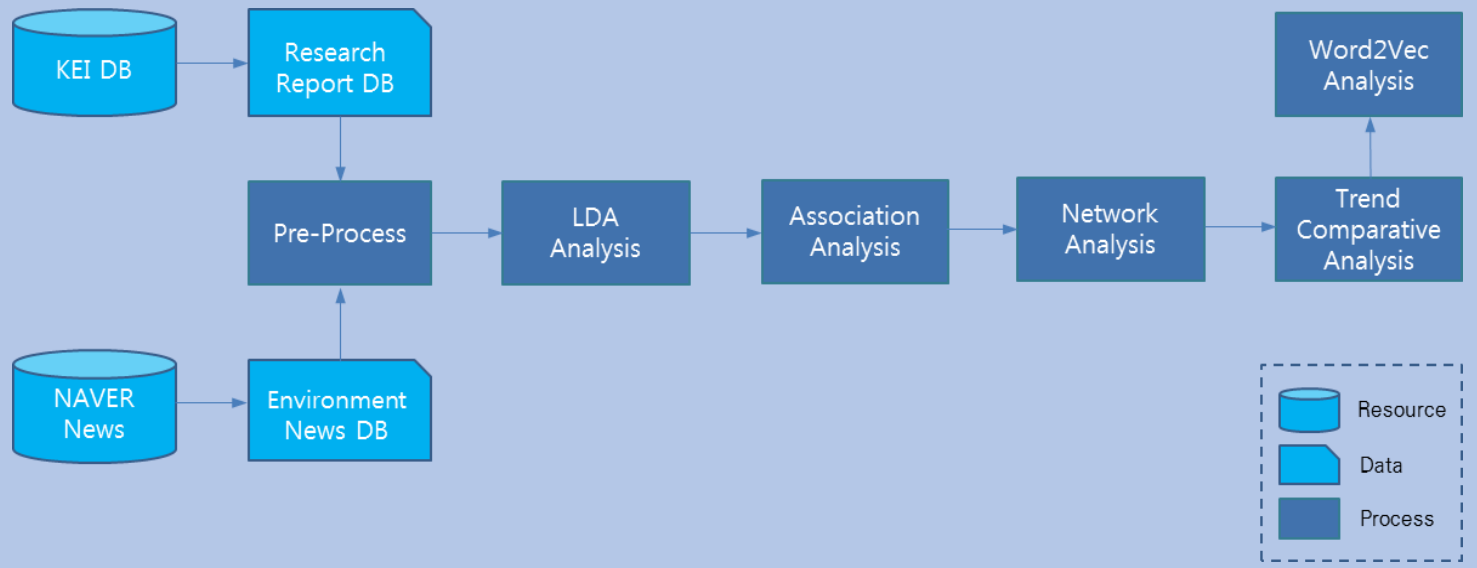
3. 연구결과: 2017년 연구성과

3. 텍스트마이닝을 이용한 KEI 연구동향분석 (김도연)

텍스트마이닝을 이용한 KEI 연구동향분석

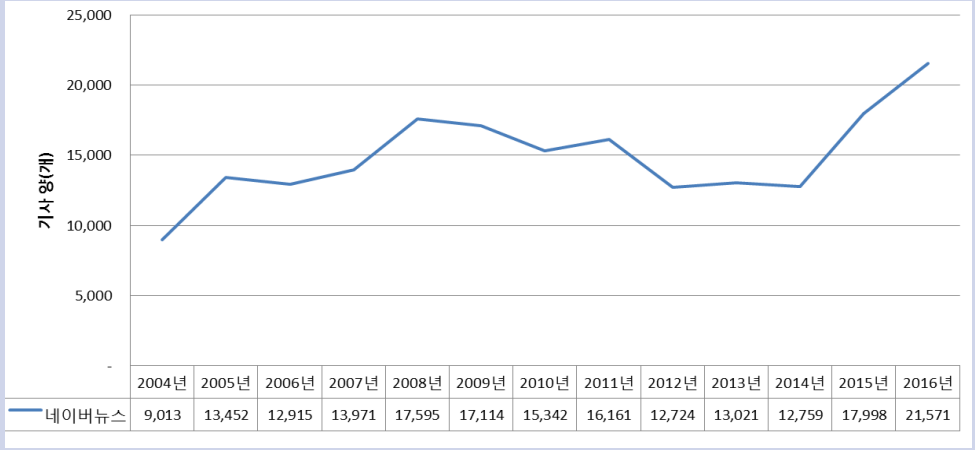
- 연구 내용
  - KEI 연구보고서(1993-2016)와 NAVER 환경뉴스(2004-2016) 데이터 이용
    LDA(Latent Dirichlet Allocation) 분석, 연관어 분석, 언어 네트워크 분석, word2vec 분석 수행
- 연구 프로세스 :



  - LDA Analysis: 매체 별 주요 토픽을 추출하고 이슈 변화를 살펴봄
  - Association & Network Analysis: 매체 별 중심 키워드를 찾고 환경 분야 이슈 발굴
  - Word2vec Analysis 분석: 환경 분야 이슈 관련 키워드 간의 문장 단위 관계를 구체적으로 살펴봄

데이터 수집 및 전처리

구분	내용
수집 도구	Java HTML Parser - jsoup
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	총 101개 (부록2 참조)
산출 양	**총 193,636개**



	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

텍스트 데이터 전처리 과정

1. 형태소분석기 수행
   - R에서 제공하는 형태소분석기 패키지 (KoNLP)를 사용하여 명사 추출
2. 특수문자, 특정단어 등 불용어 삭제
3. 단어길이 한글자 삭제
4. 출현빈도가 매우 낮은 단어(Sparse Terms) 삭제
5. Low TF-IDF 단어 삭제
6. DTM(Document Term Matrix)형태로 변형

