드 라이브러리 구축 사전 작업**
  - 본 연구에서 구축하는 환경 키워드 사전 베타 버전은 향후 환경 키워드 사전으로 발전시켜 다양한 텍스트 마이닝 분석의 인프라로 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.

초 록

본 연구는 환경정책·평가연구원(이하 KEI) 연구동향을 파악하고 환경연구에 대한 사회적 연구 수요와의 조응 여부를 탐구하였다. 1993년 이래 KEI는 다양한 환경연구를 진행하고 있지만, 연구동향이 국민적 관심에 반응하고 있는 지 여부에 대한 회의가 존재한다. 이유는 일반적으로 개별 연구자는 연구주제 선정 시 시간적 제약 및 개인적 연구 성향에 의해서 연구수요 정보 파악 범위가 제한되기 때문이다. 또한 파악된 정보에 부여되는 우선순위가 개별 연구자의 선호에 영향을 받으므로 최신 정보 및 시의성 있는 연구 수요 반영에 제약이 존재하게 된다. 따라서 연구주제 선정 시 기존의 개별 연구자의 직관에 의존하는 방식을 보완할 필요가 있다. 이에 본 연구에서는 KEI에서 발행되는 연구보고서와 환경관련 뉴스기사를 수집한 후 토픽 클러스터링, 연관어 분석, 키워드 네트워크 분석 등 다양한 텍스트 마이닝 기법을 이용하여 장기간의 KEI 연구동향과 사회적 연구수요 동향을 시기별로 각각 추출하여 비교분석하였다. 분석결과는 다음과 같다. ~ (향후 내용 추가). 분석 결과를 근거로 환경정책 수요자의 선호를 파악하여 추후 환경연구 기획에 필요한 정보를 원내외 연구진에게 제공하고자 한다.

| 차 례 |