텍스트 마이닝을 이용한 KEI 연구동향 분석

5th Bigdata Research Team Seminar : 개별 과제 Proposal

2017. 04. 13

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

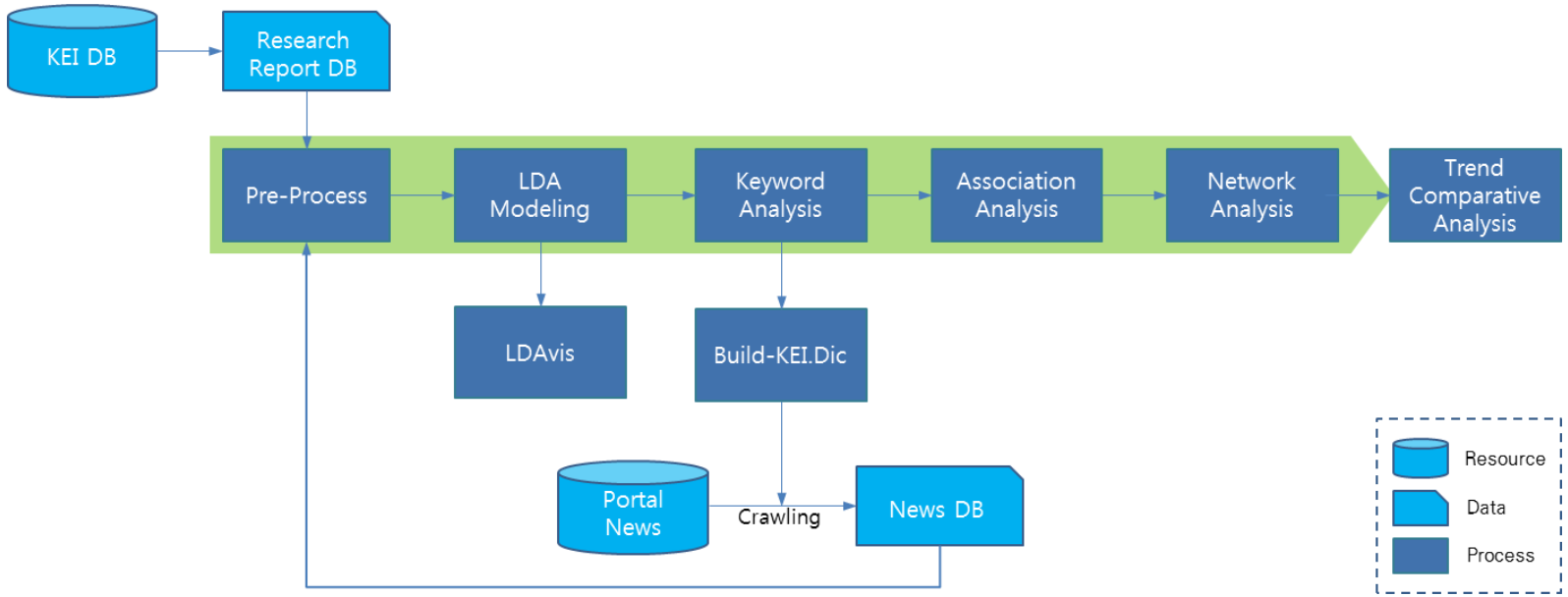
KEI 연구동향 파악

- 텍스트 마이닝 기법을 활용한 연구동향 분석
  - KEI가 설립된 1993년부터 2016년까지의 KEI DB에서 제공하는 연구보고서(제목, 목차, 요약, 날짜, 연구자명) 및 연구사업 계획서(제목, 날짜, 연구자명) 데이터를 분석에 활용
  - 24년 간(1993-2016) 연구사업계획서 2,614건, 연구보고서 1,697건 활용

매체별 환경분야 이슈 비교 분석

- 연구공급 동향과 연구수요 동향 비교 분석
  - 연구공급 동향 파악: 2000년부터 2016년까지의 KEI DB에서 제공하는 연구보고서 및 연구사업계획서 데이터 분석
  - 연구수요 동향 파악: 2000년부터 2016년까지의 네이버와 다음 등 언론매체에서 제공하는 뉴스 기사 및 댓글 데이터 분석
- 매체별 추출한 키워드를 시계열로 파악하고 두 시계열을 비교 분석

