환경 빅데이터 분석 및 서비스 개발

중간자문회의 (2017.6.29)

한국환경정책·평가연구원

강 성 원

1. 연구일반

2. 연구 진행 상황

3. 향후 계획

1. 연구 일반

개관

구분	내용	
연구성격	일반사업(연구형), 계속사업	
연구기간	2017.1 ~2017.12	
연구진	강성원 연구위원(책임) 장기복 선임연구위원 진대용 부연구위원 홍한움 부연구위원	한국진 전문원 김진형 연구원 김도연 위촉연구원 강선아 위촉연구원 정은혜 위촉연구원 이동현 한국산업기술대 교수(위탁)
자문위원	내부	명수정 연구위원 배현주 연구위원 이명진 부연구위원
	외부	김종률 과장 (환경부 정책총괄과) 우석진 교수 (명지대학교 경제학과) 강희찬 교수 (인천대학교 경제학과) 이성호 박사 (한국개발연구원)
자문일정	착수자문회의: 2017년 3월 30일 중간자문회의: 2017년 6월 29일 최종자문회의: 2017년 10월	

기간, 인력, 예산

- 기간: 2017년 1월 – 2017년 12월

- 인력: 박사급 연구원 5명(1명 원외), 전문원 1명, 연구 보조인력 4명 투입
  - 박사급 연구원 2명 채용: 진대용 부연구위원, 홍한움 부연구위원

- 예산: 3억 6백만 원 책정
  - 위탁연구비 4천 만원 책정: '딥러닝을 활용한 환경리스크 예측'
    - 위탁과제 책임자: 한국 산업기술대학교 이동현 교수

연속사업: 3년 단위 연구단계 설정

- 1단계(2017-19): 환경 빅데이터 연구 시작/연구자료 및 분석 알고리듬 공개 시작
- 2단계(2020-22): 환경 빅데이터 분석 플랫폼 설계/빅데이터 활용 공공 서비스 설계
- 3단계(2023-25): 환경 빅데이터 분석 플랫폼 자동화 시도/공공환경 서비스 시범 사업

환경 빅데이터 분석 및 서비스 개발 연차계획

	환경 빅데이터 연구	환경 빅데이터 연구 인프라	원내외 빅데이터 서비스
1기 (2017-19)	• **환경 빅데이터 연구 시행**	• 자료 및 알고리듬 축적/공개	• 원내 연구 및 경영정보 서비스
2기 (2020-22)	• 발신주기 단축	• 빅데이터 연구 과정 자동화 • 환경 빅데이터 분석 플랫폼 설계	• 연구기획 평가 및 준비 서비스 • 공공 서비스 설계
3기 (2023-25)	• 시의성 중심 발신체계 개편	• 환경 빅데이터 분석 플랫폼 지능화 시도	• 공공 서비스 시범 사업

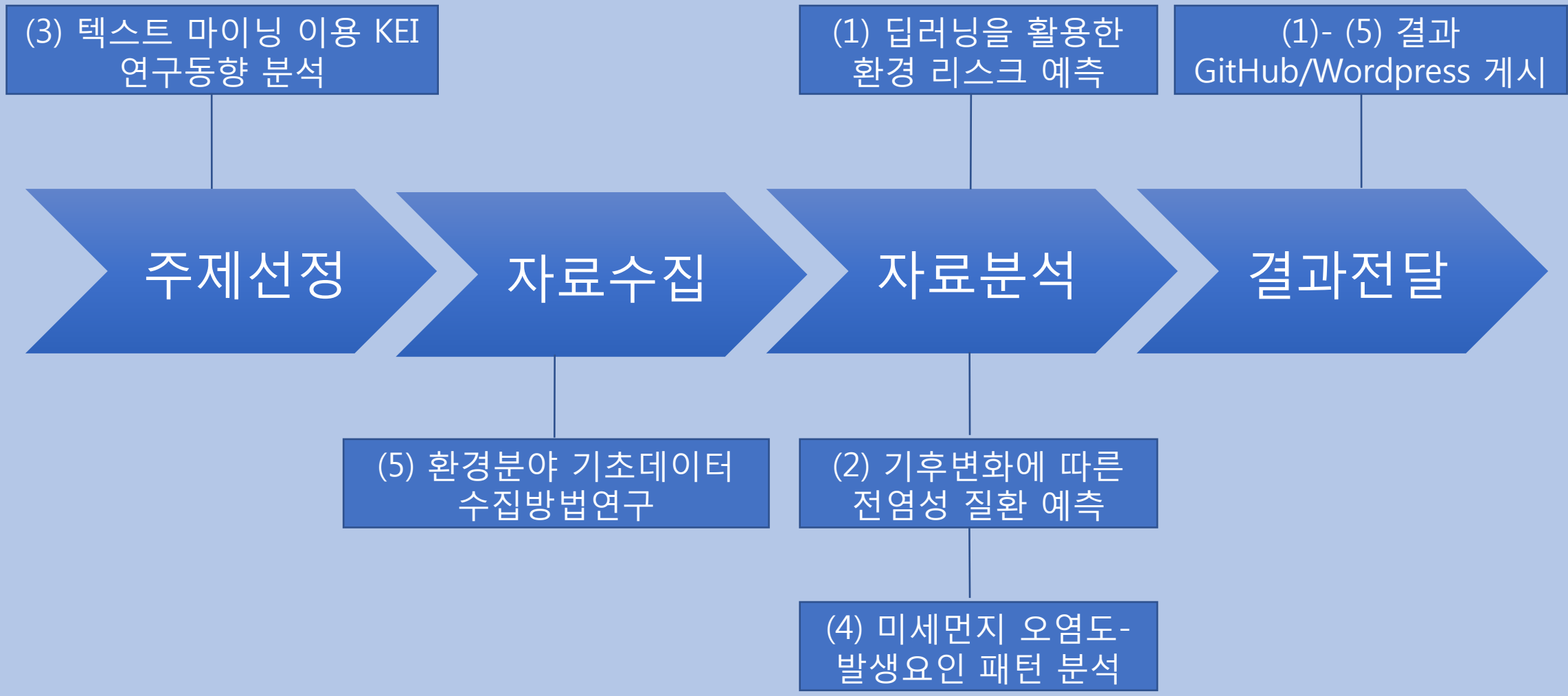
2017년: 환경위험 예측 방법론 개발

1. 환경 빅데이터 연구: 환경오염 예측 알고리듬 개발 및 학습 수준 심화
    - 전산화가 된 자료를 이용한 빅데이터 분석에 집중: 사례 개발 및 역량 축적에 중점
    - 환경오염 예측 딥러닝 알고리듬 개발 : 오염 예측의 시간-공간 해상도 제고
    - 주제 발굴, 패턴 분석, 원인 규명 등 실험적 연구 지속 추진
        - 주제 발굴: 자연언어 분석기법을 활용한 KEI연구보고서 분석
        - 패턴 분석: 기후자료-건강보험 자료 패턴 분석
        - 원인 규명: 미세먼지 발생 요인과 오염도 간 관계 규명

2. 환경 빅데이터 인프라 구축: 원내외 환경관련 자료 수집-추출 사례 축적
    - 환경 빅데이터 연구 자료 및 알고리듬 공개
    - 산재된 환경 관련 자료를 수집–추출하는 사례를 축적하여 오픈 소스로 공개

3. 원내외 빅데이터 서비스: 연구 정보 제공 서비스 개발
    - 연구 정보 추출 서비스 제공

세부과제 구성


(3) 텍스트 마이닝 이용 KEI 연구동향 분석
(1) 딥러닝을 활용한 환경 리스크 예측
(1)- (5) 결과 GitHub/Wordpress 게시
주제선정
자료수집
자료분석
결과전달
(5) 환경분야 기초데이터 수집방법연구
(2) 기후변화에 따른 전염성 질환 예측
(4) 미세먼지 오염도-발생요인 패턴 분석

착수자문회의 자문의견 반영결과

위 원	제 안 내 용	반 영 내 용 (미반영시 사유)
우석진	- 향후 환경정책에 대한 평가가 이루어져야 하기 때문에 정책수혜자, 수혜기업 DB가 구성되어 연결될 필요 있음	- 부처의 협조가 필요한 사안이므로 부처에 요청해 보겠음
이성호	- 인과관계 분석에서 tree 분석 이후 segment 별 effect 추정 / Counter factual experiment 필요	- 미세먼지 발생요인 분석 세부과제 분석과정에서 반영
김종률	- 빅데이터 분석 기법을 환경 분야에 적용하는 좋은 프로토콜을 만들어낼 수 있었으면 함 - 2025년까지의 장기적인 연구일정을 위한 전체 Roadmap 마련이 추가될 필요 있음	- 연차별 계획 보고서에 반영
강희찬	- 2017년에 가능한 부분이 보다 명확할 필요 있음 - 환경분야에 빅데이터를 적용한 사례가 많지 않기 때문에 이 부분을 보강할 방법 필요 - 각 주제별 방법론이 상이하므로 집중이 필요해 보임 - 연구인력 강화 필요 - 3년에 걸친 연구가 상호 어떻게 연결될지에 대한 고민 필요	- 2017년에는 전산화가 되어 있는 정형화된 자료 분석에 집중하고 2018년 이후에는 이미지, 문서 등 전산화가 필요한 자료로 분석 대상 확대 예정 - 환경분야에는 사례가 적지만 환경분야에서 주로 취급하는 시간-공간 시계열 자료에 대한 분석은 활발하므로 이를 원용할 계획. 2부의 연구방법론 에세이에 반영하겠음 - 딥러닝 기반 환경오염 예측 과제를 중심으로 진행 - 부연구위원 이상 급 채용 진행 중 - 1-3년차는 활발하게 사용되는 DNN, CNN, RNN 방법론을 사용하여 매체별 오염도 예측 알고리듬 구축을 목표로 진행 - 3단계: 첫단계에서는 방법론 확립. 2단계에서는 1단계의 방법론을 활용한 연구 인프라 구축, 3단계에서는 1-2단계의 연구성과를 활용하는 서비스 개발에 중점

착수자문회의 자문의견 반영결과

위 원	제 안 내 용	반 영 내 용(미반영시 사유)
명수정	- Platform 구성 시 내부사용자의 의견을 반영할 수 있도록 해야 함 - 빅데이터의 한계와 현실적 기대치를 반영할 수 있는 에세이 결과물이 나왔으면 함 - 활용의 성공사례 외에도 실패사례도 공유되었으면 함 - 법.제도적인 어려움을 개선하기 위한 제안이 있었으면 함 - 다차년도 연구인만큼 연구범위에 대한 로드맵이 있었으면 함 - 사례 분석에 있어 분야별 전문가 의견이나 자문이 필요하므로 평상시에도 의견을 듣고 도움을 받는 것이 좋을 듯함	- 내부 사용자 대상 설문조사를 연구 내용에 추가하는 방향을 검토 - 연구 시행 결과를 가감없이 전달하는 방향으로 보고서 서술 - 상동 - 보고서 2부에 반영 - 연차계획 보완하겠음 - 매월 성과 공유 결과를 온라인으로 공유할 예정
배현주	- 필요한 자료와 형태에 대한 수요조사 필요 - 기후변화와 전염병에 대한 연구가 기상청과 질병관리본부에서 진행되고 있으니 패턴 분석에 방점을 두어 차별화하는 것이 필요 - 미세먼지는 측정소의 특성에 따라서 변동성이 크게 달라지는 특징이 있으니 연구에 이러한 특성을 반영하는 것이 필요함 - 연구내용이 방대하므로 선택과 집중을 하는 것이 바람직함	- 내부사용자 대상 설문조사를 연구 내용에 추가하는 방향을 검토 - 기후변화-전염병 패턴 분석 과제에 반영 - 측정소의 특성을 사람이 분류하는 것보다 딥러닝 기법을 통해 데이터의 특성(feature)을 자동적으로 탐지할 수 있도록 하는 것이 더 좋을 수도 있음 - 딥러닝 기반 환경오염 예측 과제를 중심으로 진행
이명진	- Risk 관리가 필요. 텍스트마이닝의 경우 영어에는 잘 작동하지만 한글을 처리하는 데에는 문제가 있을 수 있음 - Random forest를 적용함에 있어 Bagging, Pruning, Bootstrapping 등 중간 단계에 거쳐야 할 작업들이 존재 - Dependent Variable이 명확한 분류가 되어 있을 경우에 분석이 용이 - 연구진 강화 필요 - 논문 등 학술 기여에 대한 내용 추가가 필요함	- 최근 한글처리기법 발달 성과 (R의 KONLP package, Python의 KONLPy package)를 적극적으로 수용하겠음 - 미세먼지 원인분석 과제 진행 시 필요한 중간과정에 대한 결과 수록 - 미세먼지과제는 오염도를 종속변수로 활용할 예정이며 기후변화-전염병 패턴 분석 과제는 5개 전염병과 관련된 지표를 활용할 예정. 프로포절 세미나를 통해서 확정하도록 하겠음 - 부연구위원급 이상 채용작업 진행 중 - 성과에 따라 논문 발신이 가능한 내용은 논문으로 발전시킬 계획 (위험부담이 높은 과제이므로 구체적 성과지표는 명시하지 않음)

연구진행 상황

중간보고
기대치 (최저)

중간보고
기대치 (최고)

세부과제	주제선정	자료수집	자료 전처리	예비분석	최종분석	결론도출
빅데이터 연구 방법론 활용방안	O	△	△	△		
딥러닝을 활용한 환경리스크 예측	O	O	O	O		
기후변화에 따른 전염성 질병 예측	O	O	O	△		
텍스트 마이닝 이용 KEI 연구동향 분석	O	O	O	O		
미세먼지 오염도-발생요인 패턴 분석	O	O	O	△		
환경분야 빅데이터 수집 방법론	O	△	△	△		

보고서 목차 및 작업계획

중간보고

장	절	3월	4월	5월	6월	7월	8월	9월	10월	11월	12월
1. 서론	1) 필요성 및 연구 목적										
	2) 선행연구										
	3) 연구내용 및 방법론										
	4) 본문 내용										
2. 환경연구와 빅데이터	빅데이터 연구 방법론 활용방안 (강성원)										
3. 환경 빅데이터 연구	1) 딥러닝을 활용한 환경리스크 예측 (이동현)									후속	조치
	2) 기후변화에 따른 전염성 질병 예측 (강선아)										
	3) 텍스트 마이닝 이용 KEI 연구동향 분석 (김도연)										
	4) 미세먼지 오염도-발생요인 패턴 분석 (김진형)										
	5) 환경분야 빅데이터 수집 방법론(한국진)										
4. 요약 및 시사점	1) 연구결과										
	2) 시사점										

2. 세부과제별 연구진행상황

2. 세부과제별 연구진행상황

2장 빅데이터 연구방법론 활용방안

빅데이터 연구방법론 활용방안

- 환경정책연구 방법론과 빅데이터 분석기법의 특징을 파악하여 적용방안 도출
  1. 환경정책연구: 환경정책을 유형화 및 유형별 관련연구 특성 파악
  2. 빅데이터 분석방법: 주요 분석기법 정리 및 장점 파악
  3. 환경정책연구 유형별, 단계별 빅데이터 분석기법 활용방안 도출
- 현재 진행상황: 환경정책 유형별 관련연구 방법론 특징 추출
  - 환경정책 유형화 : 예산서, 환경백서
  - 환경정책 유형별 유관연구 파악 : 2016년 KEI 보고서 (온라인 등재)
  - 유관 연구 유형별 방법론 분포 파악: 정량적 방법론 비중 파악
    - 정량적 방법론 사용 목적 파악 : 빅데이터 분석 방법 적용 여부 점검 시작

환경정책 유형별 관련연구 특징 추출

- 환경정책 유형화 : 부문 및 기능
  - 부문: 예산서 '관' 항목을 사용하되 '환경일반'을 세분화
    - '관' 항목: '상하수도', '수질', '폐기물', '대기', '자연환경', '환경일반'
    - '상하수도'와 '수질'은 '상하수도-수질'로 통합, '자연환경'에서 환경영향평가를 구분
    - '환경일반': '화학물질', '환경보건', '환경산업(기술, 경제)', '국제협력' '기타'

  - 기능: 환경오염물질 발생단계에 따라 정책기능 분류
    - 억제(Control): 배출원의 환경오염물질 배출을 억제하는 기능
      - 환경규제/환경관련 부담금
    - 처리(Treatment): 이미 배출된 오염물질의 영향을 저감하는 기능
      - 오염물질 처리 설비(환경기초시설) 및 시설(하수종말처리장, 폐기물매립장) 설치/환경오염 피해 보상
    - 대비(Preparation): 배출 이전에 배출량 저감 조치 유도 혹은 배출 정보 제공
      - 환경영향평가/환경오염예보
    - 환경조성(Support): 여타 3가지 기능이 원활하게 수행될 수 있는 여건 조성

환경연구(2016 환경정책·평가연구원 보고서)

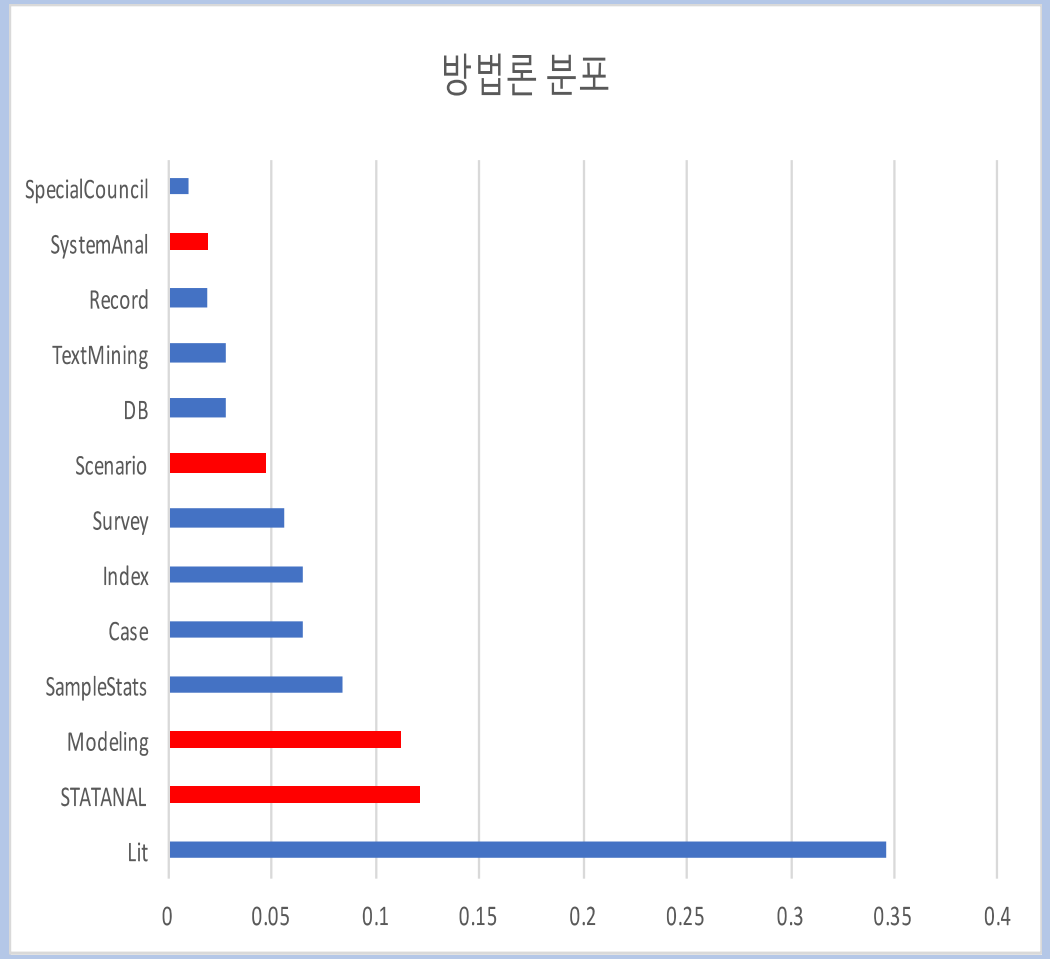
		억제 (규제. 조세)	처리 (시설, 복원, 피해보상)	대비 (예방, 영향평가)	환경조성 (계획.교육,법제도)
상하수도,수질		2	9	3	
폐기물			3		2
대기(기후)		11	8	1	3
자연환경(영향평가)		6	6	3	2
영향평가				7	1
환경 일반	화학물질	2			1
	환경보건				3
	환경산업				5
	국제협력				12
	기타	2		5	10