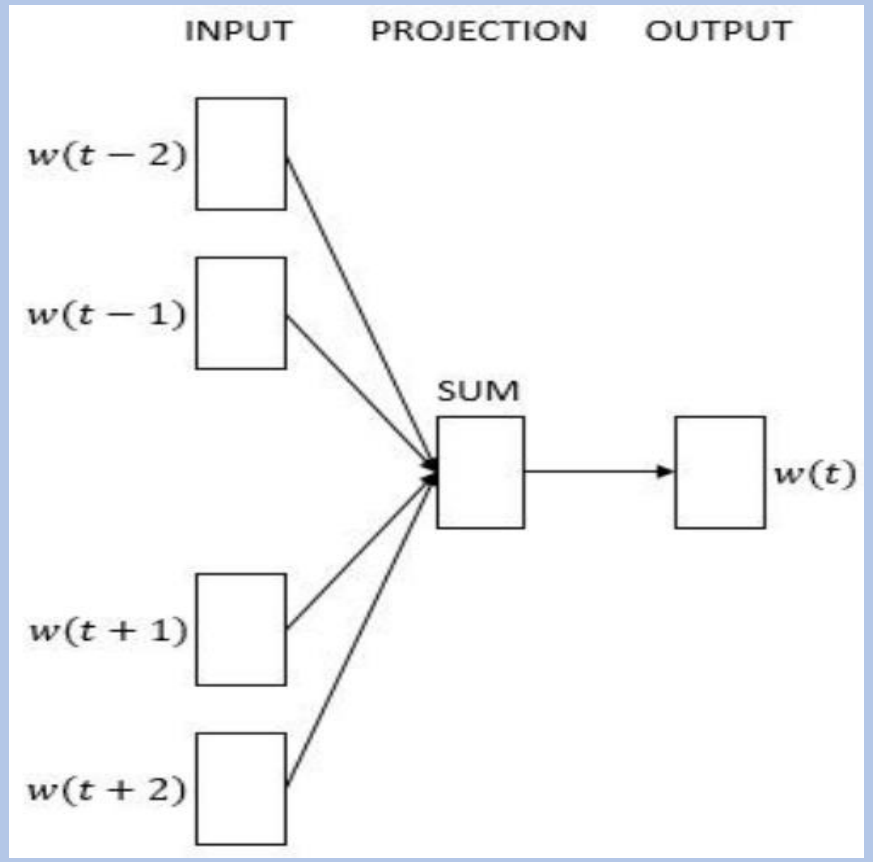
예)

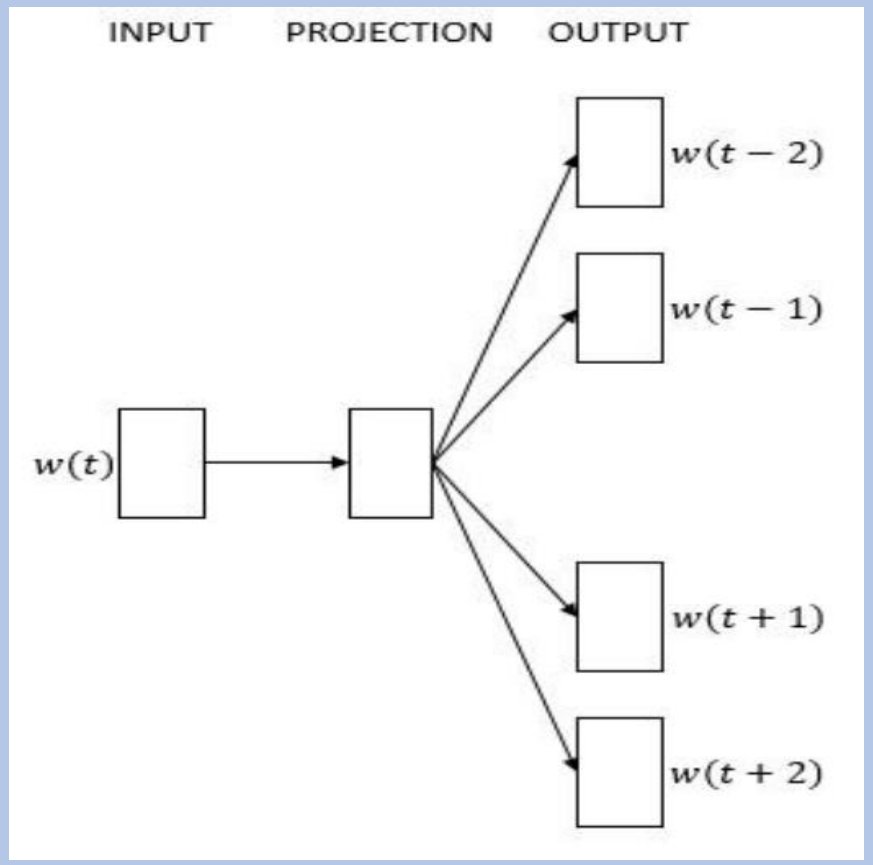
Term / Document	기후	오염	…	한강
보고서 1	0	2	…	1
보고서 2	2	1	…	1
…				…
보고서3	0	1	…	5

텍스트 마이닝 방법론

- LDA : 분석 대상 문서 단어 분포 기반 문서 토픽 추출
- 키워드 연관성 분석 : 동시 출현 빈도가 높은 단어 조합 파악
- 키워드 네트워크 분석: 단어 간 연결 관계 파악
  - 밀도: 네트워크 내 연결점 간의 관계가 나타나는 빈도
  - 중심성: 다른 연결점과의 연결 정도
- Word2vec: 문장 내 단어 간 관계 파악
  - CBOW(Continuous Bag Of Words) : 주변 단어를 이용하여 특정 단어를 예측
  - Skip-gram: 특정 단어를 이용하여 주변 단어를 예측