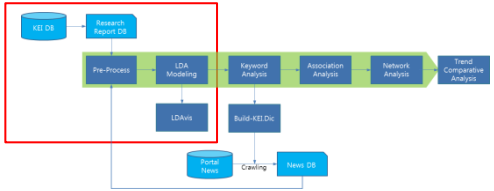
Text Mining Flowchart



Text Mining Process List

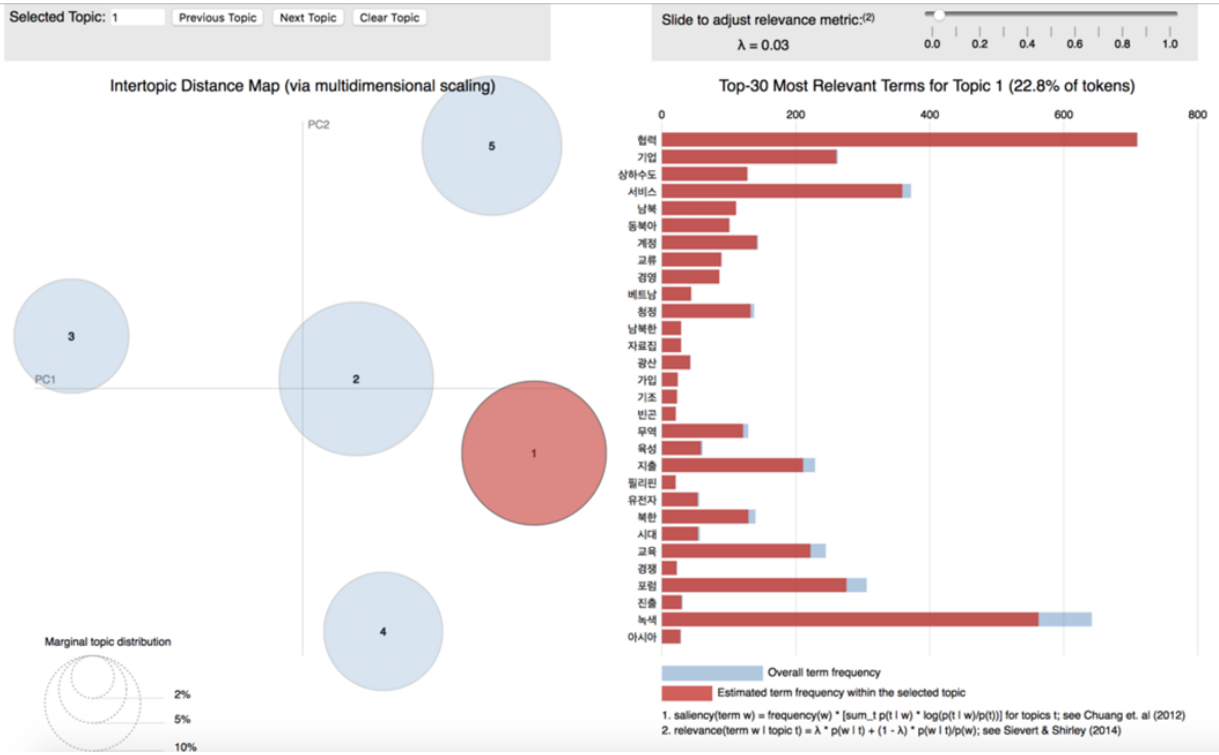
Plan 2017		Process	File Name	Description	Input	Output	Note
상반기	3월 상순	Pre-processing(1)	Text_Preprocess.R	- 형태소분석기 실행(KoNLP 등) - Low TF-IDF 값 제거 - 불용어처리 등 전처리 과정 - Word Lengths는 2글자 이상	kei.csv	DocumentTermMatrix	- 자문의견 (이명진 박사님) : 한글 처리 문제 (조사 등) -> 다양한 한글 전처리 방법(R package)을 통해 해결 가능함.
	3월 하순	LDA Modeling(1)	Running_LDA.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)	LDAvis.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	Keyword Analysis(1)					
	5월 하순	Association Analysis(1)					
	6월 상순	Network Analysis(1)					
	6월 하순	Build-KEI.Dic				KEI.Dic	- 자문의견(이명진 박사님) : 한글 라이브러리 문제 -> LDA를 활용한 토픽별 사전 구축
하반기	7월 상순	Data Collection	Naver_news.java Daum_news.java	- Web crawling	KEI.Dic	Naver_news.csv Daum_news.csv	Java-joup 사용
	7월 하순	Pre-processing(2)		상동		상동	상동
	8월 상순	LDA Modeling(2)					
	8월 상순	LDAvis(2)					
	8월 하순	Keyword Analysis(2)					
	9월 상순	Association Analysis(2)					
	9월 하순	Network Analysis(2)					
	10월	Trend Comparative Analysis					
	11월	시사점 도출 및 정책제언					
	12월	향후 계획 수립		- 뉴스기사 댓글, 환경부 관련 페이스북, 트위터, 블로그, 유튜브 등을 통해 정책수혜자 오피니언 마이닝 분석 실시			- 자문의견(우석진 교수님) : 환경정책 수혜자 DB 분석을 통해 정책적 시사점 도출 필요