환경정책연구 방법론 분류기준

명칭	내용
STATANAL*	통계적 분석 방식을 사용하여 변수간의 관계를 파악하거나 관심변수의 값을 추정하는 방법 - 결정요인분석, CVM, Conjoint, 메타연구 등 중간단계에서 계량경제학적 분석방식을 요하는 방법론 포함 - 클러스터 분석, 주성분 분석 등 전통적인 회귀식을 이용하지 않는 통계적 추론 방법론 포함
Modeling*	연역적 추론에 기반한 모형을 구축하고 이를 이용하여 관심 변수의 값을 구하는 방법 (일반균형, 산업연관분석)
SystemAnal*	사전적 모형 없이 직관적인 인과관계 네트워크 시스템 모형을 구축하여 분석하는 방법
Scenario*	파라미터 값이 상이한 시나리오를 구축하고 관심 변수의 값을 시나리오에 따라 구하는 방법 (경제성, 수익성..)
SampleStats	특정한 방법론 없이 기초자료로부터 표본통계량을 조합하여 논거를 찾아내는 방식의 연구를 의미
Index	기초통계량으로부터 관심대상 현상을 대표하는 지표를 도출하는 방식
DB	DB구축
Lit	주제와 관련된 선행연구 및 사례에 관련된 문헌을 종합하여 정리하고 시사점을 도출하는 방법 - 정량적 연구 중 결과물을 도출하지 않고 방법론을 정리한 연구도 문헌연구로 간주
Survey	설문조사(추가적인 분석 없이 설문조사 결과만 제시한 경우)
SpecialCouncil	전문가들로 구성된 패널의 의견을 종합하는 방법
Record	행사기록
Case	문헌조사 이외의 방법을 사용하여 사례를 조사하는 방법(인터뷰, 실측 등을 포함)

문헌조사가 방법론의 압도적 비중 차지

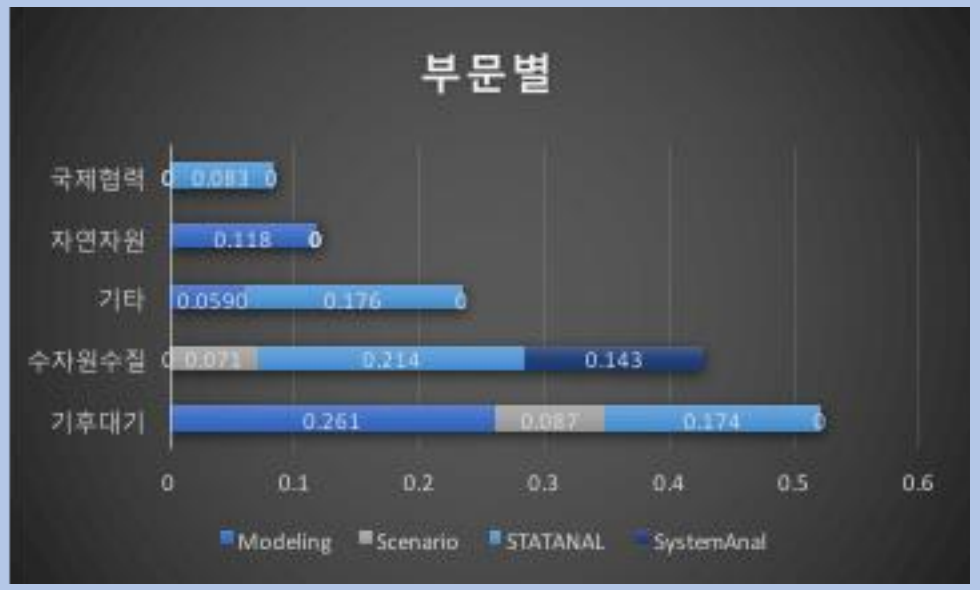
- Machine Learning 과 유사한 '정량적' 연구는 30.8%
  - STATANAL (12.1%)
  - Modeling (11.2%)
  - Scenario (5%)
  - System Anal (1.9%)

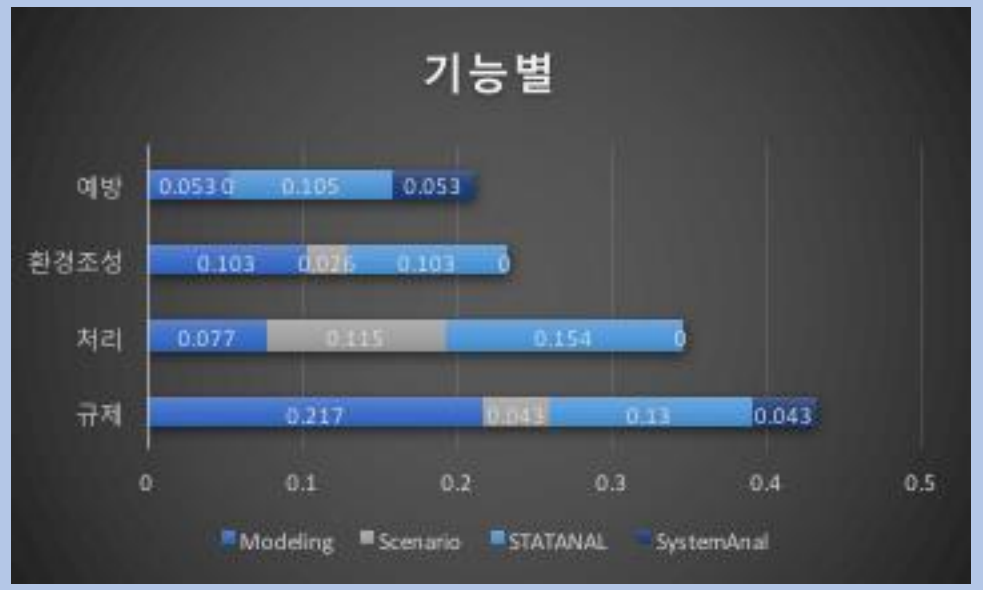


기후대기, 규제: 정량연구 비중이 높음

- 부문: 기후대기(52.2%), 상하수도 수질(42.8%) 정량연구 비중 높음
  - 전체 연구 건수가 5건 이하인 부문은 제외
- 기능: 사전적 규제 (34.7%) 정량연구 비중이 가장 높음

부문별, 기능별 정량적 연구 비중



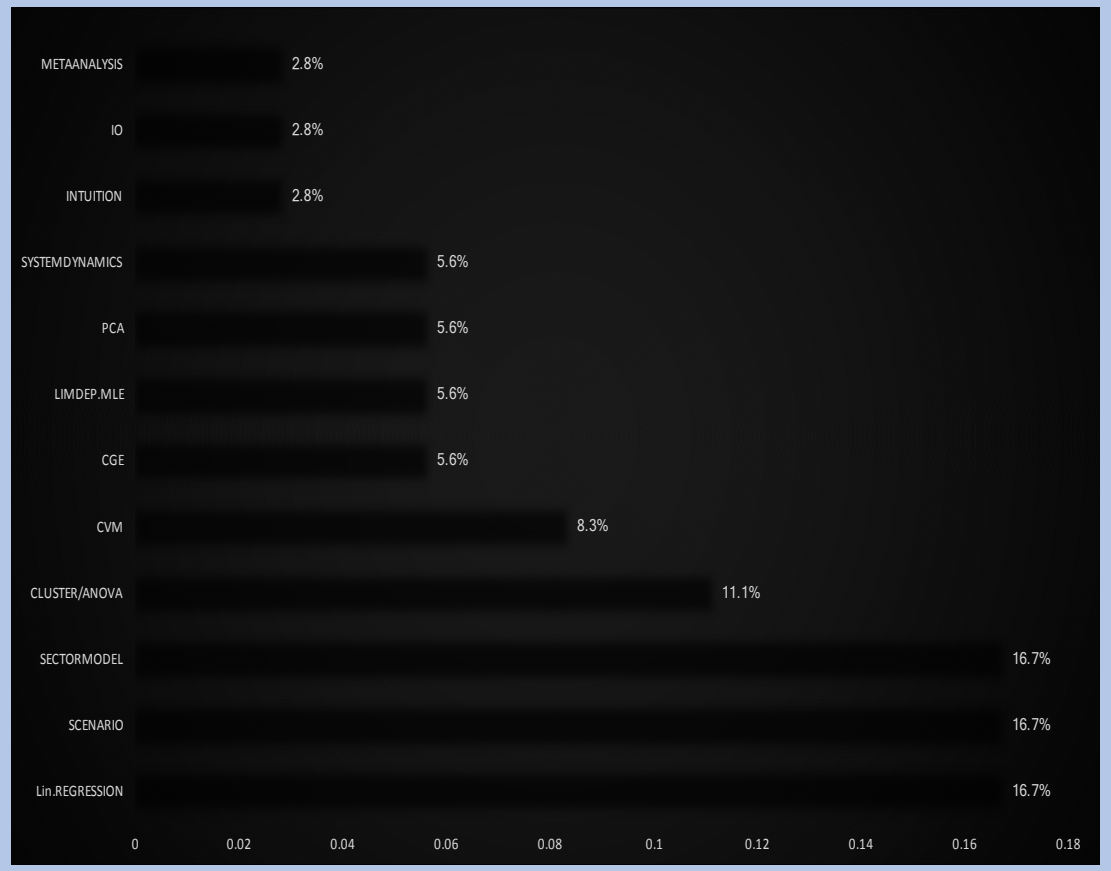


정량연구는 예측 목적으로 주로 수행

- 정량연구 방법론: 선형회귀, 시나리오분석, 부문 특화 모형 활용 활발 (16.7%)

- 정량연구 목적: 예측 (58.3%) > 인과 (27.8%) > 상관관계 파악(13.9%)
  - 예측: 관심변수의 미래 값 추정
  - 인과: 특정 변인에 따른 관심변수 변화
    - 경험적 인과관계(how much) 중점
  - 상관관계: 관심 변수 간 관계 패턴

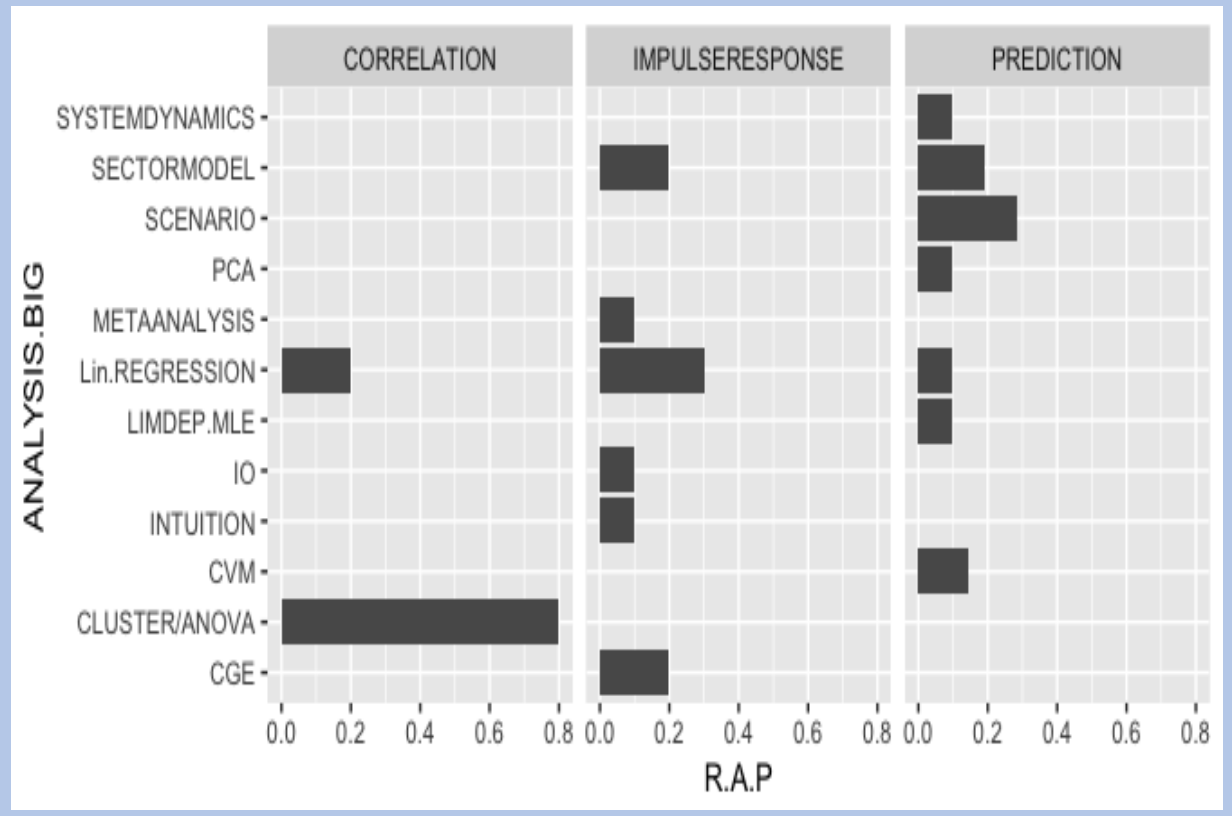
정량분석 세부 방법론 분포



빅데이터 분석기법 : 예측, 상관 분석 개선

- 예측 연구의 시나리오 분석 의존(26.8%) 완화
  - 다양한 변수의 영향을 함께 고려할 정량적 방법론이 부재한 현실 타개
  - 모형(19%), 회귀(19%), 시스템분석(9.5%) 대안
    - 데이터에 최적화된 예측방식 (Deep Learning)
- 상관관계 분석 방식 다변화
  - 군집분석(80%)이 대부분인 상관관계 분석 대안 제시: SVM, k-mean cluster
- 인과관계 연구의 대안 제시
  - 선험적 연관관계 의존 (CGE 20%, IO 10%, 모형 20%, 선형회귀 30%) 완화
    - 데이터 설명에 최적화된 변수 간 관계 파악 (Regression Tree)

정량분석 세부 방법론 분포 - 연구목적 별



빅데이터 분석기법 환경정책연구 적용

방법론	상관관계	인과관계	예측
CGE	-	Regression Tree, Counter factual	-
CLUSTER/ANOVA	Cluster, Dimension Reduction, Association Rule	-	-
CVM	-	-	14.3%
INTUITION	-	10%	-
IO	-	Regression Tree, Counter factual	-
LIMDEP.MLE	-	-	Deep Learning
Lin.REGRESSION	Cluster, Dimension Reduction, Association Rule	Regression Tree, Counter factual	Deep Learning
METAANALYSIS	-	10%	-
PCA	-	-	Deep Learning
SCENARIO	-	-	Deep Learning
SECTORMODEL	-	Regression Tree, Counter factual	Deep Learning
SYSTEMDYNAMICS	-	-	Deep Learning

향후 계획

- 환경연구 문헌 추가 수집: 학술논문
  - 기존 연구 문헌 자료 대표성 문제: 문헌연구 >> 정량적 연구
  - 예산으로 가중치, 기초연구 제외: 시도해 보았으나 결과 유지
- '정성적 연구'와 빅데이터 분석기법 간의 관련성 점검
  - Sample Stat, Index 연구 중 빅데이터 분석기법 활용 가능 연구 점검
    - 사용 목적에 대한 점검에서 시작
- 빅데이터 분석기법 (Machine Learning) 방법론 소개
  - 상관분석 : Unsupervised Learning
    - Cluster, Dimension Reduction(PCA, t-SNE), Association Rule
  - 예측 : Supervised Learning
    - Regression, Supporting Vector Mechanism, Decision Tree, Neural Network
  - 인과 : Decision Tree, Counter factual experiment.
    - 구조적 인과관계(why)는 파악하기 어려우나, 경험적 인과관계(how)는 파악 가능

2. 세부과제별 연구진행상황

3장. 환경 빅데이터 연구

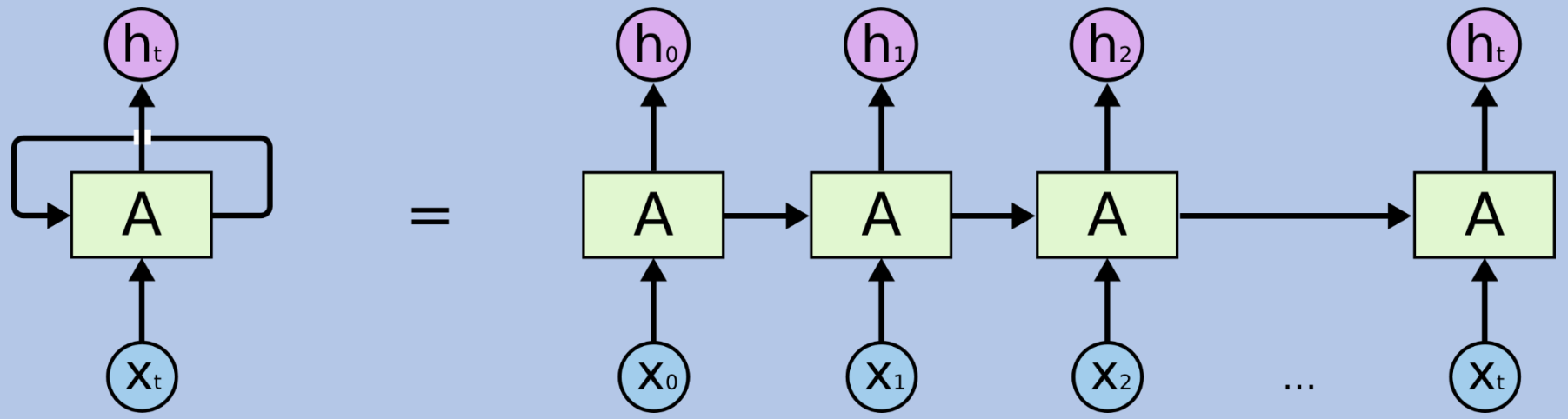
1. 딥러닝을 활용한 환경 리스크 예측 (이동현)

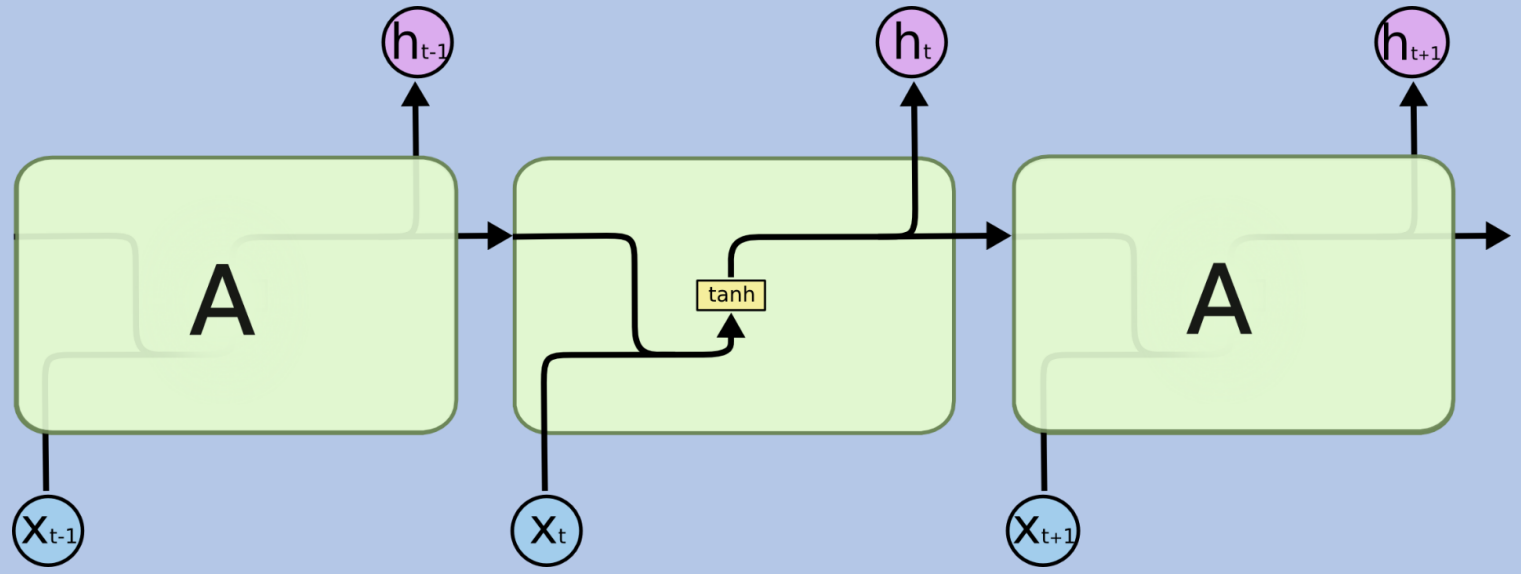
(1) 딥러닝을 활용한 환경 리스크 예측

- 딥러닝 기술을 활용하여 전국 측정소 단위에서 미세먼지의 시간당 오염도를 예측
  - 종속변수: AirKorea 미세먼지(PM_{10})
    - 2016년 1월 ~ 2016년 12월 서울시 25개 도시/14개 도로변 관측소 측정 데이터
  - 독립변수
    - AirKorea 대기오염물질 (아황산가스, 일산화탄소, 오존, 이산화질소)
    - 기상청 방재기상관측 데이터 (시간당 기온, 풍속, 풍향 등)
- 순환신경망(RNN), 단층 LSTM, 다층 LTSM 평균제곱오차: OLS 평균제곱오차(RMSE)의 45.1%-49.7%