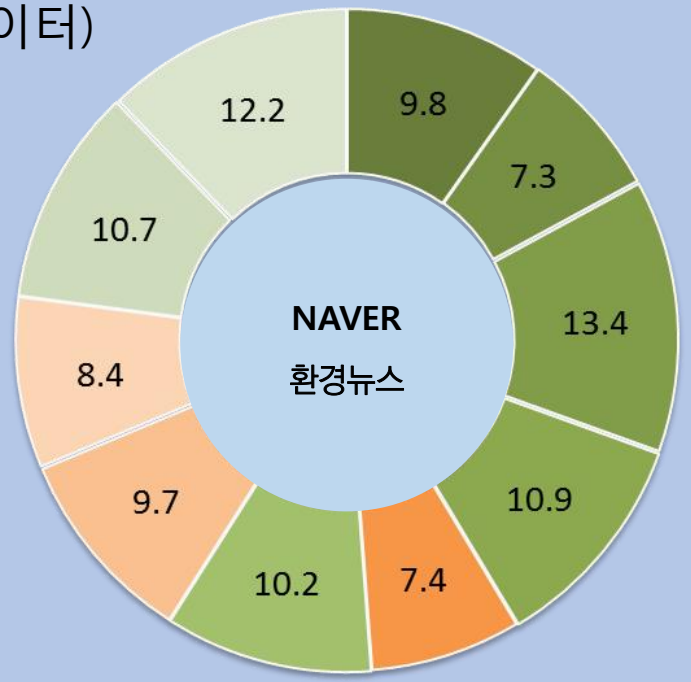
Word2Vec : CBOW , Skip-gram

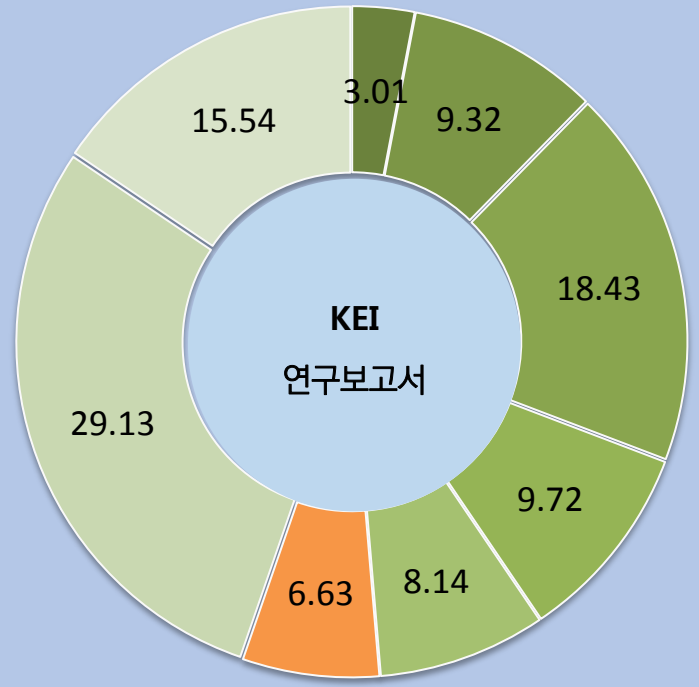




LDA 분석: 매체 별 토픽 구성 비교

- 매체별 13개년(2004-2016) 전체 데이터 분석
  - 공통 토픽(녹색) : 대기 및 기후변화, 폐기물, 환경영향평가, 에너지자원, 수질오염, 대외협력
  - 기타 토픽(주황색) : KEI 연구보고서(생태계), NAVER 환경뉴스(유전자 변형/소음, 해양/풍력, 보건/데이터)





기후변화 폐기물 환경영향평가 에너지자원 유전자변형, 소음 수질오염 해양, 풍력 보건, 데이터 대외협력 물환경

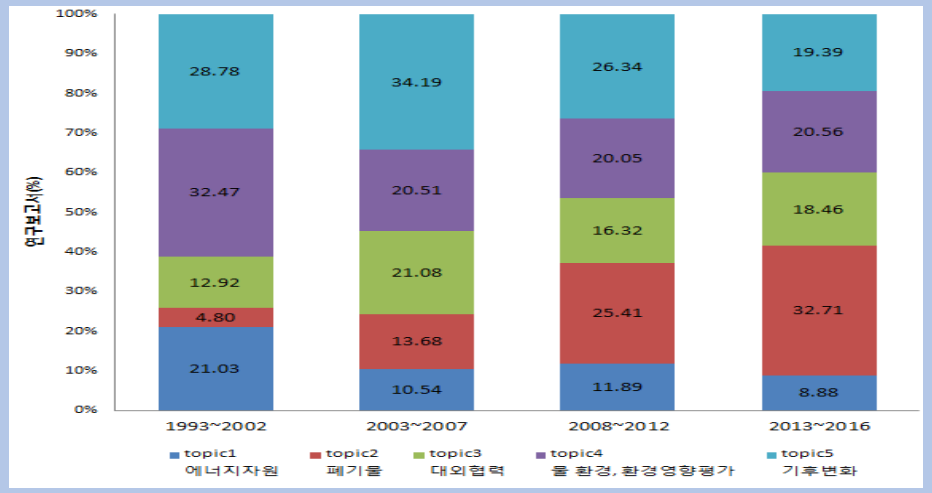
기후변화 폐기물 환경영향평가 에너지자원 대기오염 생태계 수질오염 대외협력

LDA 분석: 기간 별 토픽 구성 변화

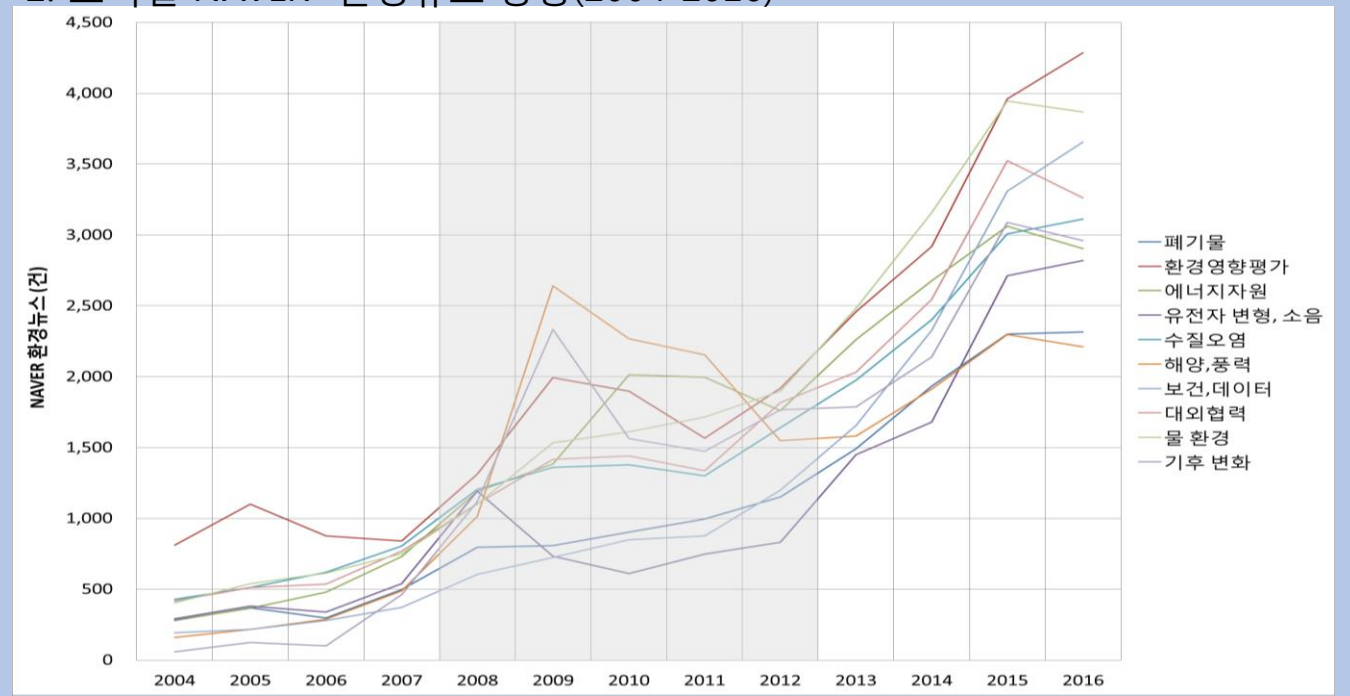
- KEI 연구보고서
  - 2구간(2003~ 2007년) : 대기 및 기후변화 관련 연구 활발
  - 3구간(2008~ 2012년) : 폐기물, 물 환경/환경영향평가 연구가 급증
  - 4구간(2013~ 2016년) : 폐기물 연구 활발, 2015년 기점으로 연구의 양 감소
- NAVER 환경뉴스
  - 2구간(2008~2012년)을 제외하고 기사 양이 증가하는 추세를 보임
    - 2008년은 '유전자 변형/소음' 관련 기사가 급증함
    - 2009년은 '해양/풍력', '기후변화', '환경영향평가' 관련 기사가 급증함
- KEI 연구: '기후변화' 선도/ '수질오염' 후행/ '에너지-자원' 부족
  - 최근 '유전자 변형, 소음' '보건, 데이터' 환경 뉴스 증가 추세가 반영되지 못함