Text Mining Process List



Plan 2017	Process	File Name	Description	Input	Output	Note
3월 상순	Pre-processing(1)	Text_Preprocess.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	Running_LDA.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	LDAvis.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

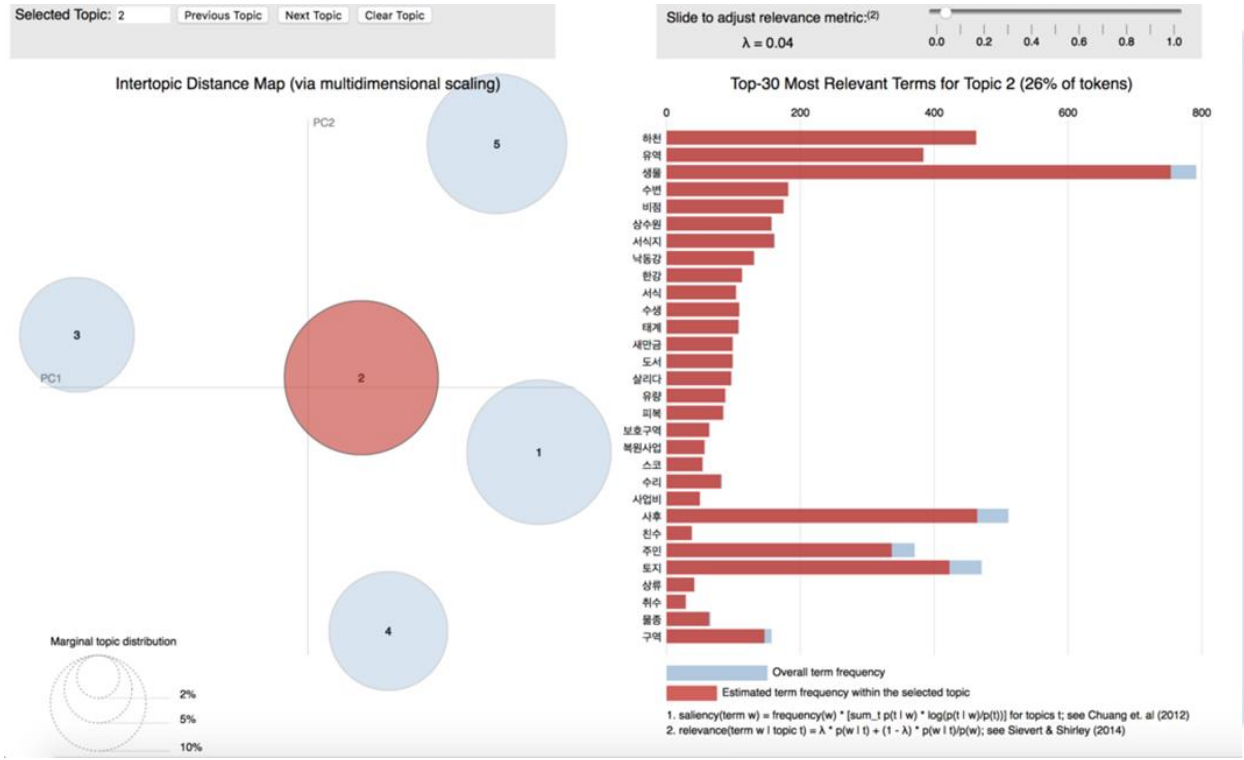
LDAvis (Topic 1)