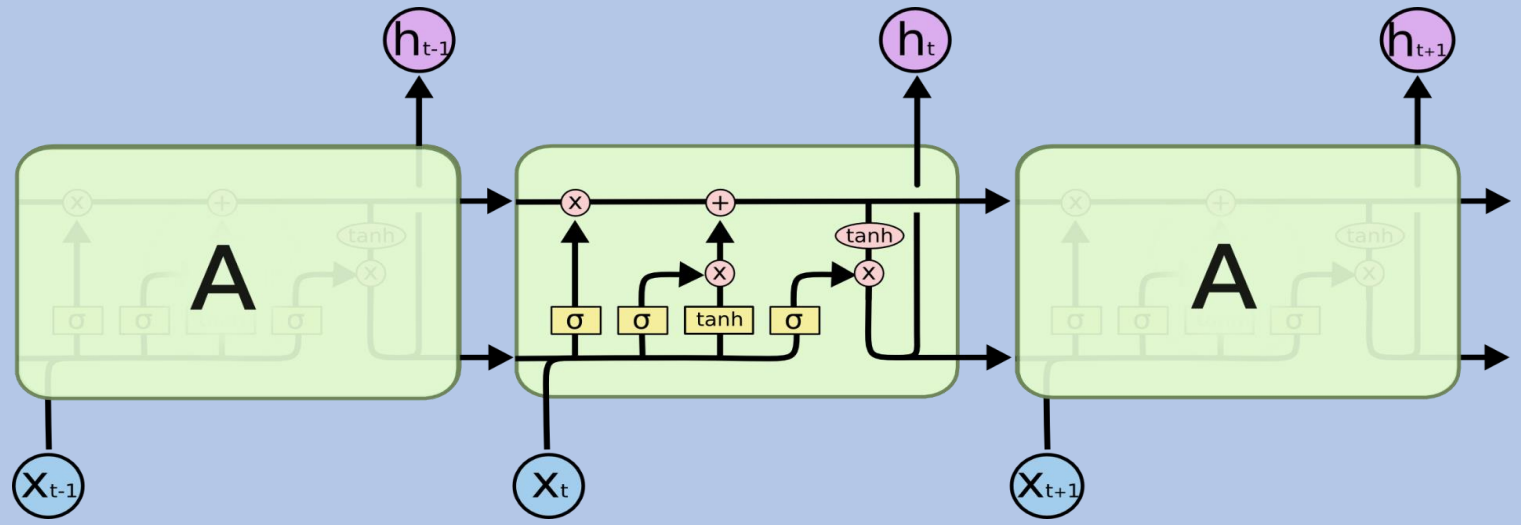
분석 모형: 순환신경망 (RNN: Recurrent Neural Network)

- 순환신경망(RNN)은 동일한 네트워크를 여러 개 복사한 것
- 각 네트워크는 자신의 뒤에 있는 네트워크에 정보 전달



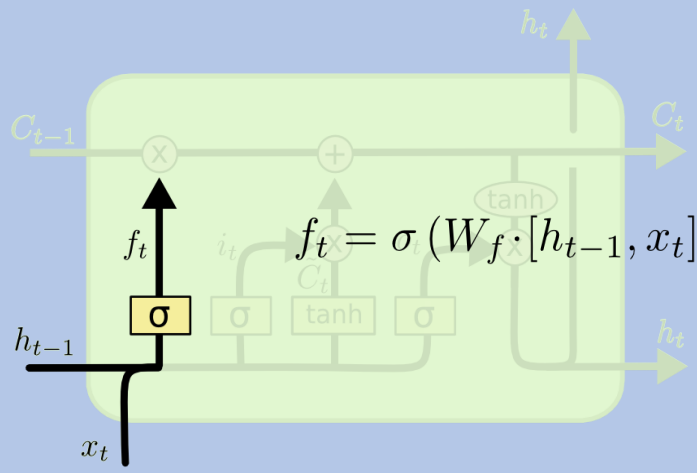


- 표준 RNN: 반복 모듈 하나에 하나의 층을 포함
  - 과거의 정보를 받아서 전달
  - Vanishing/Exploding gradient. 문제에 취약



- LSTM: 반복 모듈 하나에 네 개의 상호작용 층을 포함
  - Forget gate: 버리는 정보
  - Input gate: 갱신 정보
  - 메모리 갱신
  - 출력 gate : 전달할 정보

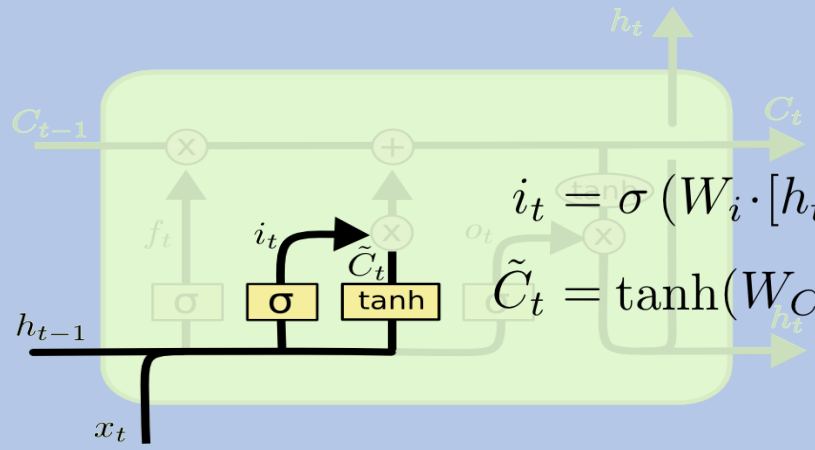
1



$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

- **버리는 정보 결정 : forgot gate**

2

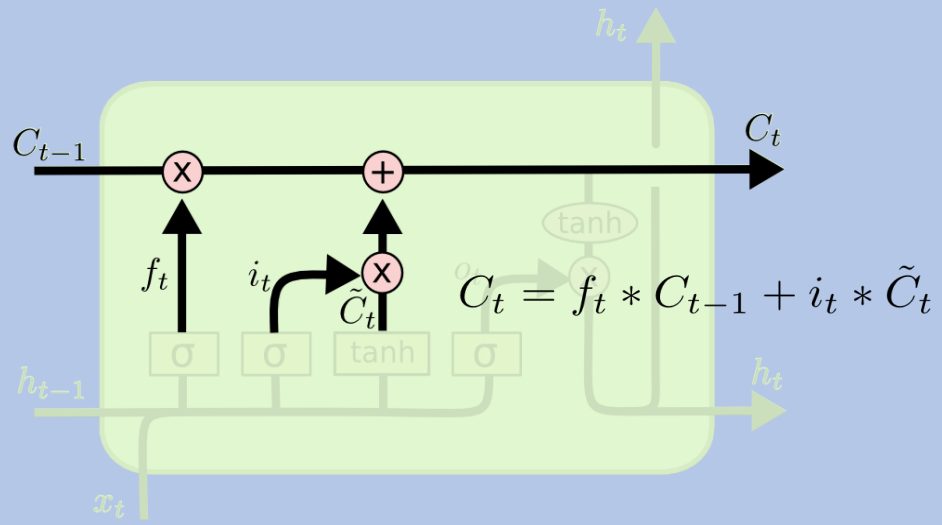


$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

- **새로운 정보를 셀 상태에 저장**
  - **1. input gate: 갱신 대상 확정**
  - **2. tanh : 갱신 정보 생성**

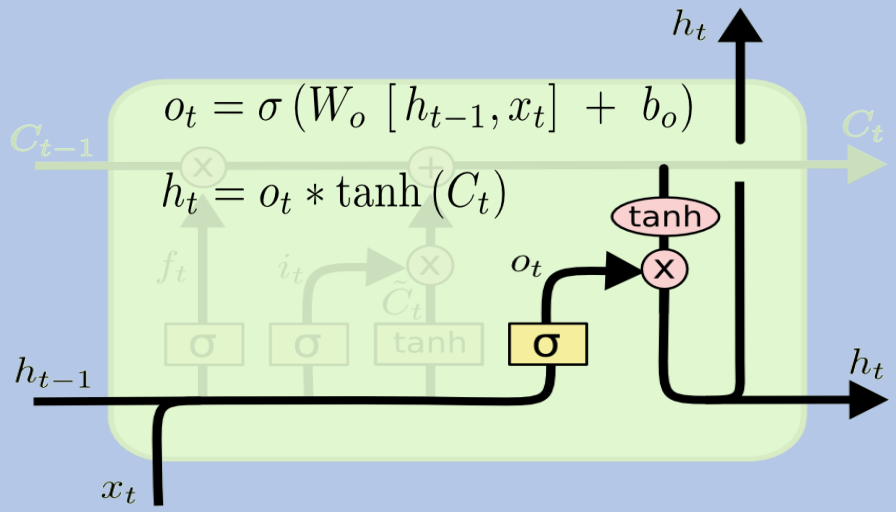
3



$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

- **이전 셀 상태를 새로운 셀 상태로 갱신**

4



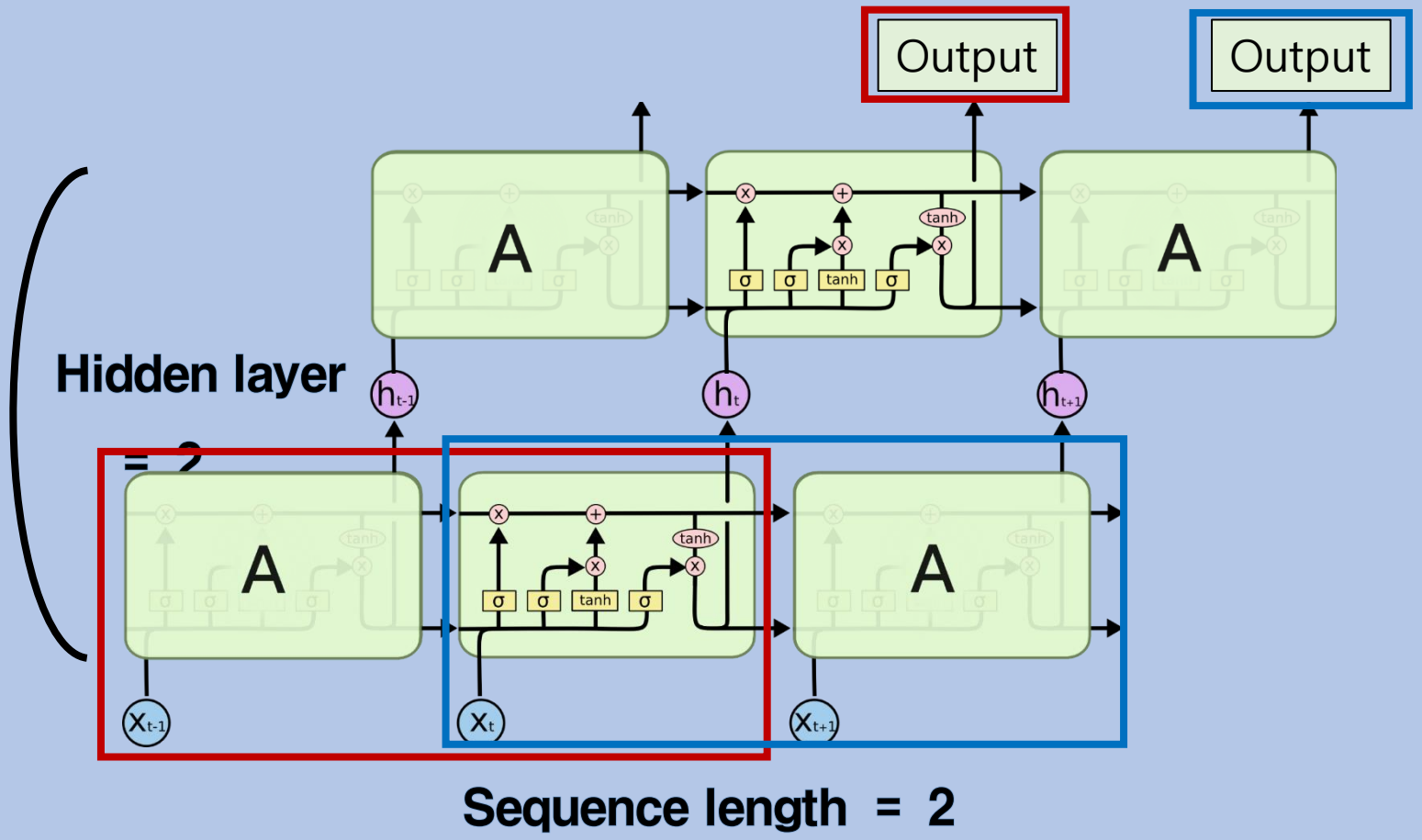
$$o_t = \sigma\left(W_o \left[h_{t-1}, x_t\right] \; + \; b_o\right)$$

$$h_t = o_t * \tanh\left(C_t\right)$$

- **무엇을 출력할지 결정**: output gate

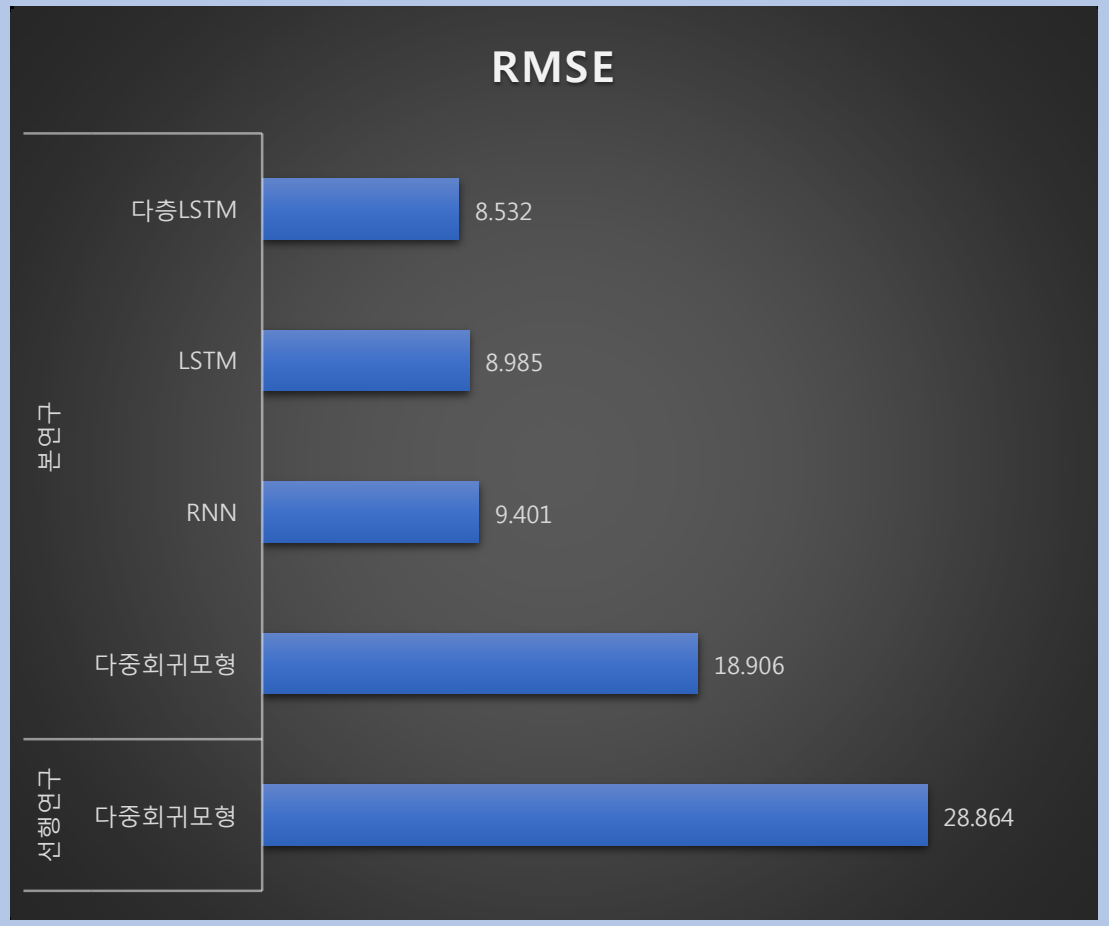
다층 LSTM: 2개 은닉층

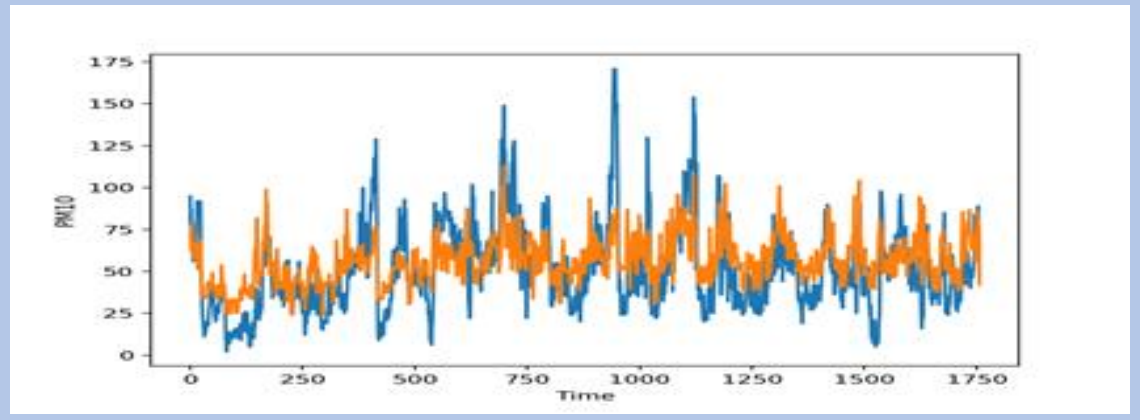
- **2시간 치의 데이터로 다음 시간 때의 미세먼지를 예측** (Many to one)