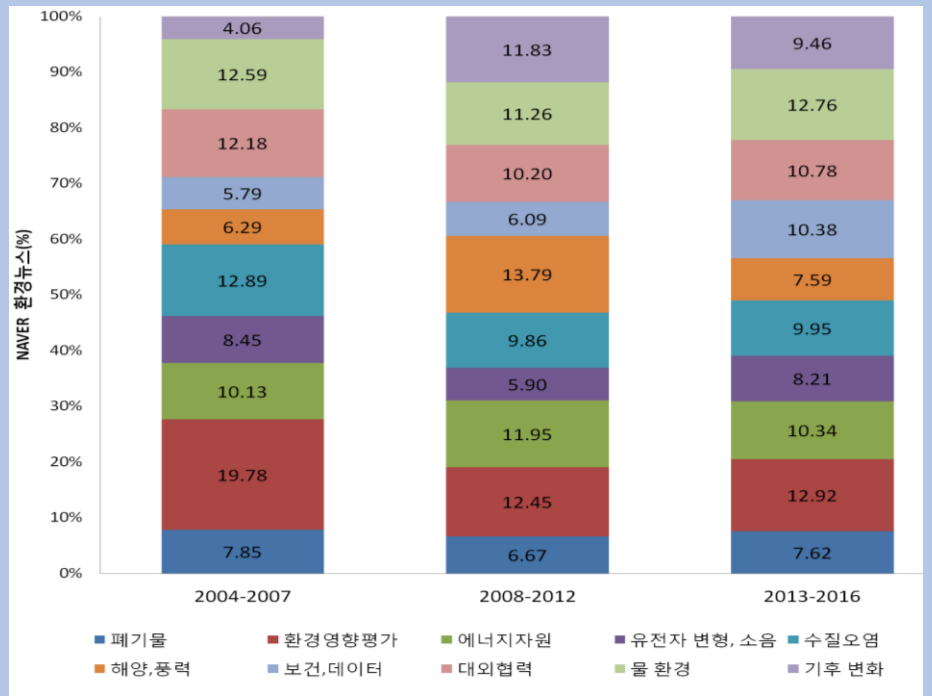
기간 별 토픽 동향 변화

1. 토픽별 KEI 연구보고서 동향(1993-2016)



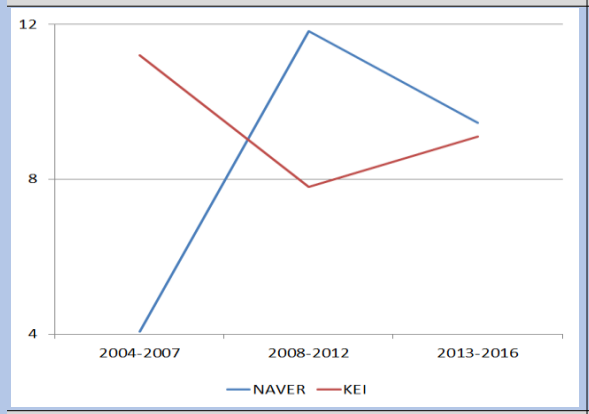
2. 토픽별 NAVER 환경뉴스 동향(2004-2016)