Term
협력
남북
동남아
교류
베트남
무역
필리핀
북한
진출
아시아

“대외 협력”

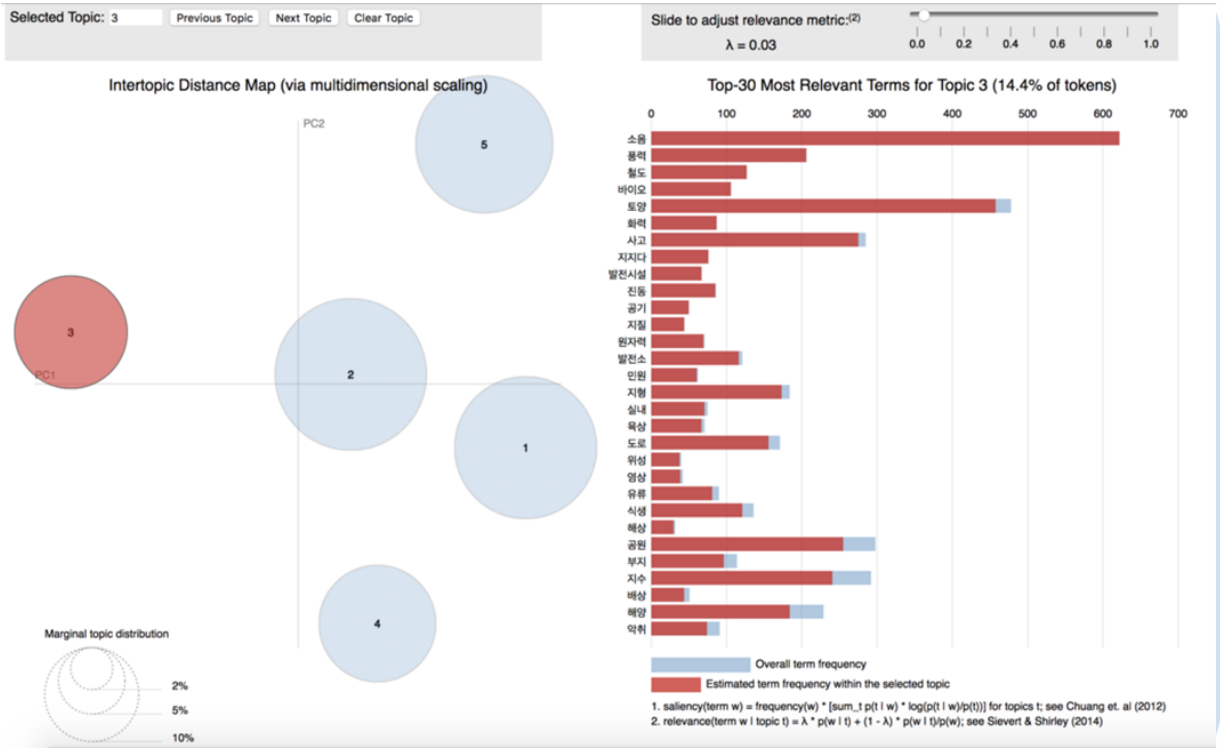
LDAvis (Topic 2)



Term
하천
유역
수변
비점
상수원
낙동강
한강
수생
새만금
친수

“물환경”