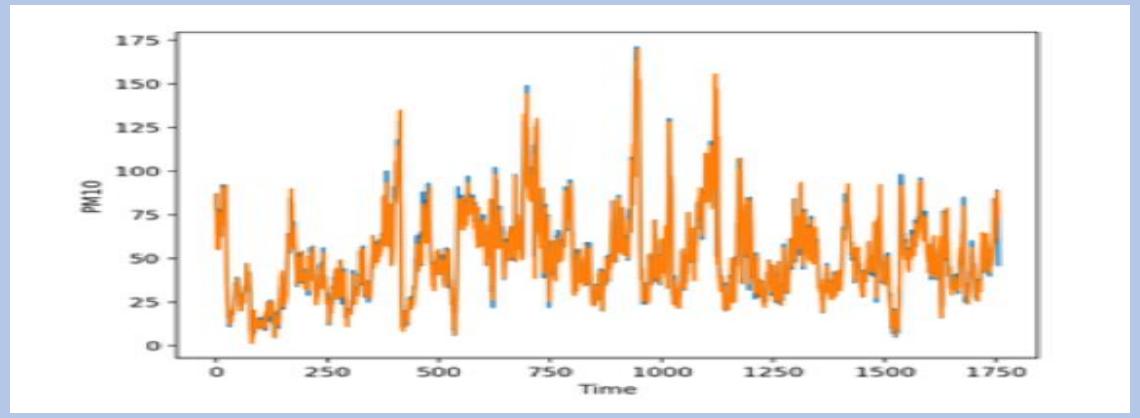


X_t
- 아황산가스
- 일산화탄소
- 오존
- 이산화질소
- 기온
- 풍속
- 풍향
- 미세먼지

분석결과: 다중회귀모형과 RNN

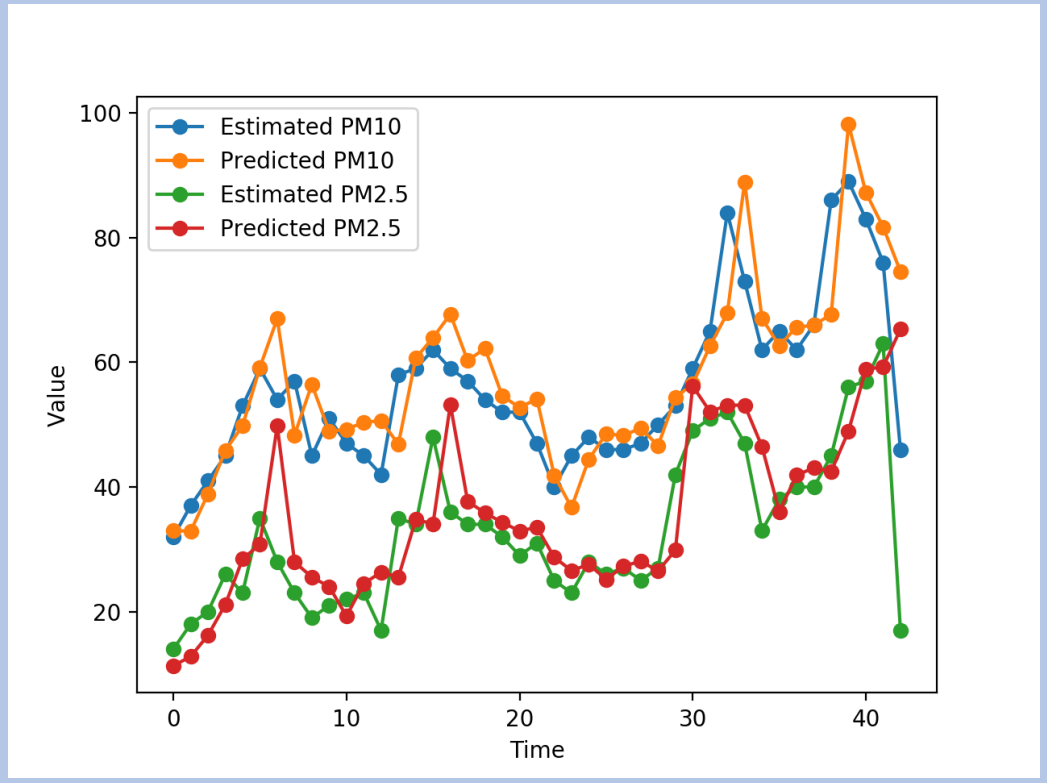




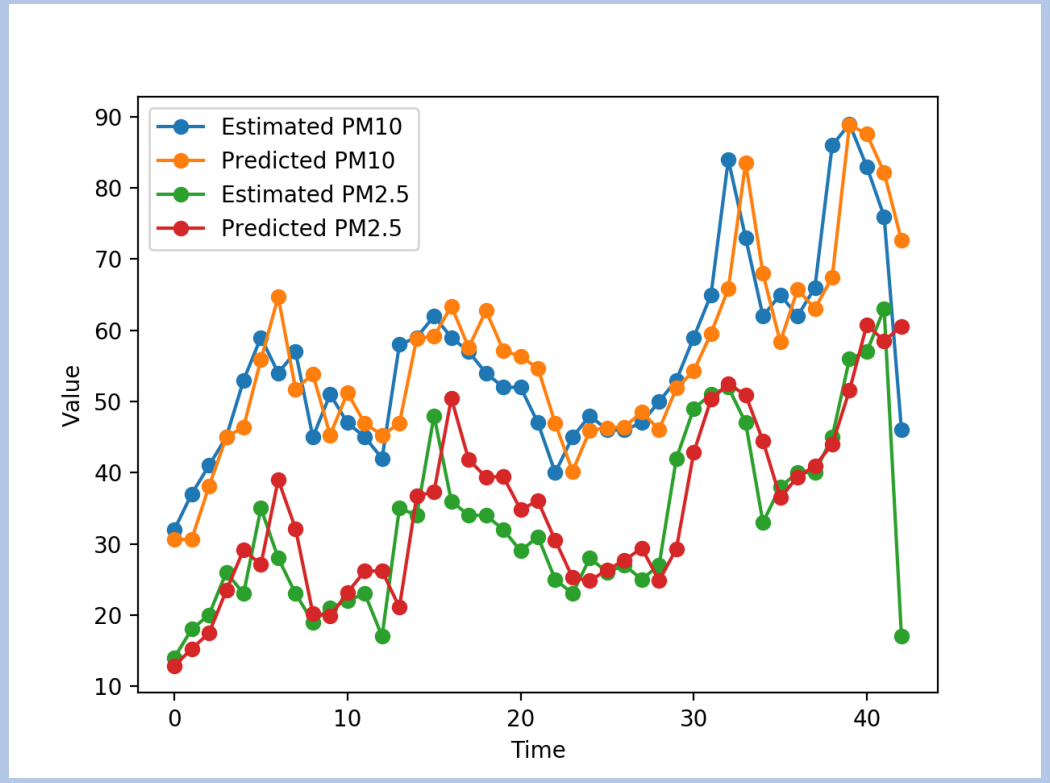


분석결과: LSTM 알고리듬 사용 예측성과

단층 LSTM



다층 LSTM



향후 계획

- 공간시계열의 특성을 반영하여 추정 결과 개선 지속
  - Lag Structure 는 반영하였으나 지역 간 dependency는 반영되지 않은 상태
  - 서울시 이외 타 지역에 대한 확장을 시도

- 기존의 계량 분석 방법론과 병행을 시도

2. 세부과제별 연구진행상황

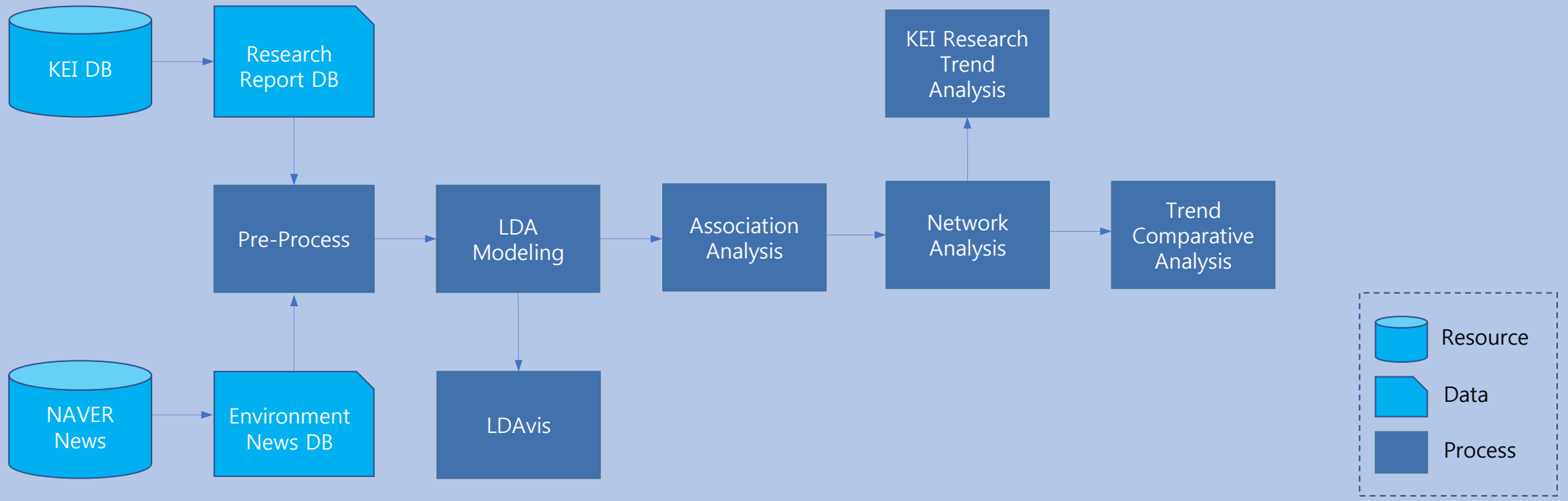
3장. 환경 빅데이터 연구

2. 텍스트마이닝을 이용한 KEI 연구동향분석 (김도연)

텍스트마이닝을 이용한 KEI 연구동향분석

- KEI 연구동향 분석
  - 24년 간(1993-2016) 축적된 1,697건의 KEI 연구보고서(제목, 목차, 요약, 날짜) 분석
- 연구공급 동향과 연구수요 동향 비교 분석
  - 연구공급 동향 파악 : 13년 간(2004-2016) 축적된 1,170건의 KEI 연구보고서(제목, 날짜) 분석
  - 연구수요 동향 파악 : 13년 간(2004-2016) 축적된 193,636건의 NAVER 환경뉴스(제목, 날짜, 출처) 분석
- 중간보고까지 진행사항
  - KEI 연구동향 분석 수행
  - 네이버 환경 뉴스기사 크롤링(약 19만개)
  - KEI, 네이버 전체 데이터 LDA 비교 분석

연구동향분석 작업 흐름도



Text data Pre-processing

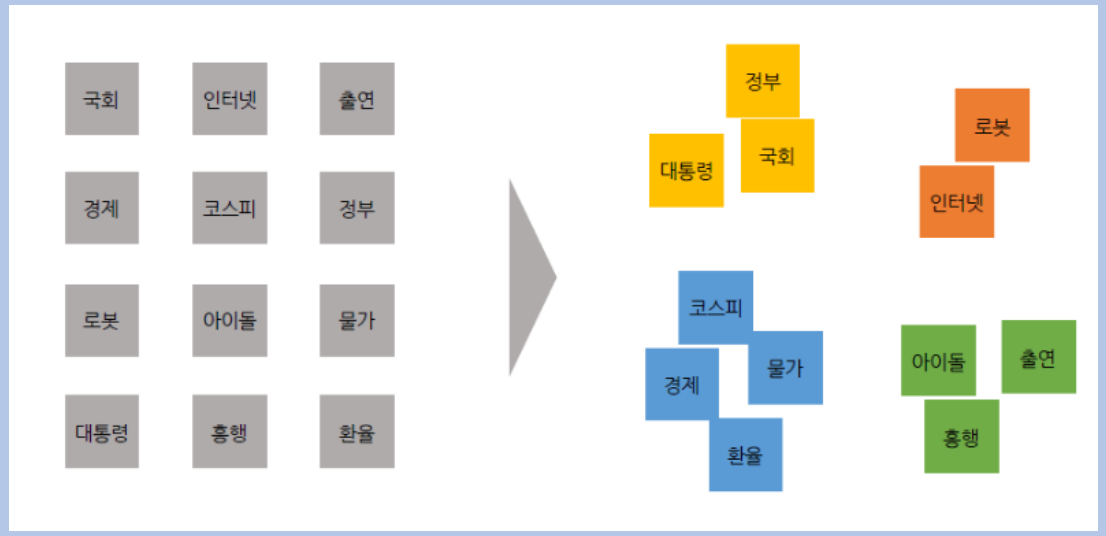
1. 형태소분석기 수행
   - R에서 제공하는 형태소분석기 패키지 (KoNLP, tctStart)를 사용하여 명사 추출
2. 특수문자, 특정단어 등 불용어 삭제
3. 단어길이 한글자 삭제
4. 출현빈도가 매우 낮은 단어(Sparse Terms) 삭제
5. Low TF-IDF 단어 삭제
6. DTM(Document Term Matrix)형태로 변형

   예) 오른쪽 표

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

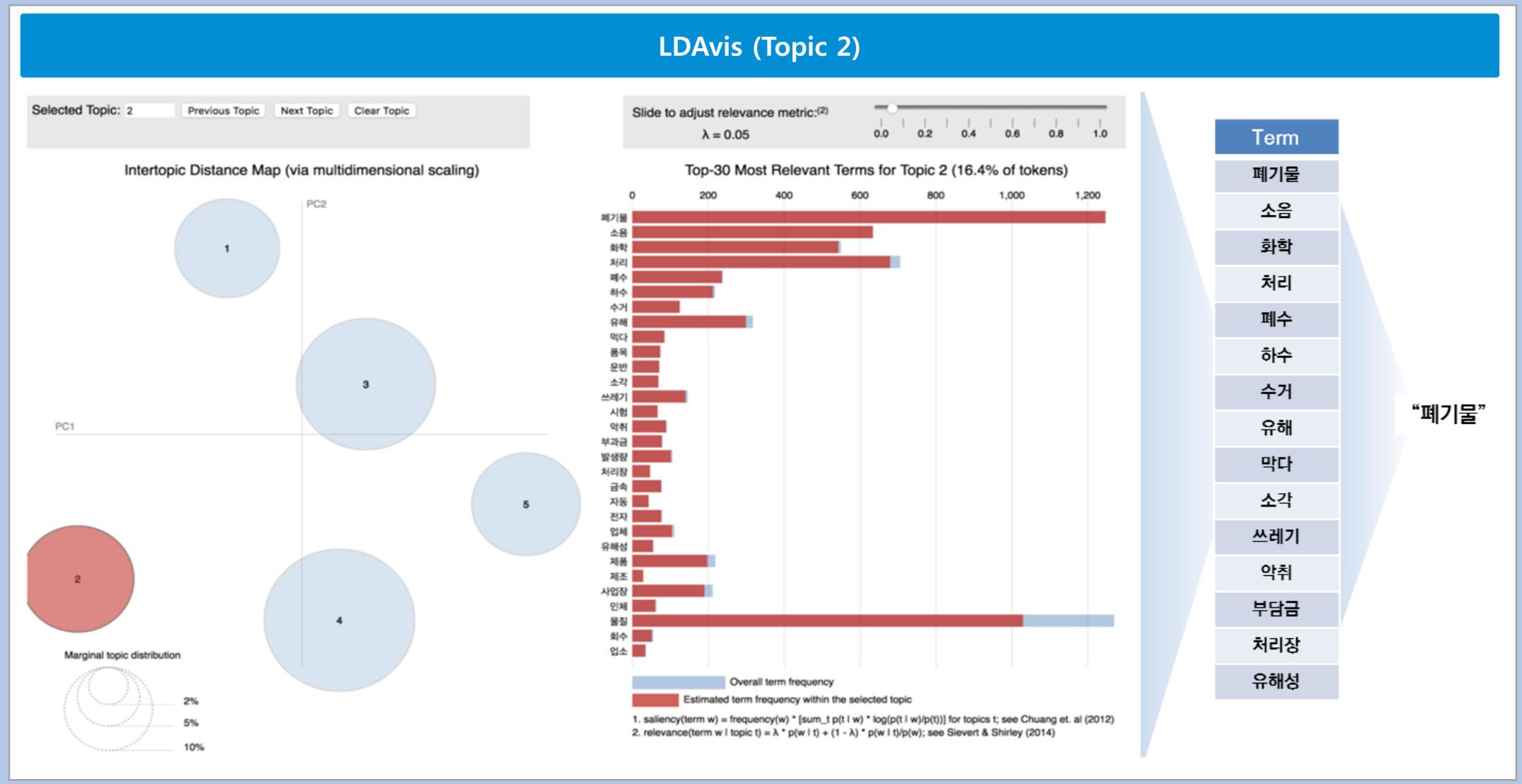
LDA 분석: 단어 분포 → 문서 주제

- 단어 분포에서 숨어있는 문서 주제 파악: Unsupervised Learning
  - 단어(word) 가 모이면 주제(topic), 주제가 모이면 문서(Document)
    - 문서[Document]← 토픽[Topic] ← 단어[Word]
  - 주제의 posterior 분포 도출 Bayesian Estimation
    - 단어의 likelihood 함수 + 주제의 prior 분포
  - 문서 별 개별 주제가 출현할 확률 중 가장 큰 값을 갖는 주제를 그 문서의 주제로 선정