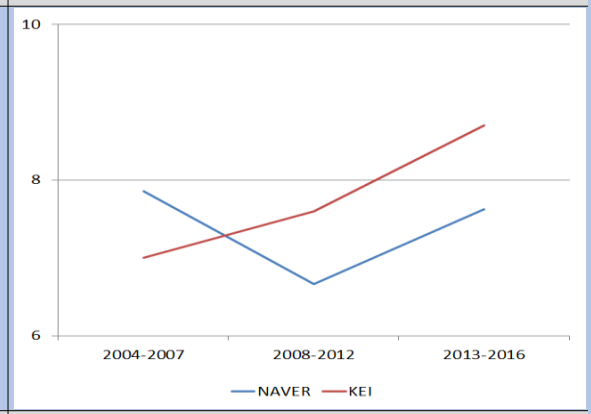


4. 공통 토픽 동향 비교

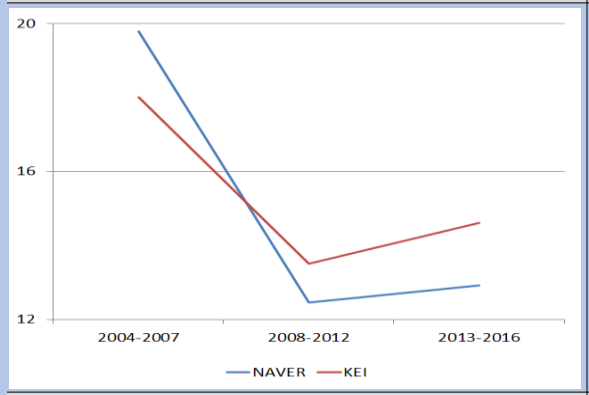
1. 기후변화



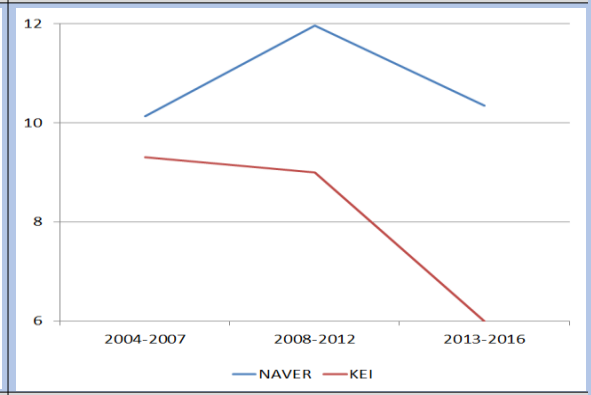
2. 폐기물



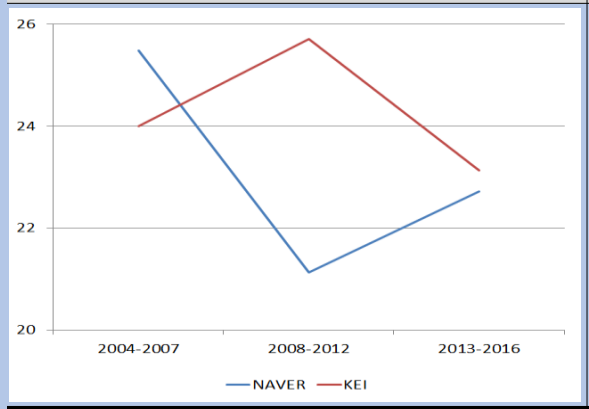
3. 환경영향평가



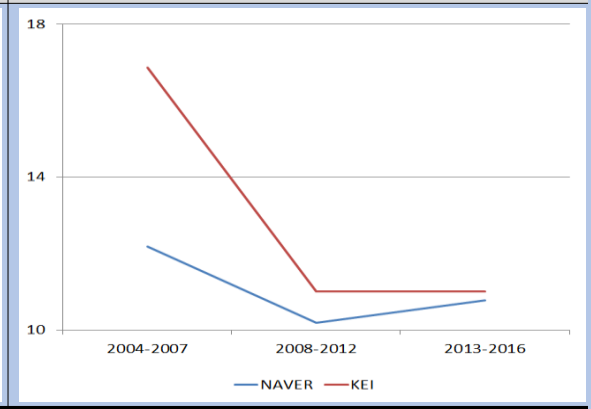
4. 에너지 자원



5. 수질오염

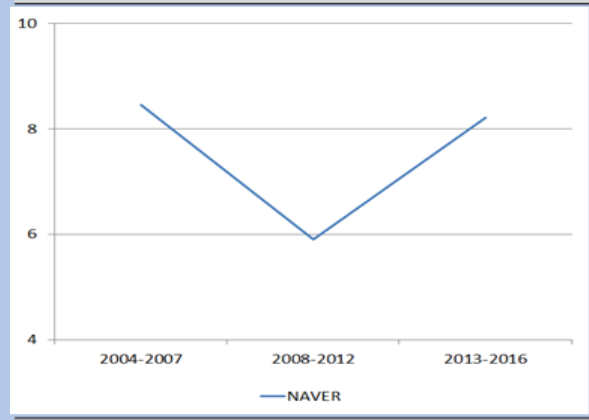


6. 대외협력

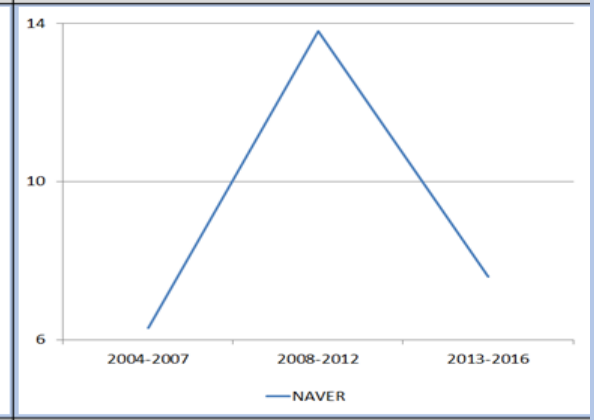


5. 기타 토픽 동향

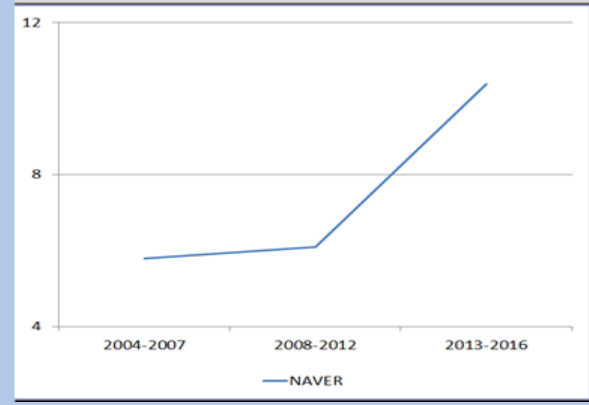
유전자 변형, 소음



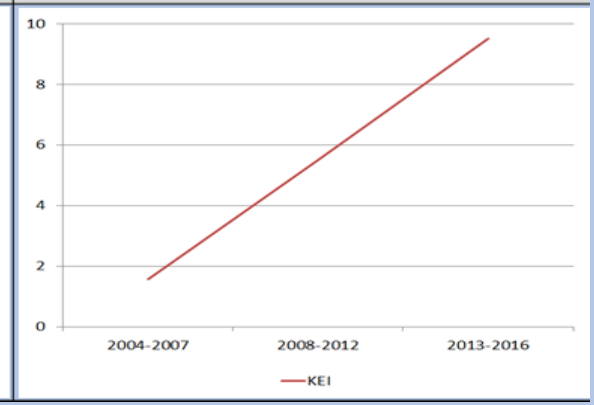
해양, 풍력



보건, 데이터



생태계



- **토픽 동향 유사:** '환경영향평가', '에너지 자원', '대외협력'
- **토픽 동향 차이**: KEI '기후변화' 선도, '수질오염' 후행, '에너지-자원' 부족
- **최근** '유전자 변형, 소음'과 '보건, 데이터' 관련 환경 이슈가 대두