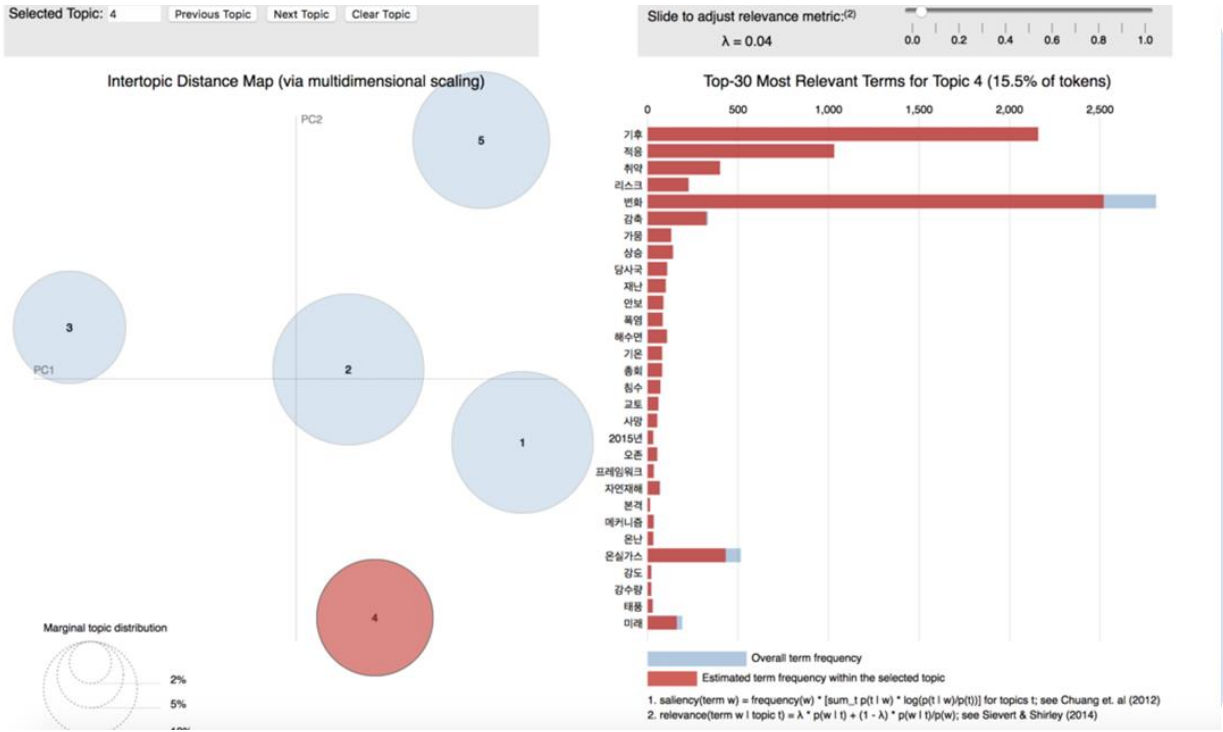
LDAvis (Topic 3)



Term
소음
풍력
철도
바이오
화력
발전시설
진동
원자력
발전소
악취

“에너지 자원”

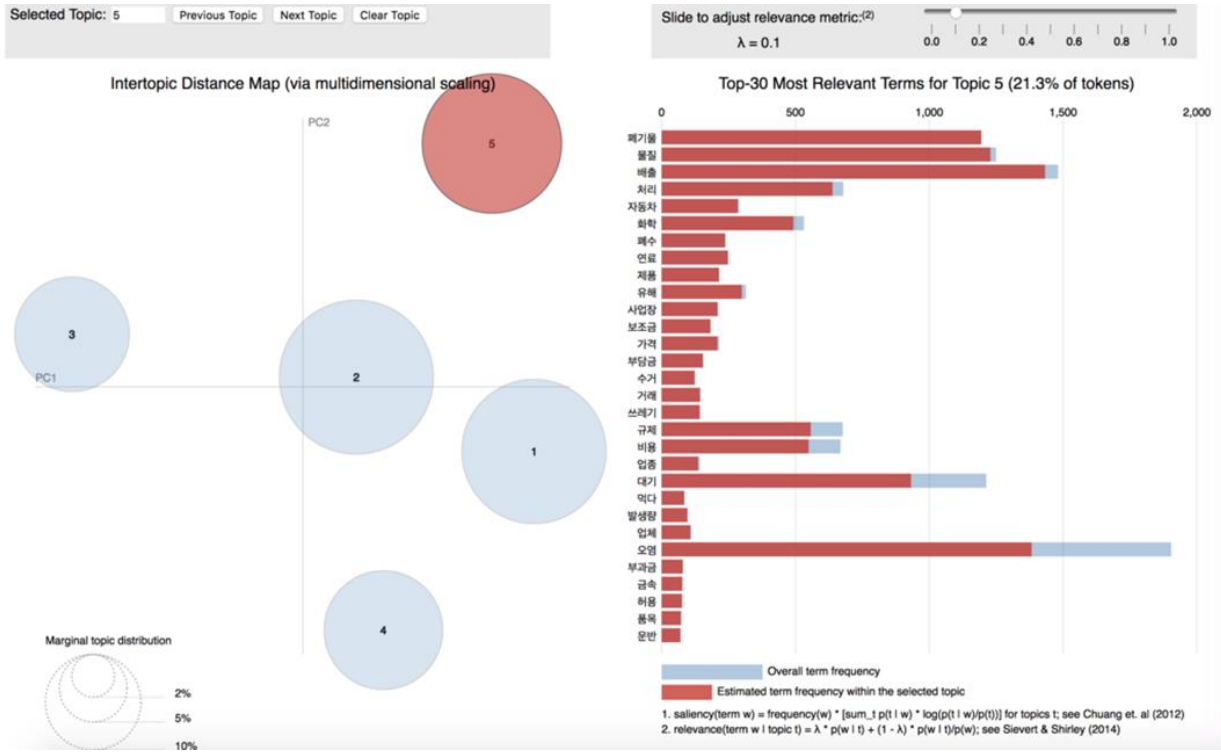
LDAvis (Topic 4)