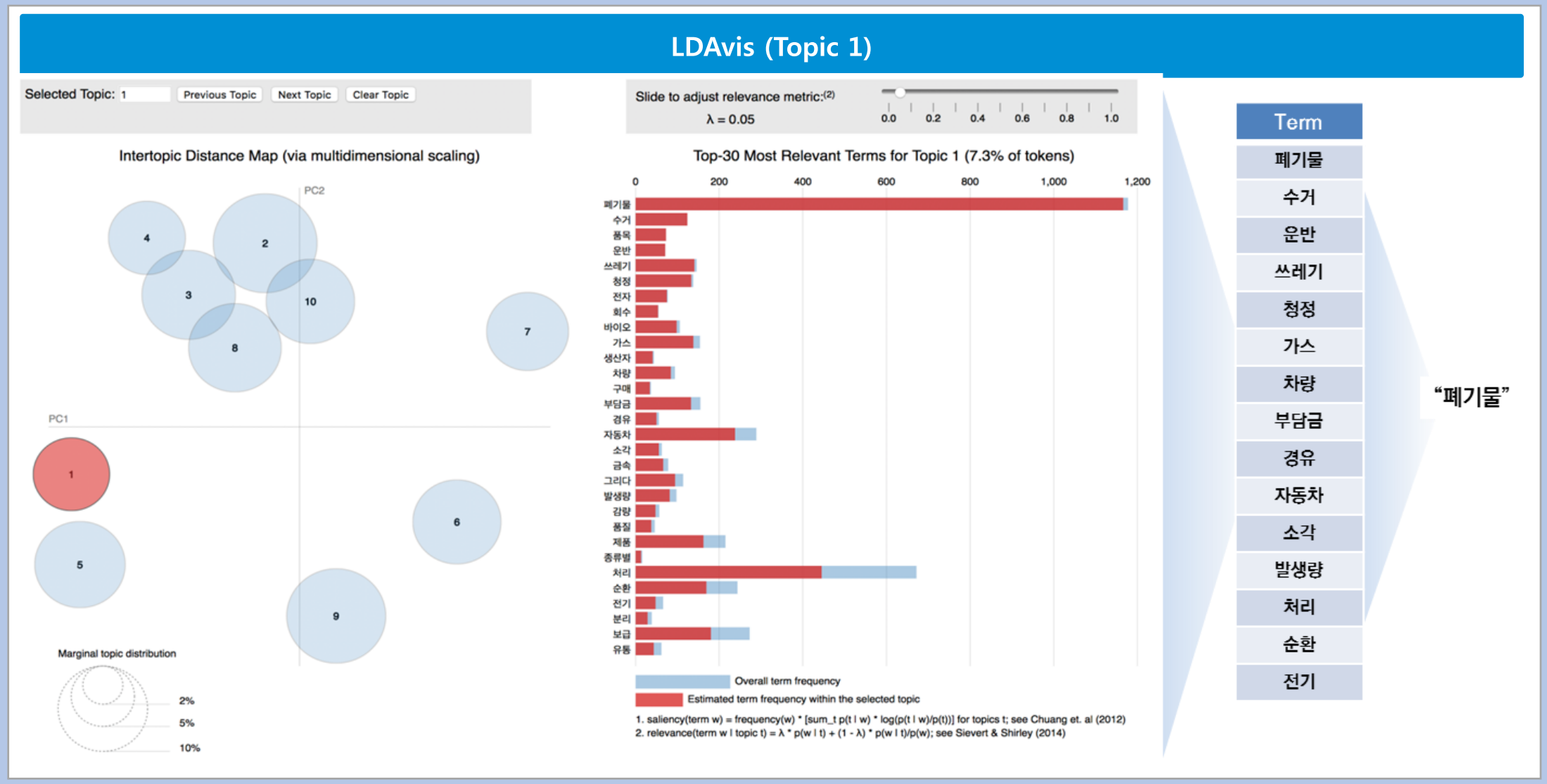
LDA 분석 결과: 도출된 주제와 연관어

예시1) **KEI 연구보고서**에서 추출한 토픽 2번: 폐기물



LDA 분석 결과: 도출된 주제와 연관어

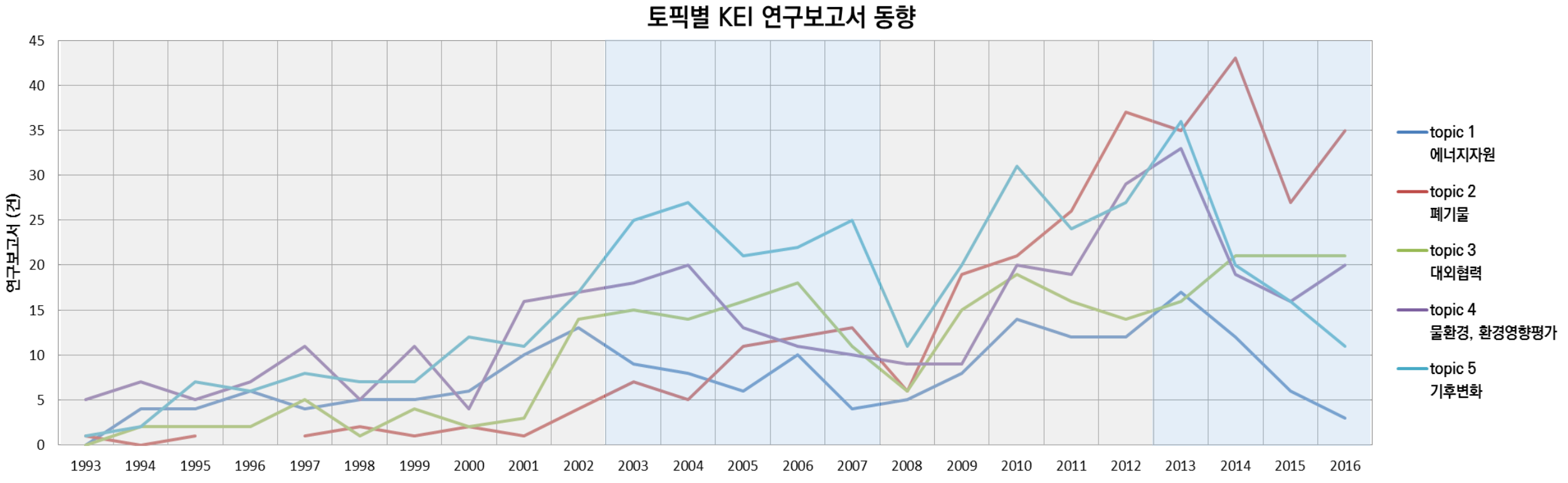
예시2) **네이버 환경뉴스**에서 추출한 토픽 1번: 폐기물



KEI 연구보고서 LDA 분석

- 1993년~2016년까지의 수집한 전체 KEI 연구보고서 LDA 분석 결과

: **5개 토픽**(에너지자원, 폐기물, 대외협력, 물환경/환경영향평가, 기후변화)으로 분류됨

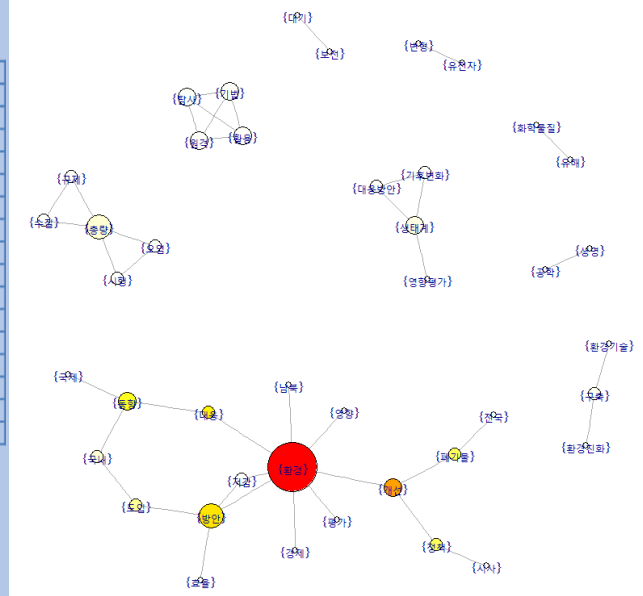


키워드 연관성 및 네트워크 분석

1993년 ~ 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

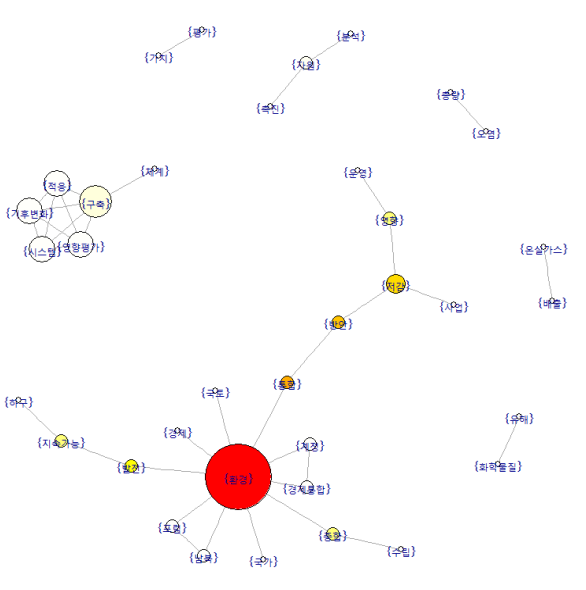
- 수질오염총량제 시행, 원격탐사기법 활용, 환경친화 기술, 유전자 변형, 전국 폐기물 개선 등의 연구가 활발했음.



2003년 ~ 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	0.0147	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

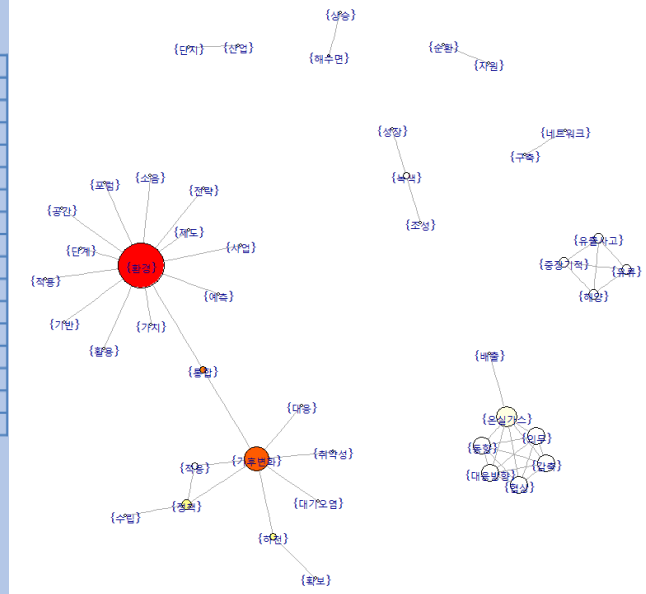
- 기후변화 영향평가 및 적응시스템 구축, 온실가스 배출, 환경경제통합계정 키워드가 새롭게 등장
- 전구간에 이어 유해화학물질, 남북 키워드 계속 등장.



2008년 ~ 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

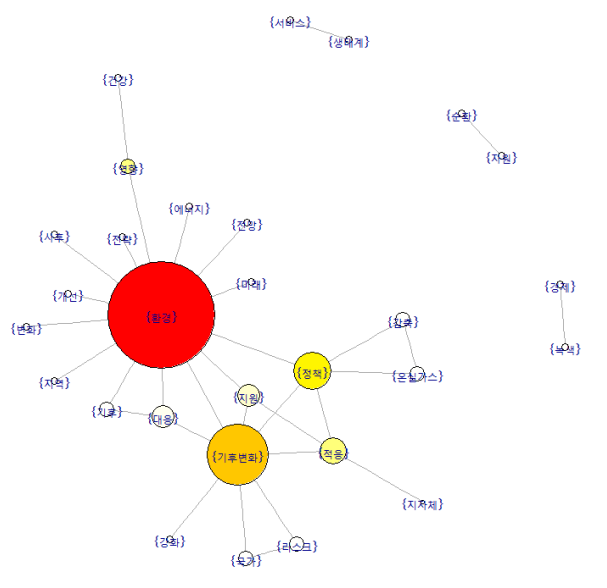
- 기후변화, 온실가스 키워드의 매개중심성 높아짐.
- 해양 유류 유출사고, 녹색성장 조성, 해수면 상승,



2013년 ~ 2016년

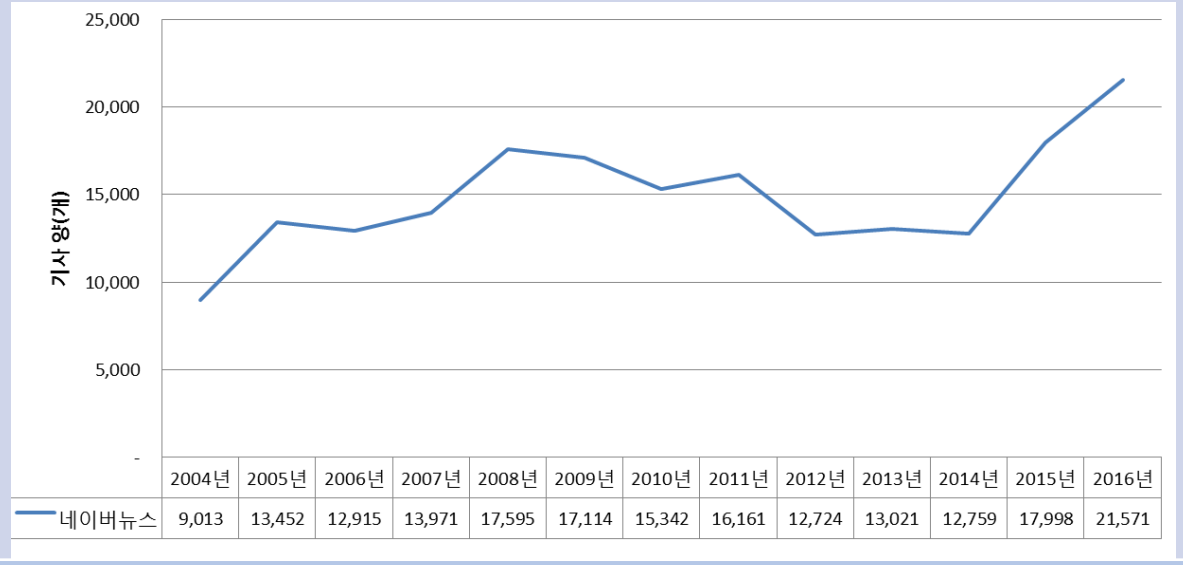
no	lhs		rhs	support	confidence	lift
1	지자체	=>	적응	0.0125	0.7143	8.4244
2	적응	=>	지자체	0.0125	0.1471	8.4244
3	감축	=>	온실가스	0.0175	0.8750	43.8594
4	온실가스	=>	감축	0.0175	0.8750	43.8594
5	감축	=>	정책	0.0125	0.6250	5.8285
6	정책	=>	감축	0.0125	0.1163	5.8285
7	온실가스	=>	정책	0.0125	0.6250	5.8285
8	정책	=>	온실가스	0.0125	0.1163	5.8285
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 매개중심성 높아짐.
- 환경 키워드와 관련하여 건강, 미래, 전망, 에너지 키워드가 새롭게 등장함.



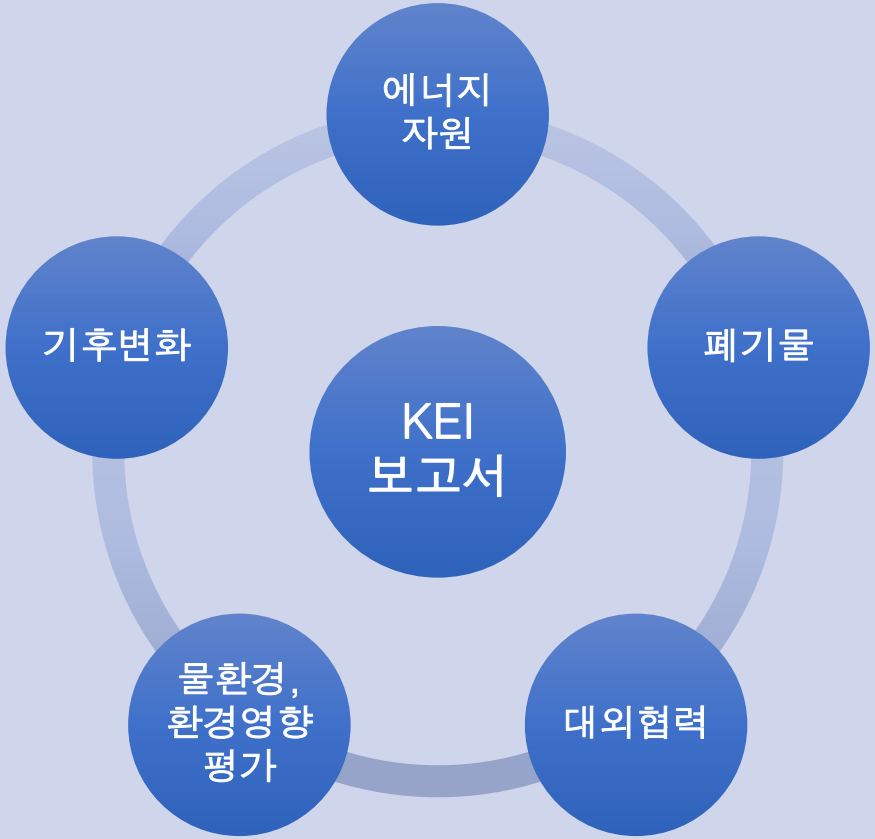
네이버 뉴스 기사 데이터 수집

구분	내용
수집 도구	JAVA Jsoup(html parser)
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	**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

LDA 결과 비교

KEI 연구보고서(1993~2016)	네이버 환경뉴스(2004~2016)
	

향후계획

- 네이버 환경뉴스에서 추출한 토픽 점검 후 토픽 시계열 분석
  - LDA분석을 통해 추출한 10개 토픽 점검 및 축소
    예) 폐기물, 오염원과 해양/풍력, 물환경을 합치는 방법 등을 고려

- 네이버 환경뉴스 키워드 연관성 및 네트워크 분석
  - 2004년부터 2016년까지 수집한 네이버 뉴스 기사(19만건) 데이터 사용
  - 1시기(2004년~2007년), 2시기(2008년~2012년), 3시기(2013년~2016년)로 분류하여 총 3개의 시기 분석

- 연구공급 동향과 연구수요 동향 비교 분석
  - KEI 연구보고서와 네이버 환경뉴스에서 시기별로 각각 추출한 키워드를 시계열로 파악하고 두 시계열을 비교 분석
  - 시기별 환경정책 동향과 KEI 연구 동향의 미스매치 파악 및 시사점 도출

- 분석 알고리듬 code 및 Data를 Open source로 공개

2. 세부과제별 연구진행상황

3장. 환경 빅데이터 연구

3. 기후변화에 따른 전염성 질환 예측 (강선아)

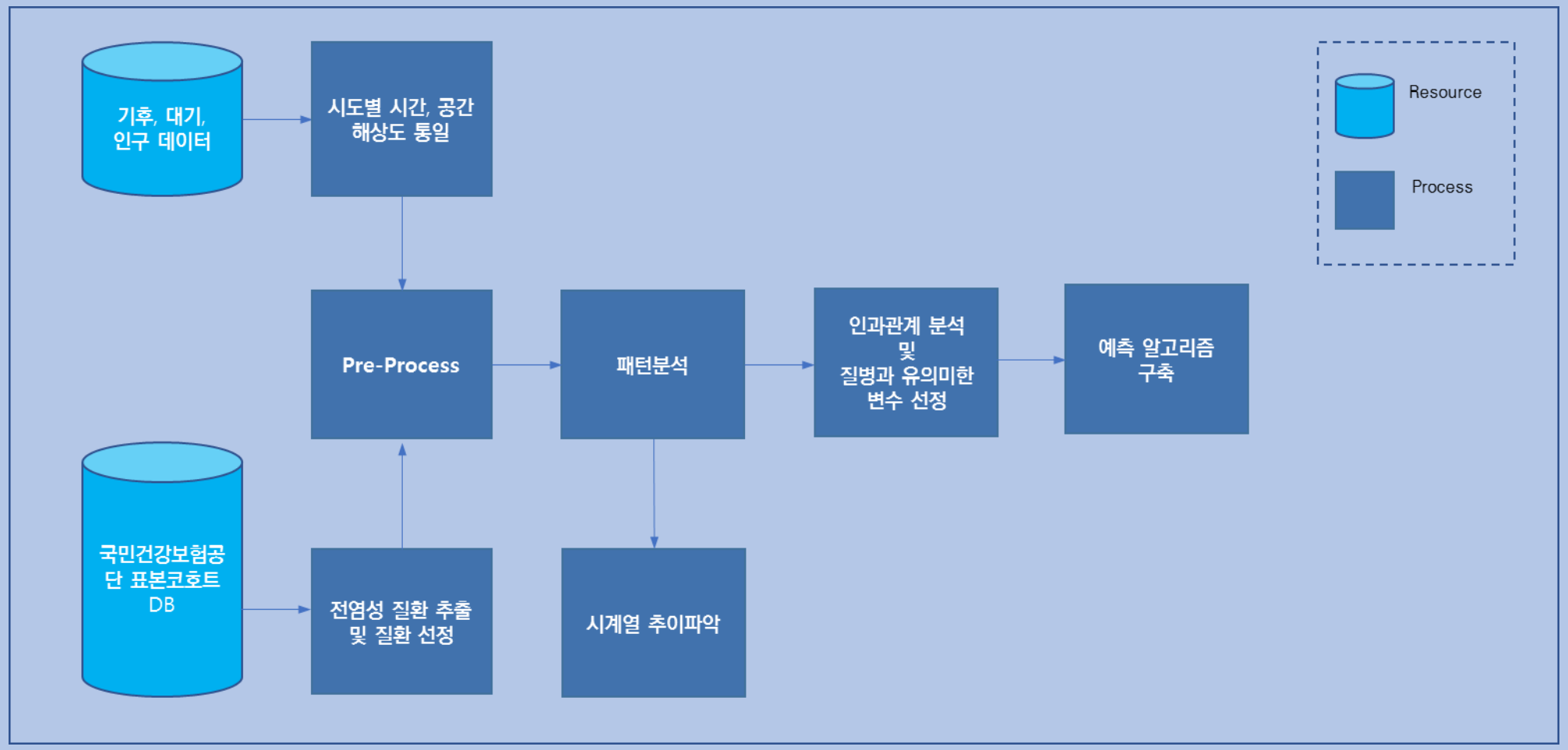
기후변화에 따른 전염성 질환 예측

- 연구내용
  - 2009~2013년(5개년)동안 발생하는 전염성 질환의 발생추이에 대한 패턴분석 및 예측 알고리즘 구축
  - 광역지자체(시도) 단위 분석: 시군구단위 전염성질환은 발생건수가 적거나 없는 경우가 빈번하여 분석에 지장

- 중간보고까지 진행사항
  - 데이터 전처리(기후, 대기, 인구 데이터 및 표본 코호트 DB데이터)
  - 예측 대상이 되는 질병 선정
  - 전염성 질환 발생 추이에 대한 시계열 분석
  - GAM 분석 및 회귀모형을 통한 질병에 영향을 미치는 유의미한 변수 추출

- 최종보고 발표 예정 : 특정 전염성 질환에 대한 예측 알고리즘 구축

연구동향 분석 작업 흐름도



데이터 전처리: 설명변수

- 설명변수: 기후, 대기, 인구 데이터 전처리
  - 대기오염물질 농도, 기상기후데이터, 인구데이터를 설명변수로 사용

설명변수 데이터

데이터	출처	기간	항목	주기	공간 단위
대기오염물질 농도	한국환경공단	2009-2013	PM_{10}, NOx, SOx, CO, O_3	1시간	측정소
기상기후 데이터	기후데이터센터	2009-2013	기온, 습도, 풍향, 풍속, 기압, 강수량, 일조, 일사 등	일 단위	측정소
인구	국가통계포털	2009-2013	주민등록인구현황, 주민등록인구세대수 등	연 단위	시군구

데이터 전처리 프로세스

시도/시군구
통일

시도/시군구를 중심으로
대표값 산출(최대, 최소, 평균값)

- Step 1. 공간 해상도
  - 측정소 데이터는 공간적으로 점(point)데이터이고, 시군구/시도의 경우 면(polygon)데이터임
  - 공간해상도를 맞추기 위하여 같은 시군구/시도에 위치한 측정소의 데이터를 평균 내어 매칭
- Step 2. 시간 해상도
  - 시간해상도는 년, 월, 일 모두 다르며, 분석을 수행할 시간해상도는 월 단위
  - 월 단위보다 시간해상도가 낮은 경우: 시간, 일 단위 데이터의 평균을 월 단위 데이터로 사용
  - 월 단위보다 시간해상도가 높은 경우: 연 데이터를 12로 나누어 사용하고, 농도, 밀도(인구)와 같은 경우 연 데이터를 그대로 사용

데이터 전처리: 질병 선정 및 건수 도출

- Step 1. 자격 DB와 진료 DB 연계: 한 달 이내 동일인 동일한 질병 방문 = 1건 간주
- Step 2. 진료 DB 주상병명과 부상병명이 다른 경우: 다른 케이스로 간주
- Step 3. 2009~2013년 연속 발생 질병만 분석 대상으로 고려
- Step 4. 질병코드(한국질병표준사인분류 기준) 소수점 그룹화 : 질병 레벨을 높여 건수를 산출

질병코드	질병이름
A00	콜레라
A01	장티푸스 및 파라티푸스
A02	기타 살모넬라감염
A03	시겔라증
A04	기타 세균성 장감염
A05	달리 분류되지 않은 기타 세균성 장감염
A06	아메바증
A07	기타 원충성 장질환
A08	바이러스성 및 기타 명시된 장감염
A09	감염성 및 상세불명 기원의 기타 위장염 및 결장염
A15	세균학적 및 조직학적으로 확인된 호흡기결핵
A16	세균학적 및 조직학적으로 확인되지 않은 호흡기결핵