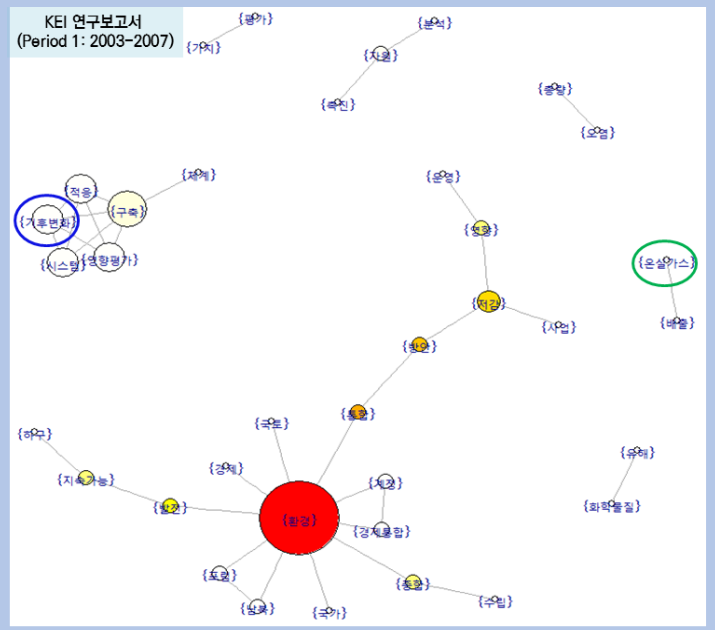
키워드 네트워크 분석: 기간별 네트워크 변화

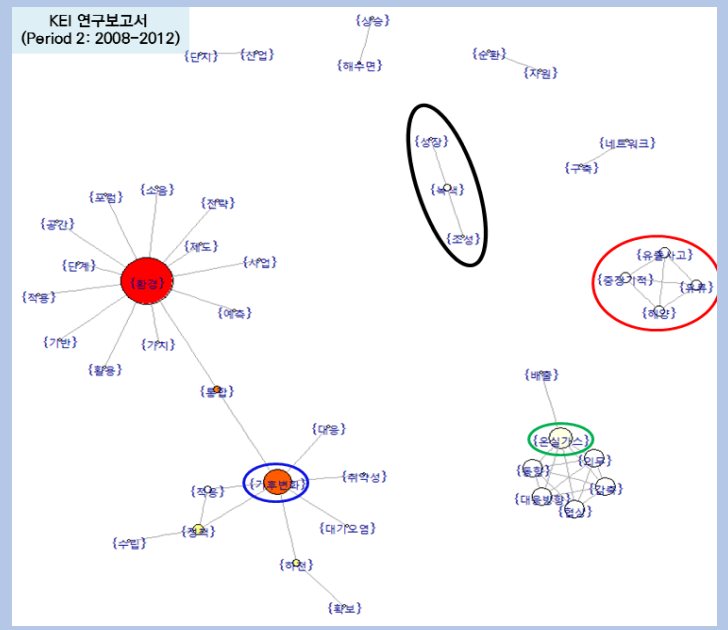
- 기후변화 : KEI 가 Trend를 선도하였으나 최근 추세 반영 미흡
  - 2003~ 07년 KEI 연구보고서 기후변화 네트워크 생성 → 3구간 NAVER 환경 뉴스 기후변화 네트워크 선도
  - 2008~16 년 KEI 연구보고서 기후변화 매개중심성 강화 vs. NAVER 환경뉴스 기후변화 관련 키워드 네트워크 분화 (태풍, 한파, 대설)
- 사건사고: 해양 오염 관련 Trend 는 NAVER 선도
  - 2003~07년 NAVER 뉴스 해양오염 키워드 네트워크 생성 : 2008~12 KEI 연구보고서 해양오염 키워드 생성