Term
기후
가뭄
폭염
상승
해수면
오존
침수
온난
온실가스
태풍

“기후 변화”

LDAvis (Topic 5)



LDAvis (Topic 전체)

No.	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Title	대외 협력	물환경	에너지 자원	기후 변화	폐기물
1	협력	하천	소음	기후	폐기물
2	남북	유역	풍력	가뭄	배출
3	동남아	수변	철도	폭염	자동차
4	교류	비점	바이오	상승	화학
5	베트남	상수원	화력	해수면	폐수
6	무역	낙동강	발전시설	오존	연료
7	필리핀	한강	진동	침수	유해
8	북한	수생	원자력	온난	쓰레기
9	진출	새만금	발전소	온실가스	발생량
10	아시아	친수	악취	태풍	오염

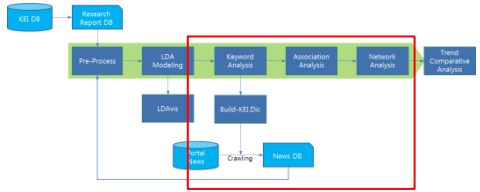
⋮

향후 …

5개의 Topic을 중심으로

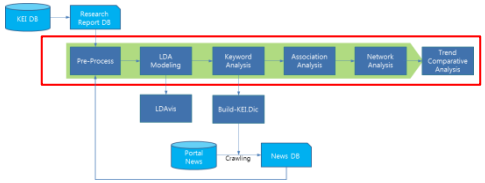
1. Keyword, Association, Network 분석 수행
2. KEI 환경 사전 구축 -> 뉴스 데이터 수집

Text Mining Process List



Plan 2017	Process	File Name	Description	Input	Output	Note
5월 상순	Keyword Analysis(1)					
5월 하순	Association Analysis(1)					
6월 상순	Network Analysis(1)					
6월 하순	Build-KEI.Dic				KEI.Dic	- 자문의견(이명진 박사님) : 한글 라이브러리 문제 -> LDA를 활용한 토픽별 사전 구축
7월 상순	Data Collection	Naver_news.java Daum_news.java	- Web crawling	KEI.Dic	Naver_news.csv Daum_news.csv	- Java-joup 사용

Text Mining Process List



Plan 2017	Process	File Name	Description	Input	Output	Note
7월 하순	Pre-processing(2)		상동			
8월 상순	LDA Modeling(2)					
8월 상순	LDAvis(2)					
8월 하순	Keyword Analysis(2)					
9월 상순	Association Analysis(2)					
9월 하순	Network Analysis(2)					
10월	Trend Comparative Analysis					
11월	시사점 도출 및 정책제언					
12월	향후 계획 수립		- 뉴스기사 댓글, 환경부 관련 페이스북, 트위터, 블로그, 유튜브 등을 통해 정책 수혜자 오피니언 마이닝 분석 실시			- 자문의견(우석진 교수님) : 환경정책 수혜자 DB 분석을 통해 정책적 시사점 도출 필요

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.