장 감염 질환을
분석 질병으로 선정

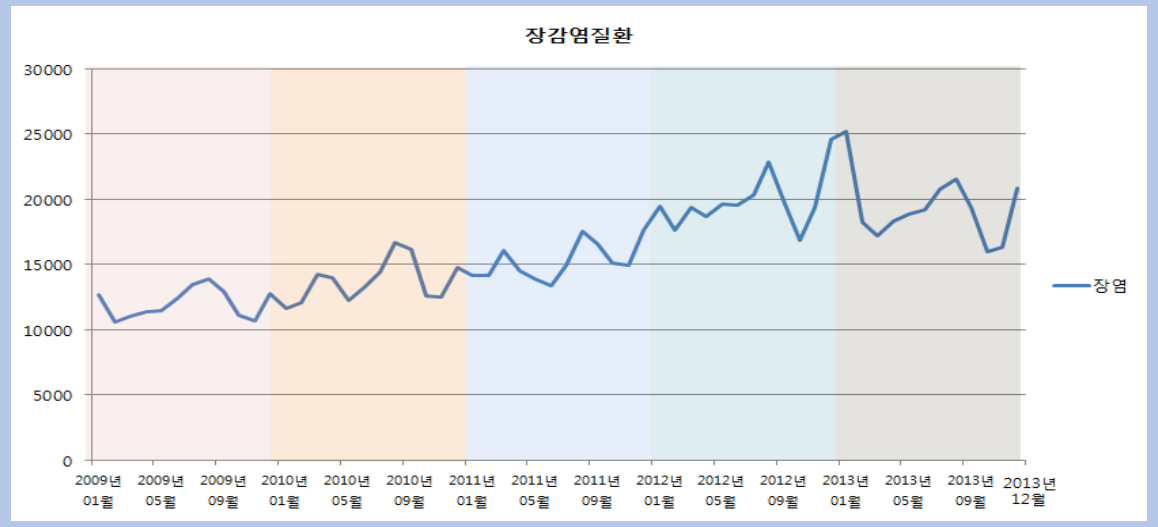
전염성 질환 데이터 시공간 시계열 분석

- 장감염질환의 5개년 시계열 분석
  - 연도별 월별 질환 발생 건수 추이 분석: 계절 /날씨 영향 파악
    - 12-1월과 7-8월에 질환이 급격하게 늘어남
    - 발생빈도가 약간씩 증가하는 경향을 보임

연도별 최다빈도 건수

연도	월	건수
2009	8월	13,855
2010	8월	16,675
2011	8월	17,516
2012	12월	24,574
2013	12월	25,200

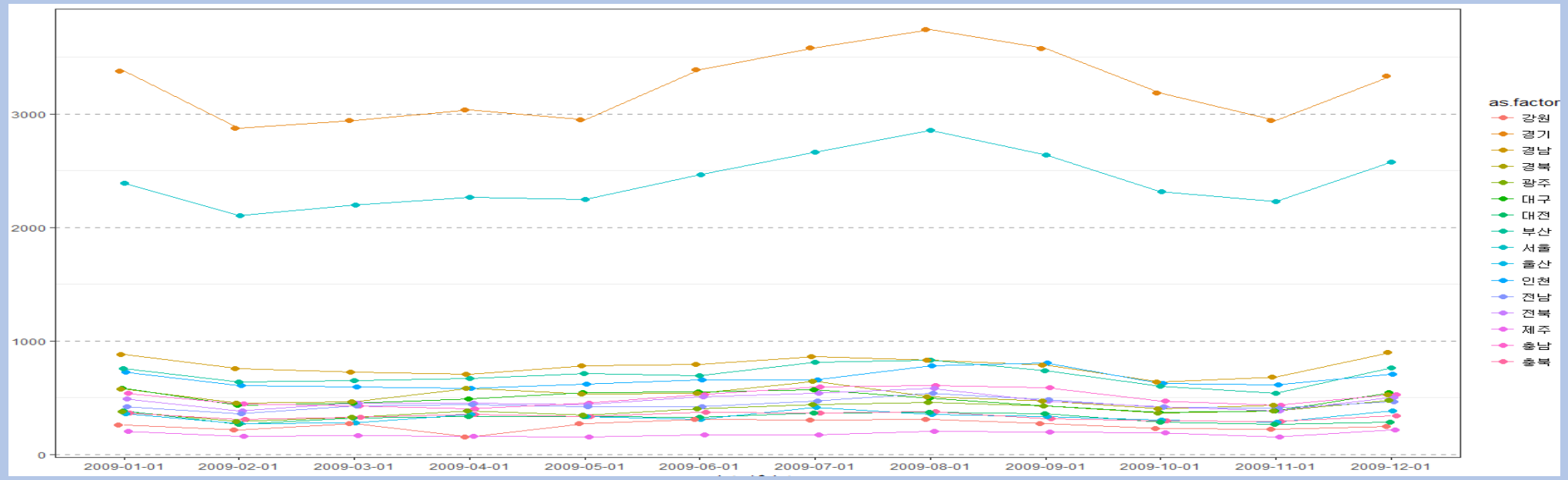
장감염질환 5개년 시계열 분석



전염성 질환 데이터 시공간 시계열 분석

- 7-8월, 12-1월에 발생 빈번: 시도별 추이 유사
- 인구밀도 대비 전염성 질환의 발생 건수: 경북이 빈번

장감염질환 공간적 시계열 분석(2009년)



인구밀도 대비 발생 건수

시도	2009	2010	2011	2012	2013
강원	33.7	37.5	44.0	55.5	53.3
경기	34.4	38.6	42.3	53.3	51.6
경남	30.3	36.2	40.0	51.6	50.3
경북	43.9	52.0	57.6	72.0	69.1
광주	1.6	2.0	2.2	3.0	3.0
대구	2.1	2.3	2.6	3.6	3.5
대전	1.4	1.6	1.7	2.6	2.6
부산	1.8	2.1	2.4	3.3	3.3
서울	1.7	1.8	1.9	2.5	2.5
울산	3.8	4.4	4.6	6.2	5.8
인천	3.0	3.5	4.1	5.1	4.7
전남	33.8	36.7	42.8	55.5	58.1
전북	24.5	26.9	30.0	42.7	41.7
제주	7.1	8.1	8.8	9.8	9.8
충남	25.5	29.0	33.0	38.2	36.4
충북	19.7	21.7	24.1	33.0	30.8

유의미한 변수 도출 및 인과관계 분석

- 장감염 질환에 영향을 미치는 유의미한 변수 도출
  - GAM(Generalized Additive Model), OLS 이용

GAM 분석 결과				다중선형회귀 분석 결과			
		변수명	계수추정치	sig		변수명	계수추정치	sig
양의관계	기후	최소상대습도	8.571e-03	***	양의관계	기후	평균현지기압	2.574e+00	*
		평균풍속	2.396e-01	***			일조율_max	3.581e+01	*
		일조시간합_mean	1.564e-01	***					
		일조율_min	3.694e-01	***					
	대기	CO_mean	4.495e-01	***		대기	CO_mean	3.524e+02	***
		O3_mean	7.111e+00	***			O3_mean	2.958e+03	***
		No2_mean	9.626e+01	***			SO2_max	9.904e+02	***
음의관계	기후	최저기온	-1.868e-03	*	음의관계	기후	평균최저기온	-1.616e+01	***
		일최다강수량	-3.134e-03	***			일조시간합_max	-1.044e+01	*
		일조율_mean	-5.131e-01	***			일조율_min	-3.643e+01	*
	대기	CO_max	-4.638e-02	***		대기	PM_{10}_mean	-6.731e+00	***
							NO_2_max	-6.313e+02	***

Signif: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

GAM model

$$Y \sim (\mu, \quad)$$
$$g(E(Y)) = \beta_0 + \sum_p s_p(X_p)$$

s_i : smooth function estimated from data

향후 계획

- 변수 선정 및 인과관계 분석 보완
  - GAM을 이용한 인과관계 분석 보완
- 예측 알고리즘 구축(RNN을 이용)

과업수행내용	3월	4월	5월	6월	7월	8월	9월	10월	11월	12월
1. 문헌조사										
- 기후와 전염성 질환과의 관계에 대한 관련 연구 검토										
- RNN을 비롯한 머신러닝을 바탕으로 전염성 질환 예측과 관련된 연구 검토										
2. 전염성 질환의 시공간적 패턴분석										
- 데이터 전처리										
- 분석 질병선정										
- 시공간적 시계열 분석										
3. 전염성 질환 예측 알고리즘 구축										
- 예측 알고리즘 구축										
- 예측 정확도 비교										
■ 최종보고서 평가, 보완 및 작성 완료										

2. 세부과제별 연구진행상황

3장. 환경 빅데이터 연구

4. 미세먼지 오염도-발생요인 패턴분석 (김진형)

미세먼지 오염도-발생요인 패턴분석

- 연구내용
  - 미세먼지(PM_{10})에 영향을 미친다고 알려진 변수들을 영향을 의사결정나무 분석을 사용하여 정량적으로 평가
  - 종속변수: 2001년~2016년 9월까지 189개월 동안 측정된 미세먼지(PM_{10}) 데이터
  - 설명변수: 기상기후 데이터, 대기오염물질 배출량 데이터, 황사 및 중국 미세먼지(PM_{10}) 데이터, 인구밀도 데이터

- 중간보고까지의 진행사항
  - 데이터 수집 및 전처리 (대기오염물질 농도, 기상기후 데이터)
  - 예측 변수 기술 통계 작성
  - 의사결정나무 분석 수행할 툴 및 패키지 리서치

- 최종보고 발표 예정 : 의사결정나무 분석 결과 및 해석
  - 설명 변수 추가
  - 의사결정나무 분석 수행
  - 분석 결과 해석

수집 데이터 특성

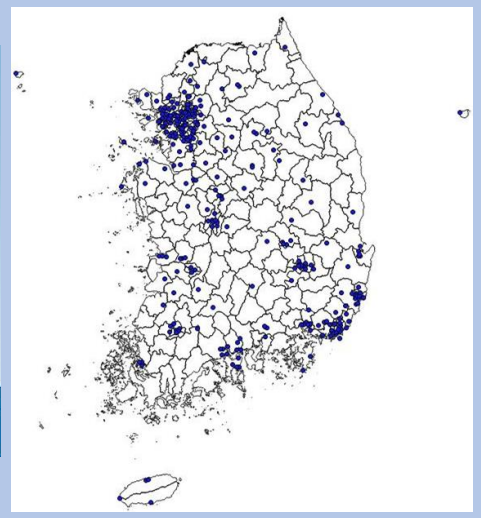
시간해상도 -> 월
공간해상도 -> 시군구

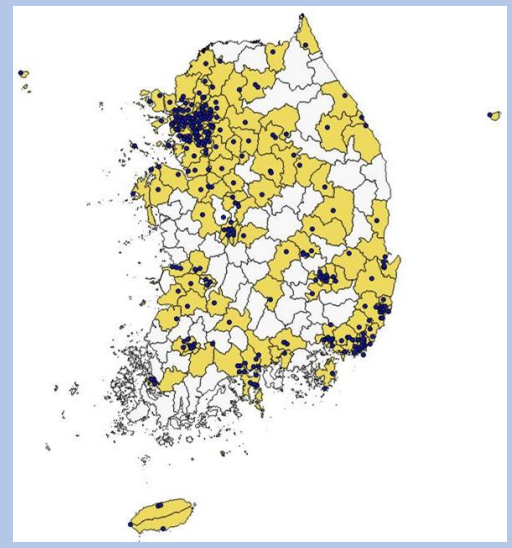
데이터명	출처	기간	항목	시간 해상도	공간 해상도
대기오염물질 농도	에어코리아	2001-2016	PM_{10}, NO_x, SO_x, CO, O_3, $PM_{2.5}$	1시간	측정소
대기오염물질 배출량	국립환경과학원	1999-2013	PM_{10}, NO_x, SO_x, CO, O_3, $PM_{2.5}$, VOC, NH_3	연 단위	시군구
기상기후 데이터	기후데이터센터	1995-2016	기온, 습도, 풍향, 풍속, 기압, 강수량, 일조, 일사, 황사일수 등	일, 월, 연 단위	측정소
인구	국가통계포털	-2016	주민등록인구현황, 주민등록인구세대수 등	월, 연 단위	시군구
행정구역별 면적	국토교통부 통계누리	1992-2016	면적, 지번수, 지목별 현황 등	연 단위	시군구
중국 대기질	주중 미국 대사관	2008-2016	AQI	1시간	측정소
	중화인민공화국 환경보호부 데이터센터	2014-2016	$PM_{2.5}$	일 단위	주요 도시

분석 대상 시군구

대기오염물질 측정소
CODE
TYPE
* NAME
LOCATION
ADDRESS
LAT
LNG
ALT

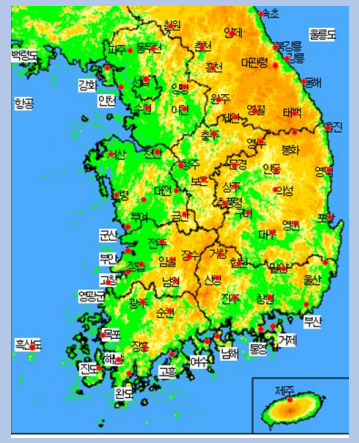
대기오염물질 측정 데이터
* DATE_TIME
* NAME
YEAR
MONTH
DAY
HOUR
SO2
CO
O3
NO2
PM10
PM25

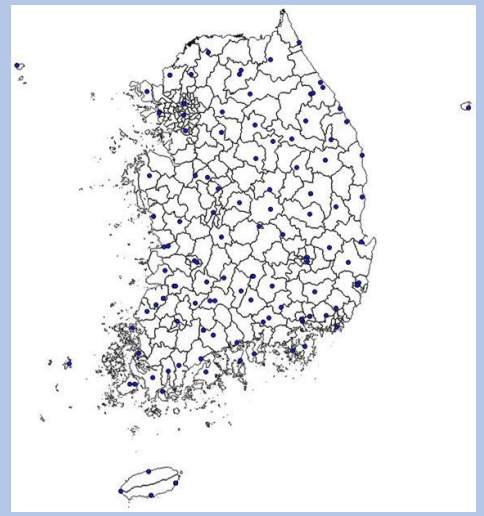




기상기후 측정소
* CODE
TYPE
NAME
LOCATION
ADDRESS
LAT
LNG
ALT

기상기후 데이터
* DATE_TIME
* CODE
YEAR
MONTH
HOUR
TEMP
PRESSURE
PRECIPITATION
⋮

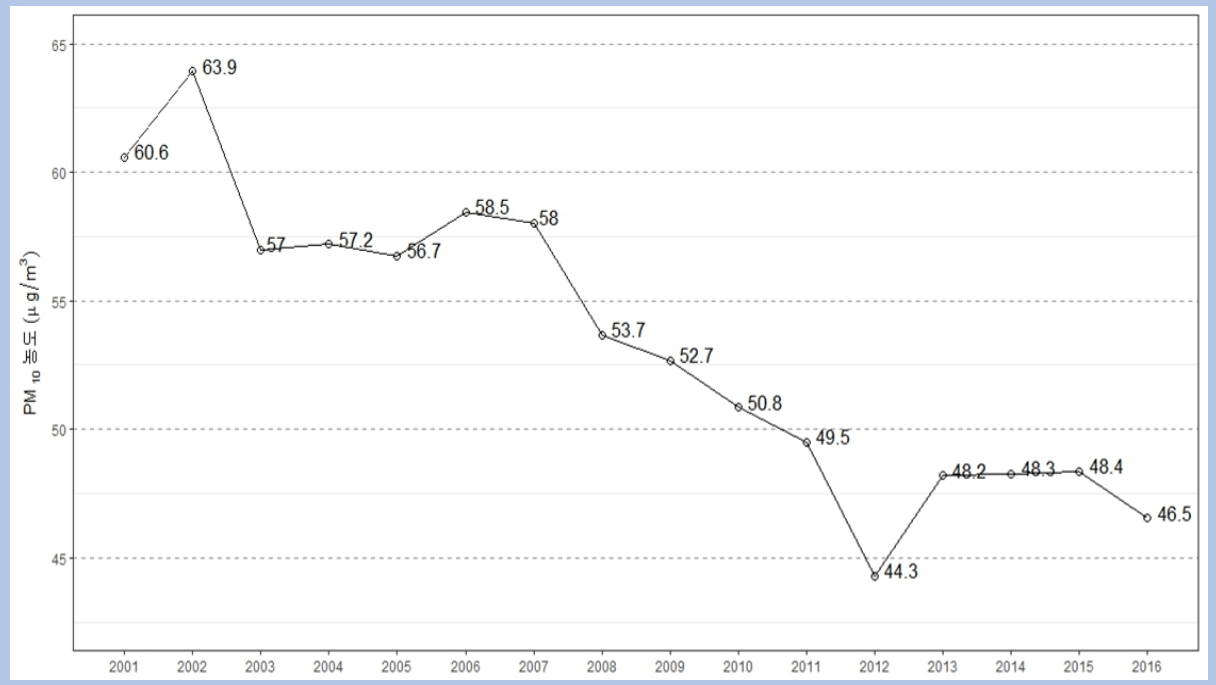




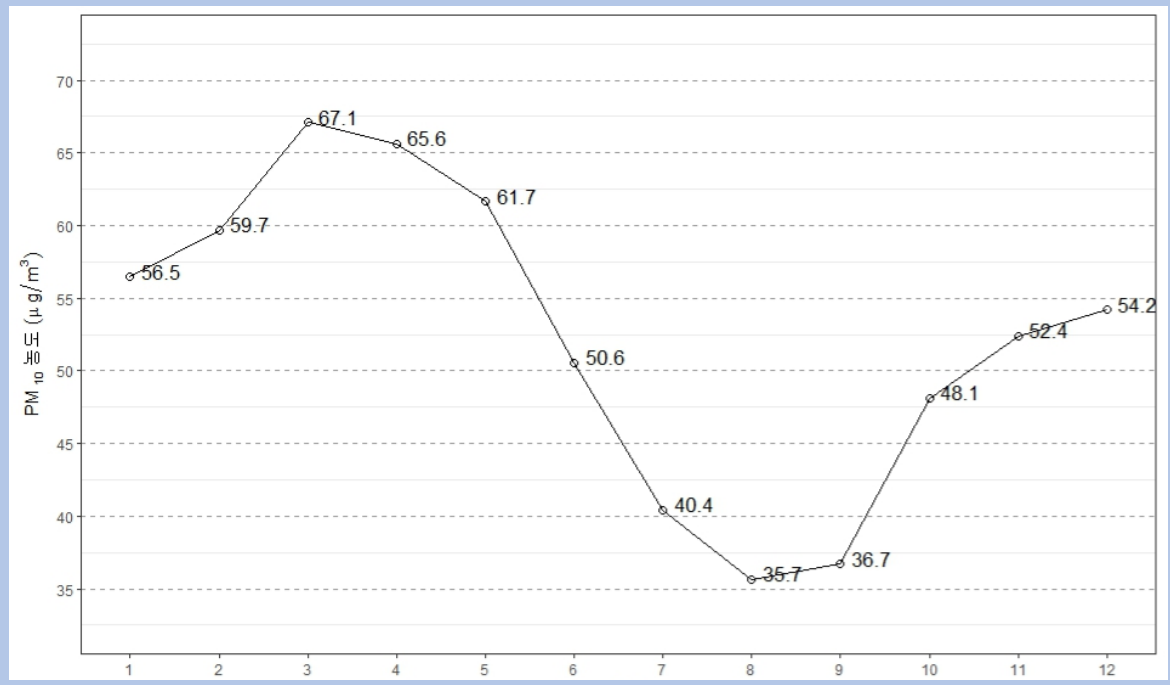
예측변수 기술통계

- 미세먼지(PM_{10}) 연평균 농도: 2001년부터 2016년 9월 전반적 감소 추세
- 분석기간의 월평균 농도는 봄철에 높고 여름철에 낮음