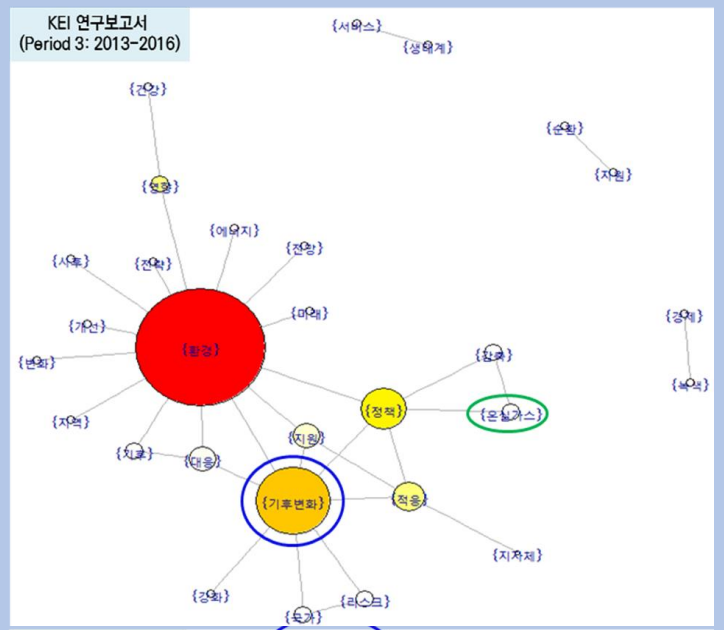
키워드 네트워크 변화

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

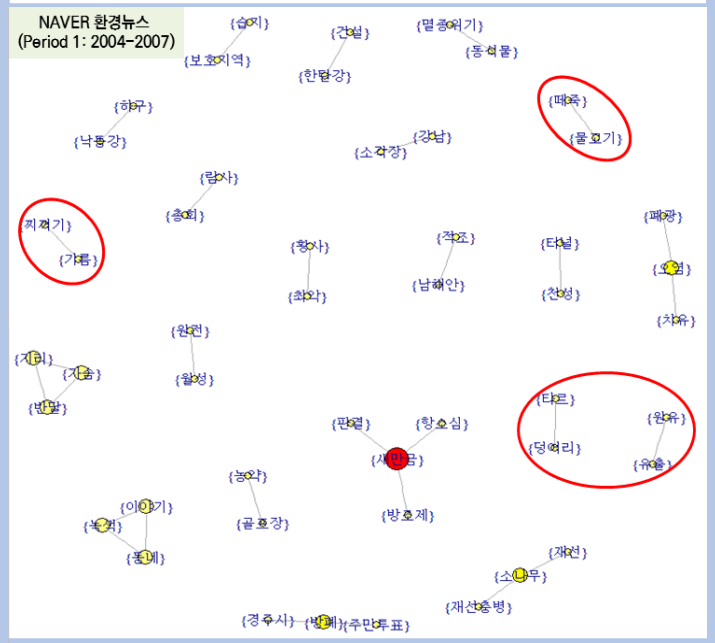
KEI 연구보고서

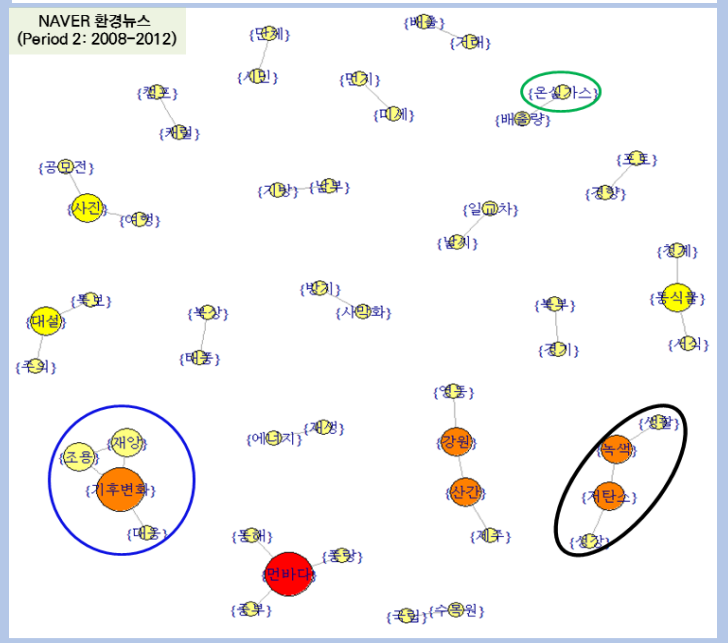


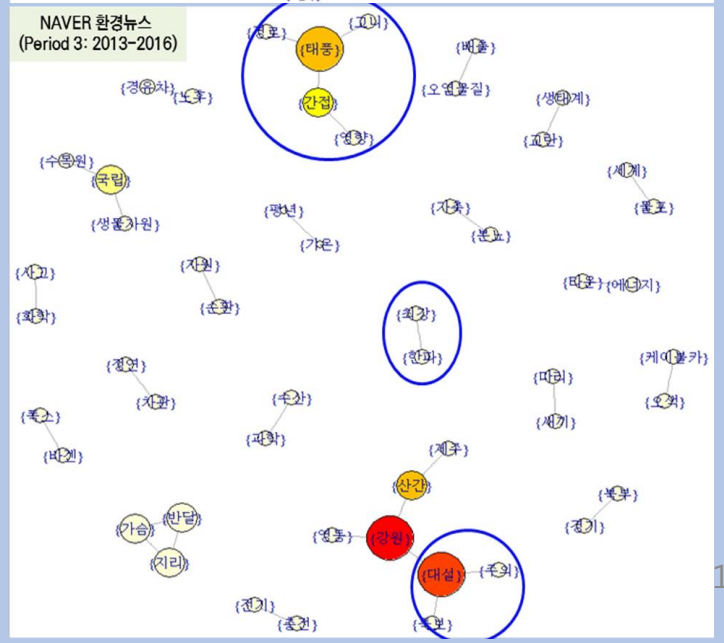




NAVER 환경 뉴스





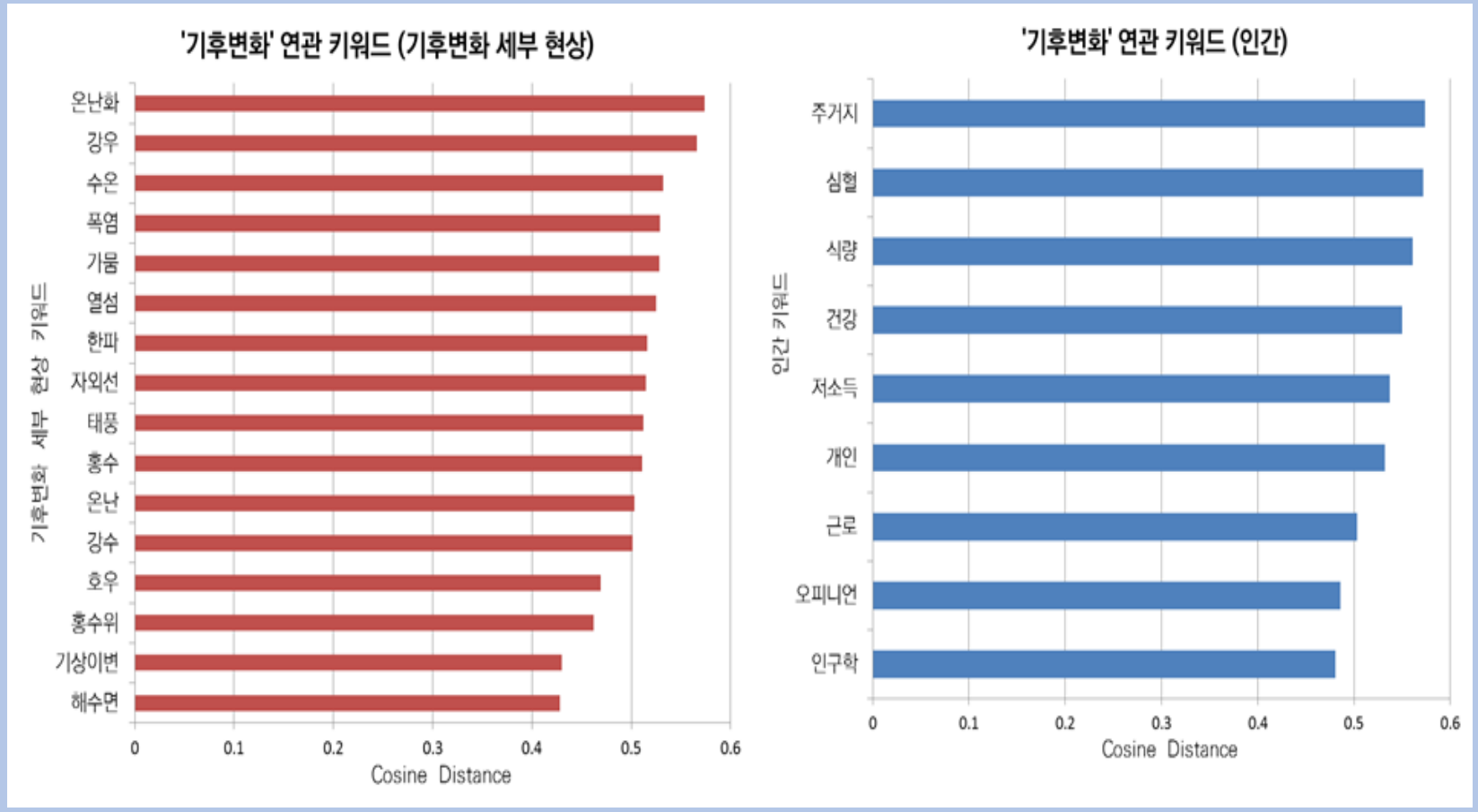


Word2Vec 분석: Keyword 간 문장 내 관계

- Skip-gram 모델: '기후변화' 관련 키워드를 넣어 주변에 있는 단어들을 예측
- 기후변화 세부 현상 키워드: '기후변화' Skip gram 분석 결과로부터 '온난화', '홍수', '가뭄' 선정
  - 3가지 키워드에 대한 Word2Vec 분석을 매체별 수행 후 비교 분석
    - KEI 연구보고서(1993~2016) 제목, 초록, 목차
    - NAVER 환경뉴스(2004-2016) 제목
- 분석 결과 : 연구보고서(국민의 삶의 질) 와 뉴스(관심사)의 강조점 차이 발견
  - 온난화: 인간 관련 단어(KEI) vs. 생물 및 식량 관련 단어(NAVER)
  - 홍수: 대한민국 지역 (KEI) vs. 중국 지역(NAVER)
  - 가뭄: 인간 관련 단어(KEI) vs. 농업 관련 단어(NAVER)