연평균 농도



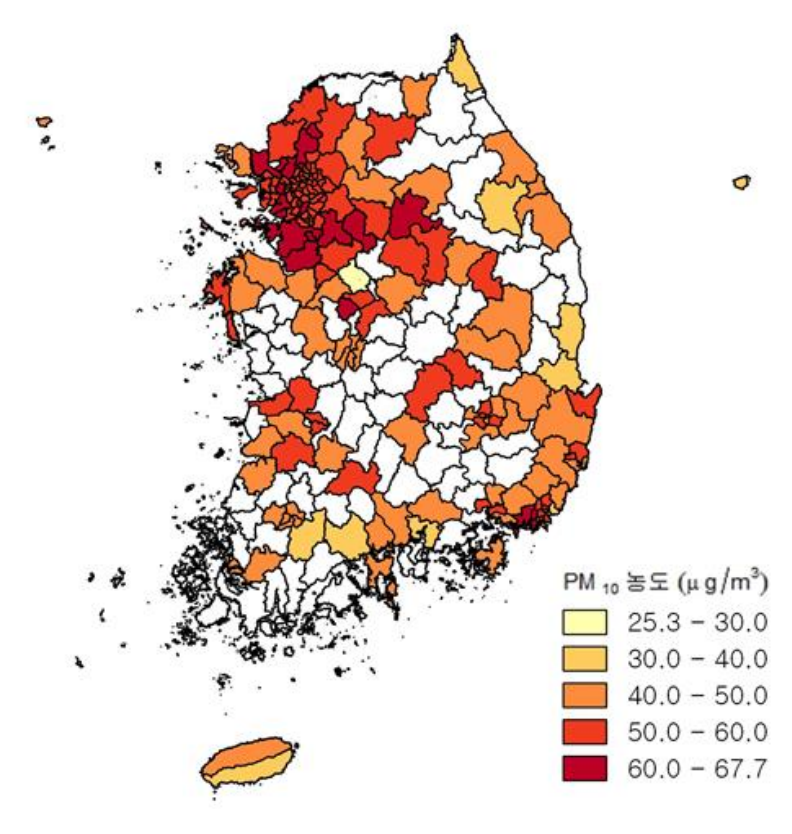
월평균 농도



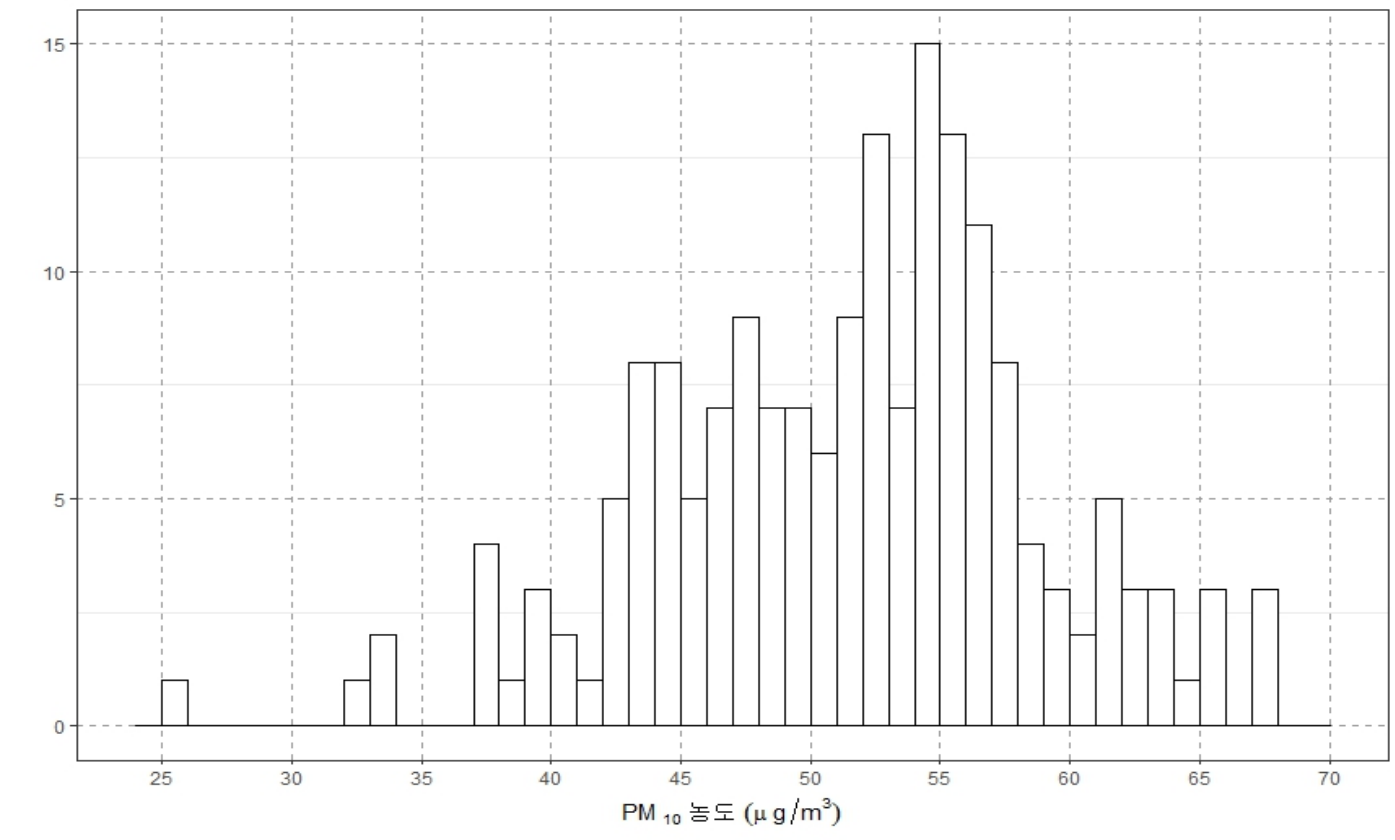
예측변수 기술통계

- 수도권과 광역시의 농도가 높은 편
- 43이하 구간에 20개, 43-58 구간에 133개, 58이상 구간에 28개의 시군구

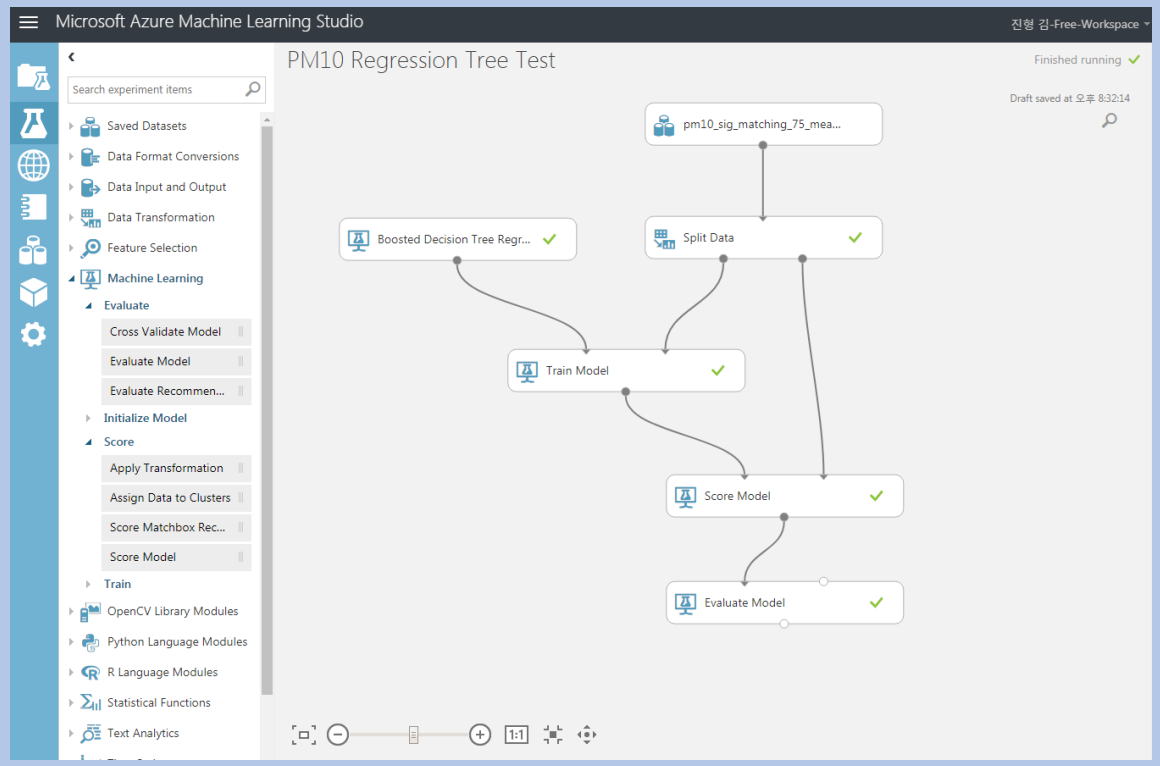
시군구 평균 지역 분포

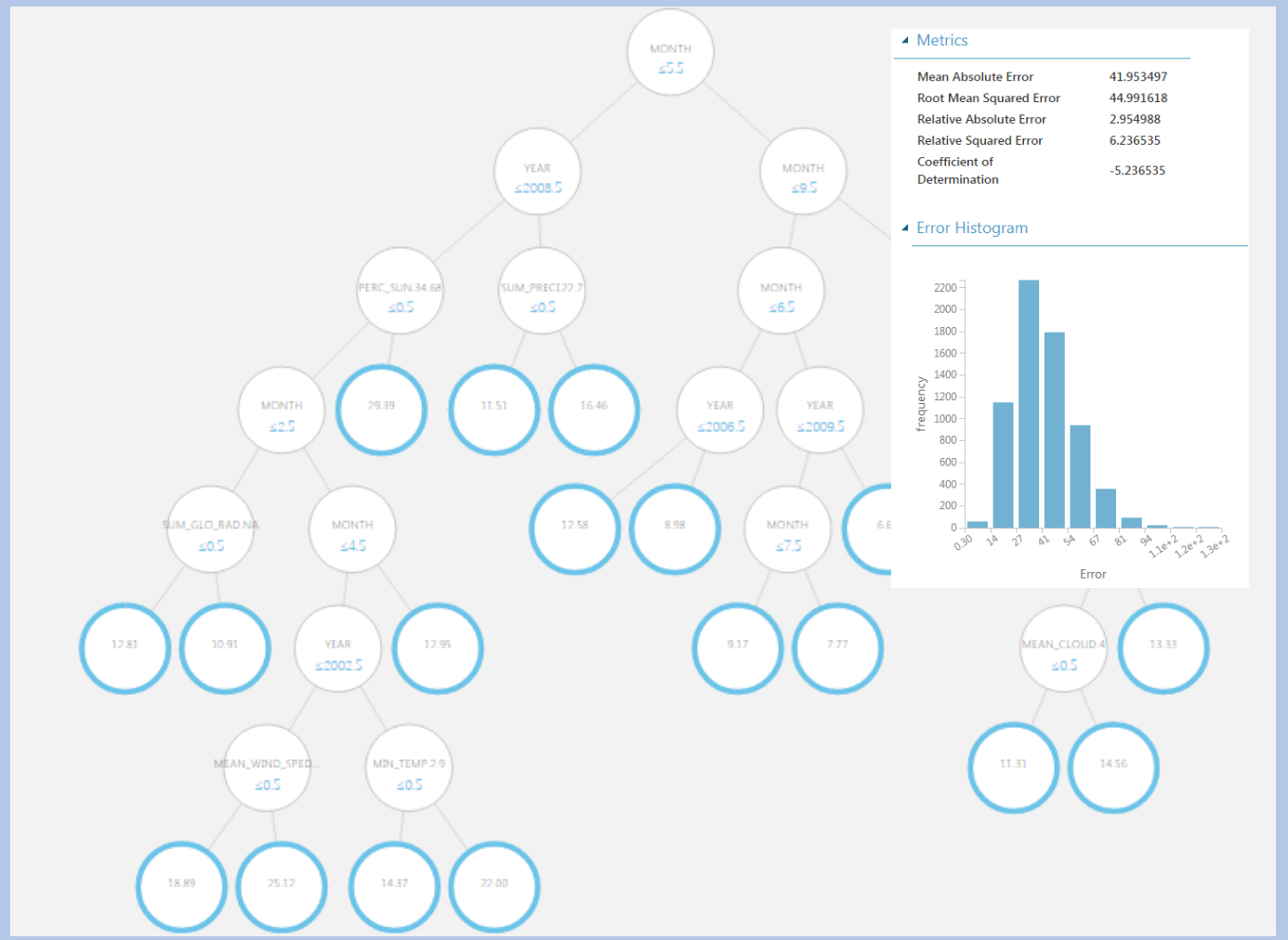


시군구 평균 histogram



의사결정나무 분석 Tool





- Microsoft Azure: Drag & Drop
- 대기오염물질 농도와 기상기후 데이터만으로 테스트 (향후 결과 해석 예정)
- 향후 변수 추가, 변형 등을 통해 분석 결과 해석 예정

향후 계획

- 변수 추가
  - 향후 데이터 수집 및 전처리를 통해 변수 추가
- 분석 시행
  - 의사결정나무 분석 시행
  - boosting, random forest 등 시행
- 결과 해석
  - 변수 설명력의 정량적 평가
  - 정책적 제언

2. 세부과제별 연구진행상황

3장. 환경 빅데이터 연구

5. 환경분야 빅데이터 수집방법연구 (한국진)

환경분야 빅데이터 수집방법연구

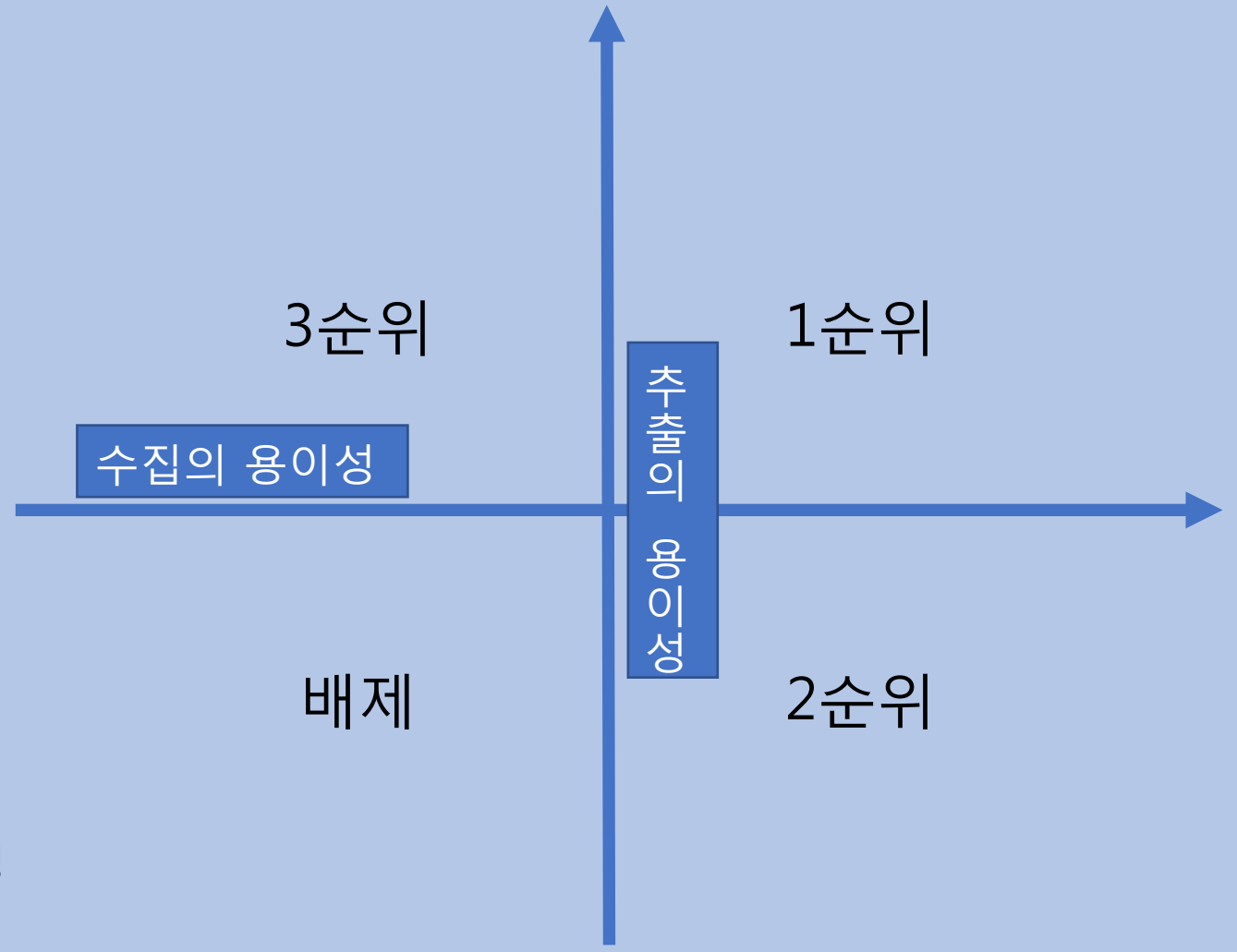
- 저작권 문제가 발생하지 않는 국내·외 데이터 서비스의 다양한 데이터 중 환경분야 활용 가능한 데이터의 수집방법 마련
  - 공공데이터포털: 16개 대분류로 (비)선택적으로 데이터셋을 공개
  - 환경기상: 파일데이터(1,202건), 오픈API(129건), 표준데이터(2건)

- 공공데이터포털에서 발표되는 월별 공공데이터 활용신청 TOP10, 연간 누적데이터 TOP 20을 기준으로 대기 및 기상 데이터를 검토
  - 한국환경공단: 대기오염정보, 측정소정보
  - 기상청: 동네예보정보, 생활기상지수조회

데이터 구분 기준에 따라 우선순위 부여

- 수집의 용이성
  - 수집대상 데이터 목록 자동화 가능
    - 파일 형태 제공
    - **특정 조건 검색 결과 파일 형태(CSV, XLS 등) 저장이 용이**

- 추출의 용이성
  - 수집된 데이터의 분리가 가능
    - 원본 데이터: CSV(텍스트), XLS(바이너리)
    - **PDF, HWP, html: [1차 분리추출 ; 파싱 Parsing] →[2차 분리정제]**



1순위 사례 : PM_{10}, $PM_{2.5}$ 검색 → 추출

- 검색조건 검색을 통한 PM_{10}, $PM_{2.5}$ 데이터 추출
  - Python urllib, BeautifulSoup, pandas

```
import urllib.request
from urllib.request import urlretrieve
from bs4 import BeautifulSoup
import pandas as pd
```

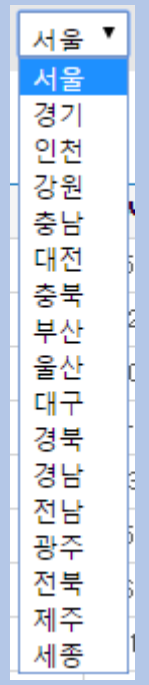
① 라이브러리 활성화

```
searchDate="2017-03-20"
#district=02
#itemCode=10007
#searchDate_f=201706
dst_url = "http://www.airkorea.or.kr/pmRelaySub?strDateDiv=1&searchDate=" + searchDate + "&distr
print(dst_url)
webpage = urllib.request.urlopen(dst_url)
ori_content = webpage.read().decode("utf-8")
content = BeautifulSoup(ori_content, "html.parser")
content
```

② 일자 설정(searchDate)

③ 지역 설정(district)

④ strDateDiv : 시간(1), 일평균(2)



PM_{10}, $PM_{2.5}$ 검색 → 추출: 추출 데이터 검토

```
#content
df = pd.read_html("http://www.airkorea.or.kr/pmRelaySub?strDateDiv=1&searchDate=" \
                  + "2017-03-20&district=02&itemCode=10007&searchDate_f=201706", encoding='utf-8')
df
```

```
[      측정망          측정소명   1시   2시   3시   4시   5시   6시   7시   8시 ...   15시  \
 0    도시대기      [서울]강남구  101   91   79   76   68   76   75   68 ...    70
 1    도시대기      [서울]강동구  123   93   79   78   91   78   73   80 ...    74
 2    도시대기      [서울]강북구   53   52   53   55   55   54   64   61 ...    55
 3    도시대기      [서울]강서구  117  111   97   91   79   98  109  100 ...   109
 4    도시대기      [서울]관악구   86   76   77   66   70   63   72   67 ...    81
```

```
df[0]
```

	측정망	측정소명	1시	2시	3시	4시	5시	6시	7시	8시	...	15시	16시	17시	18시	19시	20시	21시	22시	23시	24시
0	도시대기	[서울]강남구	101	91	79	76	68	76	75	68	...	70	92	93	110	154	123	151	176	155	152
1	도시대기	[서울]강동구	123	93	79	78	91	78	73	80	...	74	81	74	77	134	143	165	179	157	147
2	도시대기	[서울]강북구	53	52	53	55	55	54	64	61	...	55	55	61	83	112	122	139	131	120	120
3	도시대기	[서울]강서구	117	111	97	91	79	98	109	100	...	109	-	135	128	129	147	127	124	166	147
4	도시대기	[서울]관악구	86	76	77	66	70	63	72	67	...	81	80	98	119	116	134	141	133	133	138
5	도시대기	[서울]광진구	90	77	80	69	69	58	64	68	...	66	57	61	71	123	120	139	147	146	145
6	도시대기	[서울]구로구	102	89	86	88	96	88	76	74		97	-	-	167	134	141	162	158	149	143

PM_{10}, $PM_{2.5}$ 검색 → 추출 : 추출 결과 저장

저장 Code

```
filepath = "./Files/PM10/"+ "PM10_KEC_서울_2017-03-20.csv"
#filepath
#help(pd.to_csv)
#dir(df)
#df

#df[0].to_csv(filepath, sep=',', encoding="utf-8", index=false, hea
df[0].to_csv(filepath, sep=',', encoding='utf-8', index_col=false)
#sep='\t'
#df.to_csv(filepath, sep=',')
```

저장 결과

2순위 사례 : ASEC 리포트

- 파일 리스트가 존재하는 웹페이지 html code 확보
  - Python urllib, BeautifulSoup

```
import urllib.request
from urllib.request import urlretrieve

from bs4 import BeautifulSoup
```

```
dst_url = "http://www.ahnlab.com/kr/site/securityinfo/asec/asecReportList.do"
webpage = urllib.request.urlopen(dst_url)
#print(webpage.geturl()) #URL
#print(webpage.info()) #헤더
#print(webpage.getcode()) #상태
ori_content = webpage.read().decode("utf-8")
content = BeautifulSoup(ori_content, "html.parser")
content
```

ASEC : AhnLab Security Emergency Response Center

ASEC 리포트 : 파일 리스트 파싱 → 다운로드

파일 리스트 파싱

```
reportlist = content.find_all('a', {"class": "btnPdf"})
#len(reportlist)
#reportlist
#reportlist[0]
download_url = []

for pdfurl in reportlist:
    #print(pdfurl['href'])
    download_url.append(pdfurl['href'])

print("추출된 URL수 : ", len(download_url))
download_url = list(set(download_url))
print("정제된 URL수 : ", len(download_url))

download_url.sort
#print(download_url)
```

다운로드

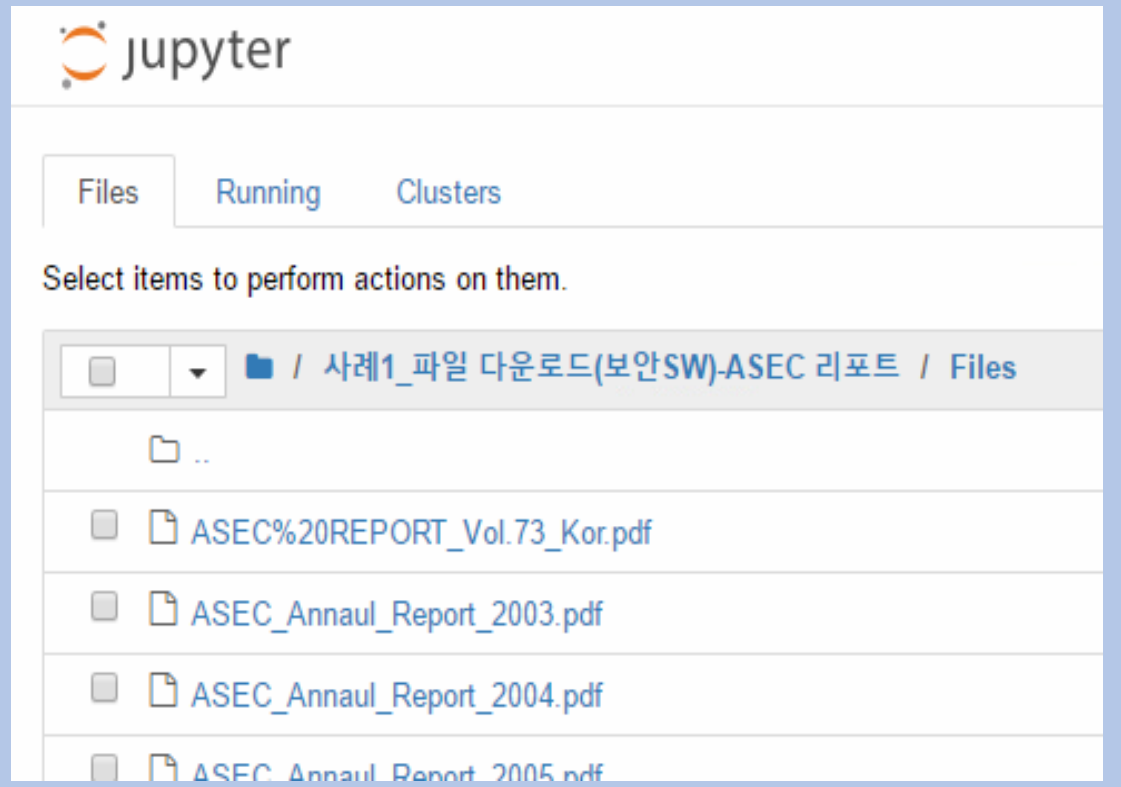
```
#download_url[0] #파일URL
#http://download.ahnlab.com/asecReport/ASEC_Report_200811.pdf
#len(download_url[0].split('/'))
#download_url[0].split('/')[-1] # 파일명
#from urllib.request import urlretrieve

print("다운로드를 시작합니다.\n")

for download_file in download_url:
    download_filename = download_file.split('/')[-1]
    print(download_filename, " 파일을 다운로드 합니다.")
    print("다운로드 URL : ", download_file)
    download_path = "./Files/" + download_filename
    #print(download_path)
    urlretrieve(url=download_file, filename=download_path)
    print(download_filename, " 파일이 저장되었습니다.\n")
    print("다운로드 경로 : ", download_file)

print("다운로드가 완료되었습니다.")
```

ASEC 리포트 : 파일 아카이브

수집된 파일 저장 (과정)

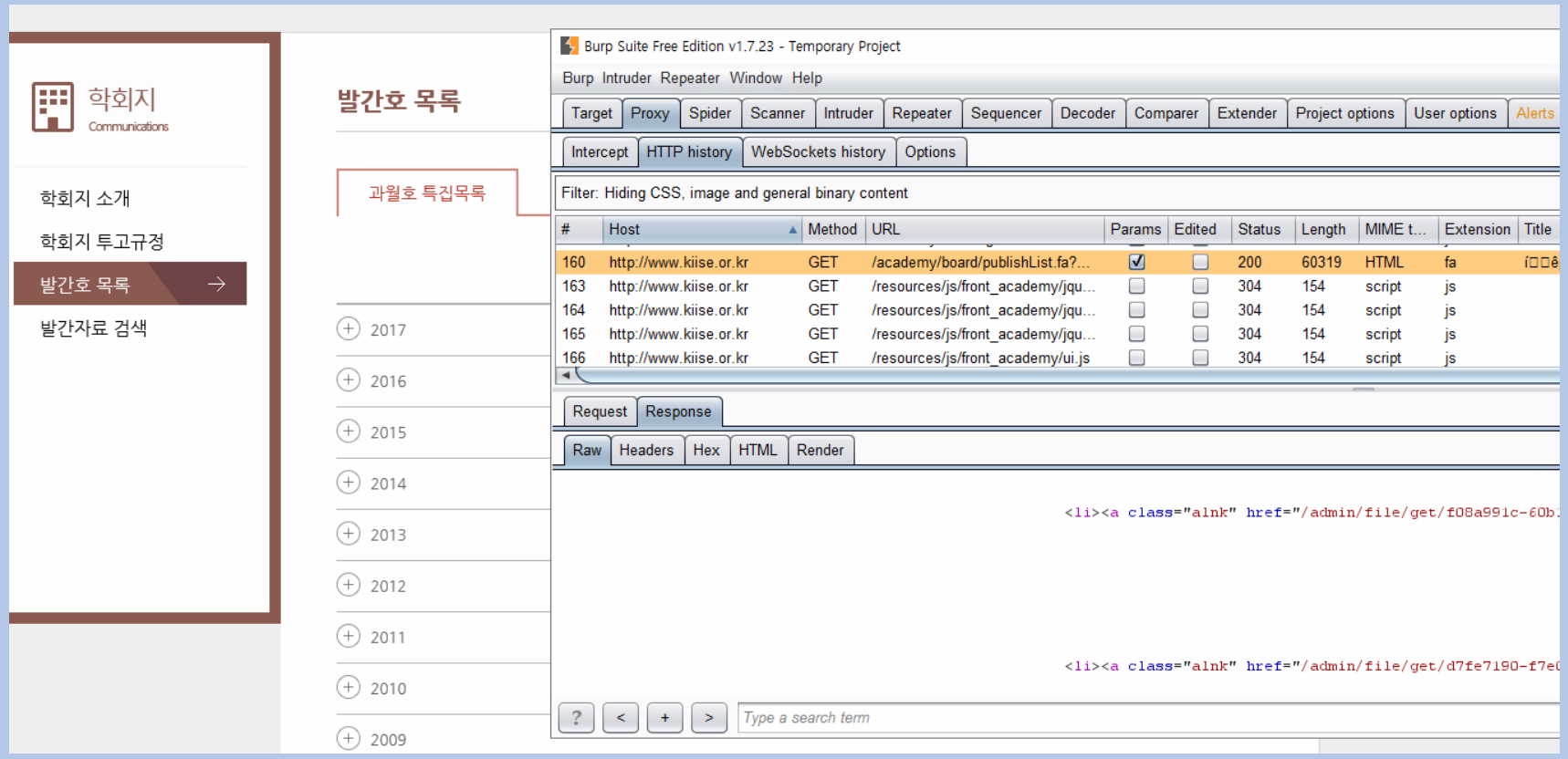
```
ASEC_Report_Vol.64_Kor.pdf  파일이 저장되었습니다.

다운로드 경로 :  http://download.ahnlab.com/asecReport/ASEC_Report_Vol.64_Kor.pdf
ASEC_REPORT_vol.80.pdf  파일을 다운로드 합니다.
다운로드 URL :  http://image.ahnlab.com/file_upload/asecissue_files/ASEC_REPORT_vol.80.pdf
ASEC_REPORT_vol.80.pdf  파일이 저장되었습니다.

다운로드 경로 :  http://image.ahnlab.com/file_upload/asecissue_files/ASEC_REPORT_vol.80.pdf
ASEC_Report_201005.pdf  파일을 다운로드 합니다.
다운로드 URL :  http://download.ahnlab.com/asecReport/ASEC_Report_201005.pdf
ASEC_Report_201005.pdf  파일이 저장되었습니다.

다운로드 경로 :  http://download.ahnlab.com/asecReport/ASEC_Report_201005.pdf
ASEC_Report_200905.pdf  파일을 다운로드 합니다.
다운로드 URL :  http://download.ahnlab.com/asecReport/ASEC_Report_200905.pdf
ASEC_Report_200905.pdf  파일이 저장되었습니다.

다운로드 경로 :  http://download.ahnlab.com/asecReport/ASEC_Report_200905.pdf
다운로드가 완료되었습니다.
```

저장 결과



3순위 사례 : 한국정보과학회지

- HTML code로부터 URL 패턴 파악
  - HTML code 로 부터 직접 url을 추출하기 어려운 경우
  - Python urllib, BeautifulSoup / 웹분석SW Burp Suite



한국정보과학회지 : 파일 리스트 추정

URL 생성 규칙에 따라 목록 생성

```
#HTML 페이지를 확인해보니 2013년 1월부터 제공함
#해당 페이지는 로그인 필요함(http://www.kiise.or.kr/academy/board/publishList.fa?MENU_ID=050300#no
#Selenimum으로 처리할 수 있으나 로그인 후 주소가 바로 드러나지 않음(메일링리스트를 통해 추출)
#http://www.kiise.or.kr/e_communications/2017/06/2017_06.pdf
destYear = 2013
destMonth = 1

#http://www.kiise.or.kr/e_communications/2017/06/2017_06.pdf
dst_url = "http://www.kiise.or.kr/e_communications/"

download_urls = []

while(destYear <= datetime.today().year):
    #print(destYear)
    for destMonth in range(12):
        if(destMonth == datetime.today().month and destYear == datetime.today().year):
            break

        #print(str(destYear), str(destMonth+1).zfill(2))
        dst_filename = str(destYear) + "_" + str(destMonth+1).zfill(2) + ".pdf"
        download_url = dst_url + str(destYear) + "/" + str(destMonth+1).zfill(2) + "/" + dst_fil
        print(download_url)
        download_urls.append(download_url)

    destYear += 1
```

```
http://www.kiise.or.kr/e_communications/2013/01/2013_01.pdf
http://www.kiise.or.kr/e_communications/2013/02/2013_02.pdf
http://www.kiise.or.kr/e_communications/2013/03/2013_03.pdf
http://www.kiise.or.kr/e_communications/2013/04/2013_04.pdf
http://www.kiise.or.kr/e_communications/2013/05/2013_05.pdf
http://www.kiise.or.kr/e_communications/2013/06/2013_06.pdf
http://www.kiise.or.kr/e_communications/2013/07/2013_07.pdf
http://www.kiise.or.kr/e_communications/2013/08/2013_08.pdf
http://www.kiise.or.kr/e_communications/2013/09/2013_09.pdf
http://www.kiise.or.kr/e_communications/2013/10/2013_10.pdf
http://www.kiise.or.kr/e_communications/2013/11/2013_11.pdf
http://www.kiise.or.kr/e_communications/2013/12/2013_12.pdf
http://www.kiise.or.kr/e_communications/2014/01/2014_01.pdf
http://www.kiise.or.kr/e_communications/2014/02/2014_02.pdf
http://www.kiise.or.kr/e_communications/2014/03/2014_03.pdf
http://www.kiise.or.kr/e_communications/2014/04/2014_04.pdf
http://www.kiise.or.kr/e_communications/2014/05/2014_05.pdf
http://www.kiise.or.kr/e_communications/2014/06/2014_06.pdf
http://www.kiise.or.kr/e_communications/2014/07/2014_07.pdf
http://www.kiise.or.kr/e_communications/2014/08/2014_08.pdf
http://www.kiise.or.kr/e_communications/2014/09/2014_09.pdf
http://www.kiise.or.kr/e_communications/2014/10/2014_10.pdf
http://www.kiise.or.kr/e_communications/2014/11/2014_11.pdf
http://www.kiise.or.kr/e_communications/2014/12/2014_12.pdf
http://www.kiise.or.kr/e_communications/2015/01/2015_01.pdf
http://www.kiise.or.kr/e_communications/2015/02/2015_02.pdf
http://www.kiise.or.kr/e_communications/2015/03/2015_03.pdf
http://www.kiise.or.kr/e_communications/2015/04/2015_04.pdf
http://www.kiise.or.kr/e_communications/2015/05/2015_05.pdf
http://www.kiise.or.kr/e_communications/2015/06/2015_06.pdf
http://www.kiise.or.kr/e_communications/2015/07/2015_07.pdf
http://www.kiise.or.kr/e_communications/2015/08/2015_08.pdf
http://www.kiise.or.kr/e_communications/2015/09/2015_09.pdf
http://www.kiise.or.kr/e_communications/2015/10/2015_10.pdf
http://www.kiise.or.kr/e_communications/2015/11/2015_11.pdf
http://www.kiise.or.kr/e_communications/2015/12/2015_12.pdf
http://www.kiise.or.kr/e_communications/2016/01/2016_01.pdf
http://www.kiise.or.kr/e_communications/2016/02/2016_02.pdf
http://www.kiise.or.kr/e_communications/2016/03/2016_03.pdf
http://www.kiise.or.kr/e_communications/2016/04/2016_04.pdf
http://www.kiise.or.kr/e_communications/2016/05/2016_05.pdf
http://www.kiise.or.kr/e_communications/2016/06/2016_06.pdf
http://www.kiise.or.kr/e_communications/2016/07/2016_07.pdf
http://www.kiise.or.kr/e_communications/2016/08/2016_08.pdf
```

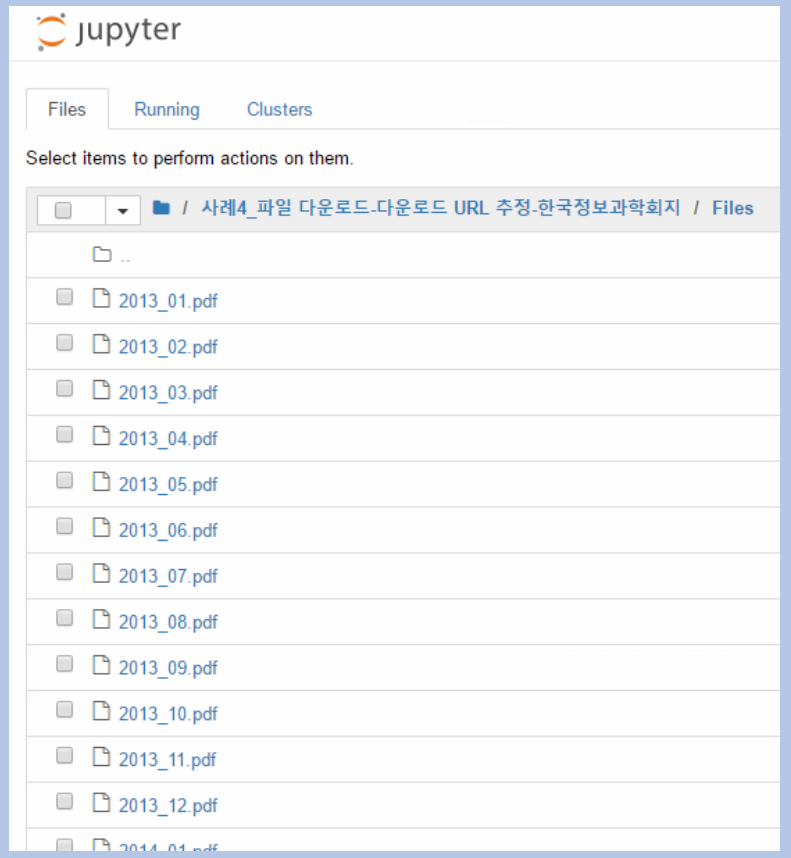
한국정보과학회지 : 파일 다운로드

URL 리스트에 맞춰 파일 다운로드

```
print("다운로드를 시작합니다.\n")

for download_file in download_urls:
    print("다운로드 URL : ", download_file)
    #print(download_file.split('/')[-1])
    urlretrieve(url=download_file, filename="./Files/" + download_file.split('/')[-1])
    print(download_file.split('/')[-1], " 파일이 저장되었습니다.\n")

print("다운로드가 완료되었습니다.")
```

```
다운로드를 시작합니다.

다운로드 URL :  http://www.kiise.or.kr/e_communications/2013/01/2013_01.pdf
2013_01.pdf  파일이 저장되었습니다.

다운로드 URL :  http://www.kiise.or.kr/e_communications/2013/02/2013_02.pdf
2013_02.pdf  파일이 저장되었습니다.

다운로드 URL :  http://www.kiise.or.kr/e_communications/2013/03/2013_03.pdf
2013_03.pdf  파일이 저장되었습니다.

다운로드 URL :  http://www.kiise.or.kr/e_communications/2013/04/2013_04.pdf
2013_04.pdf  파일이 저장되었습니다.
```

저장된 파일



향후 과제

- Data Source
  - 원내 설문조사 등 활용 가능한 환경 데이터 발굴
- 수집 방법
  - Python(탐색)/Elastic Stack(자동화) 등 제약 극복
  - OpenAPI 활용시 코딩 부담 절감
- 수집된 데이터 활용
  - 수집된 데이터를 RDBMS(Relational Data Base Management System)로 전환
    - PostgreSQL 사용
  - 도입 예정인 데이터 분석 플랫폼 서버 탑재

3. 향후 계획

연구 관리

- 월 1회 진도 점검 세미나 진행: 원내 → 원내 + 원외
  - 프로포절 세미나(4월), 진도점검 세미나(5월) 기 개최
  - 중간자문회의 후속처리(7월 1주차), 진도점검(7월 말, 8월, 9월) 개최 예정

- 기존 연구성과 온라인 공개
  - Homepage: https://keibigdata.github.io/index.html

Homepage

KEI Bigdata Research Team Blog

Bigdata 연구방법론 활용방안

[Bigdata 연구방법론 활용방안] Proposal (pdf)

Posted by **Sung Won Kang** *on April 13, 2017*

[Bigdata 연구방법론 활용방안] Progress Report (May) (pdf)

Posted by **Sung Won Kang** *on May 24, 2017*

딥러닝을 이용한 기후변화에 따른 전염성 질환 발생 패턴 분석

[전염성 질환 발생 패턴 분석] Proposal (pdf)

Posted by **Suna Kang** *on April 13, 2017*

[전염성 질환 발생 패턴 분석] Progress Report (May) (pdf)

Posted by **Suna Kang** *on May 24, 2017*

감사합니다