기후변화 Skip-gram: 기후/인간 관련 단어

- KEI 연구보고서 자료 활용
- 기후변화 세부현상 지칭 단어 및 인간 관련 단어가 기후변화와 관계
- 기후변화 세부현상 지칭 단어 중 '온난화','홍수','가뭄' 선정
  - '태풍' : 태풍 이름 관련 단어가 관계 깊은 단어로 자주 출현 → 제외
  - '한파','폭설','폭염','폭우': 기상 관련 불용어(33도, 영하, 곳곳, 오후) 가 관계 깊은 단어로 자주 출현 → 제외

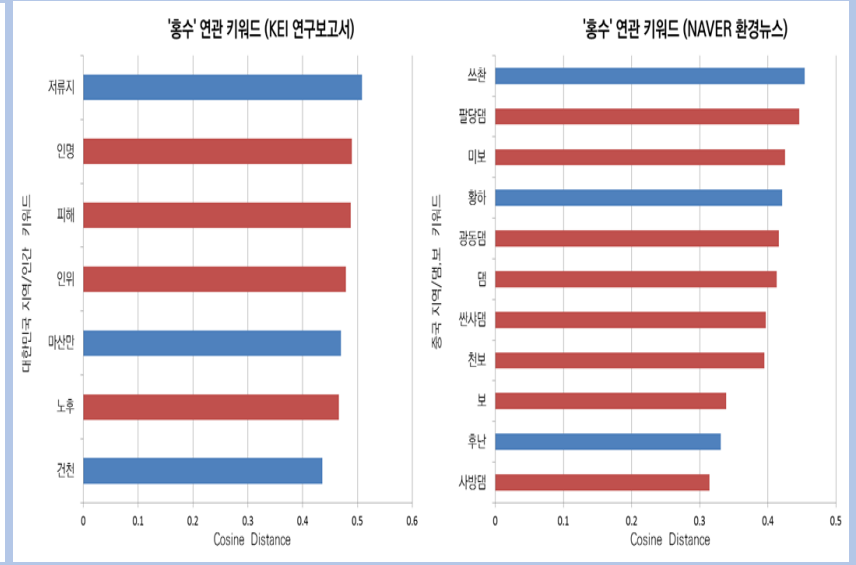


Word2Vec 분석: Keyword 간 문장 내 관계 분석

1. 온난화

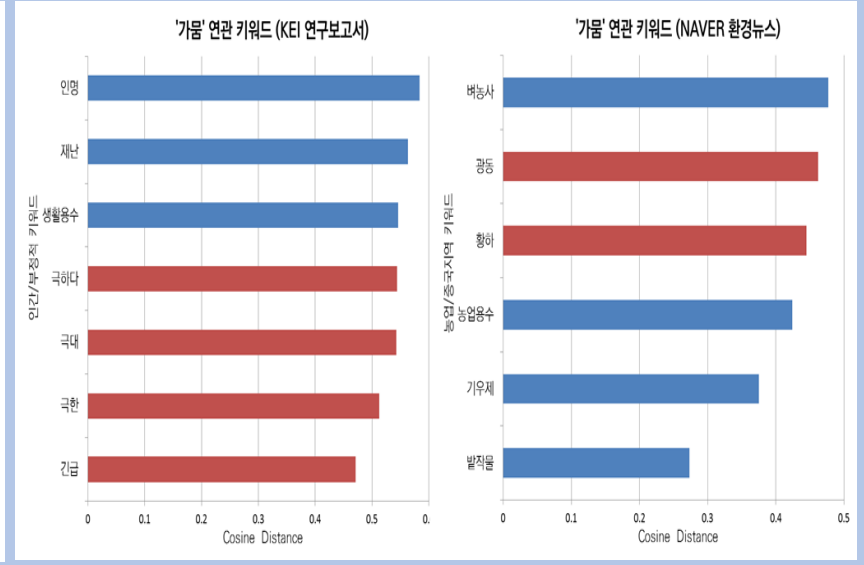
- KEI 연구보고서:
  인간 키워드(거주, 인류, 인명, 호흡기계, 심혈) 통해 온난화 관련 환경연구는 호흡기계 질환 및 심혈관계 질환과 관련된 연구가 활발히 진행되었음을 확인함
- NAVER 환경뉴스:
  생물 및 식량 키워드(플랑크톤, 산호초, 말매미, 식량) 통해 온난화의 영향을 많이 받는 플랑크톤, 산호초, 말매미와 관련된 환경문제 이슈가 대두되었음을 확인함

2. 홍수



- KEI 연구보고서:
  대한민국 지역 키워드(건천, 마산만) 통해 홍수 피해를 입은 국내 지역에 관한 환경연구가 활발히 진행되었음을 확인함
- NAVER 환경뉴스:
  중국 지역 키워드(후난, 황하, 쓰촨) 통해 환경뉴스의 관심사는 대규모 홍수 피해의 심각함에 중점을 두고 있음을 파악함

3. 가뭄



- KEI 연구보고서:
  인간 키워드(생활용수, 재난, 인명) 통해 환경 연구의 관심사는 가뭄으로 인한 생활용수 부족 문제에 중점을 두고 있음을 파악함
- NAVER 환경뉴스:
  농업 키워드(밭작물, 기우제, 농업용수, 벼농사) 통해 환경뉴스의 관심사는 가뭄으로 인한 농업용수 부족 문제에 중점을 두고 있음을 파악함

결론

LDA 분석 결과

- NAVER 환경뉴스와 KEI 연구보고서 전체 데이터(2004-2016)를 비교 분석한 결과 공통적으로 '기후변화', '폐기물', '환경영향평가', '에너지자원', '수질오염', '대외협력' 토픽을 중요하게 다루고 있음을 확인
- **환경뉴스는 '보건/데이터', '유전자 변형/소음' 관련 환경기사가 최근 많이 보도 되어 향후 성장 가능성 있는 연구 주제로 판단**

연관어 및 네트워크 분석 결과

- 환경뉴스는 기후변화의 세분화된 주제를 중심으로 보도가 되고 있으며, 환경연구는 기후변화 일반을 중심으로 연구
- **따라서 향후 기후변화의 세부주제에 대한 연구가 유망한 연구 주제로 파악됨**
- 포괄적인 관점에서 본 연구는 환경연구의 주제가 환경뉴스의 주제를 시차를 두고 추종하는 경향이 있음을 확인함
- **따라서 이러한 시차를 메울 수 있는 연구에 대한 요구가 앞으로도 지속될 전망**

Word2Vec 분석 결과

- 환경연구의 관심사는 국민의 삶의 질에 중점을 두고 있는 반면, 환경뉴스의 관심사는 기후변화로 인한 피해의 심각함에 중점
- **정책 성과 제고를 목적으로 하는 환경연구문헌과 사실보도를 목적으로 하는 언론 간의 매체의 차이를 반영**