환경 빅데이터 분석 및 서비스 개발

최종자문회의(2018.10.18)

한국 환경정책·평가연구원

강 성 원

1. 연구 일반

2. 환경 빅데이터 분석

3. 환경 빅데이터 플랫폼

4. 환경 빅데이터 서비스

5. 요약 및 시사점

1. 연구 일반

개관

구분	내용	
연구성격	일반사업(연구형), 계속사업	
연구기간	2018.1 ~2018.12	
연구진	강성원 선임연구위원(책임) 진대용 부연구위원(부책임) 명수정 연구위원 홍한움 부연구위원	한국진 선임전문원 김진형 연구원 김도연 위촉연구원 강선아 위촉연구원 이동현 한국산업기술대 교수(위탁)
자문위원	내부	공성용 선임연구위원 김호정 연구위원 하종식 연구위원 신동원 부연구위원
	외부	김종률 정책관 (환경부 대기환경정책관) 강희찬 교수 (인천대학교 경제학과) 이성호 박사 (대한상공회의소) 오세영 박사 (한국행정연구원)
자문일정	착수자문회의: 2018년 3월 중간자문회의: 2018년 6월 최종자문회의: 2018년 10월	

기간, 인력, 예산

- 기간: 2018년 1월 – 2017년 12월
- 인력: 박사급 연구원 4명(1명 원외), 선임전문원 1명, 연구원 1명, 위촉연구원 2명 투입
- 예산: 2억 7천6백만 원 책정
  - 위탁연구비 4천 만원 책정: '컨벌루션 신경망(CNN)을 통한 미세먼지 예측'
    - 위탁과제 책임자: 한국 산업기술대학교 이동현 교수

환경 빅데이터 연구 목적

◆ **빅데이터 연구 단계**

주제선정 → 자료수집 → 자료분석 → 결과전달

1. 환경빅데이터 연구
   - 주제선정 ~ 자료 분석 단계를 적용한 환경연구 수행
2. 환경 빅데이터 인프라
   - 자료수집~자료분석 단계를 수행할 수 있는 작업 공간 구축
3. 환경 빅데이터 서비스
   - '결과 전달' 단계를 확장하여 수요자 중심 서비스 개발

연속사업: 3년 단위 연구단계 설정

- 1단계(2017-19): 환경 빅데이터 연구 시작/ 환경 빅데이터 분석 플랫폼 설계
- 2단계(2020-22): 환경 빅데이터 분석 플랫폼 구축/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**	• **환경 빅데이터 분석 플랫폼 설계**	• 원내 연구정보 서비스
2기 (2020-22)	• 발신주기 단축	• 환경 빅데이터 분석 플랫폼 구축	• 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

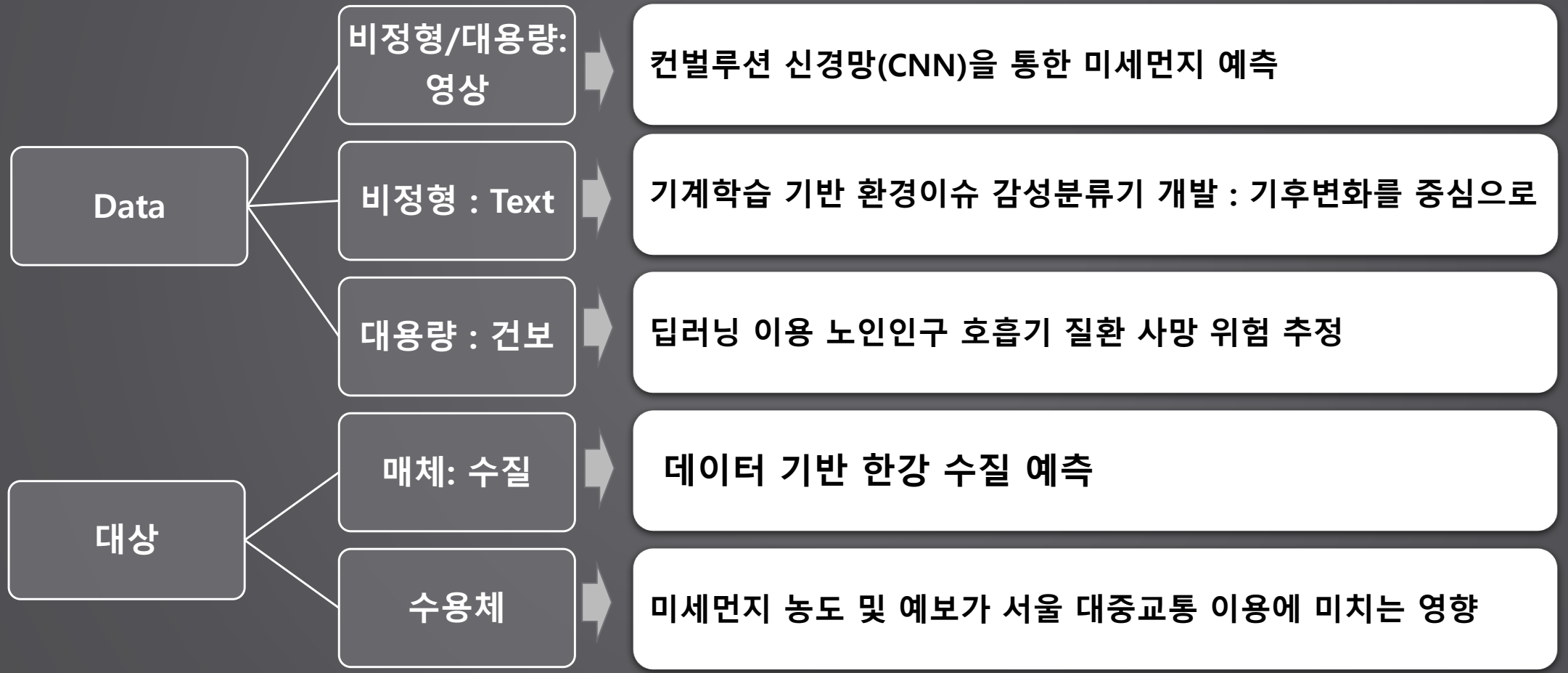
2017-19년 연차계획

	환경 빅데이터 연구	환경 빅데이터 연구 인프라	원내외 빅데이터 서비스
1단계	**환경 빅데이터 연구 시행**	**환경 빅데이터 플랫폼 설계**	**원내 연구정보 서비스**
2017	• 환경연구 알고리듬 개발 - 전산화된 자료 + Deep Learning	• 환경분야 기초데이터 수집방법 • 자료 및 알고리듬 축적/공개	• 연구동향 파악 서비스
2018	• 환경연구 알고리듬 개발: - 비정형자료 + Deep Learning	• 환경 빅데이터 플랫폼 설계 - 대용량 자료 저장-분석 기능 구비 - 연구결과 자료 및 알고리듬 공유 - 환경 기초데이터 수집 결과 축적	• 연구동향 파악 서비스 원내 • 환경 데이터 포털(Open Data Map) 구축
2019	• 환경연구 알고리듬 개발 지속 • 딥러닝 기반 연구수요 분석 상시화	• 환경 빅데이터 플랫폼 설계 완료 - 연구결과 자료 및 알고리듬 공유 지속 - 환경분야 기초데이터 수집 1단계 완료	• 연구동향 파악 서비스 원외공개 • 환경 데이터 포털(Open Data Map) 원내 공개
2단계	**발신주기 단축**	**연구 과정 자동화/플랫폼 구축**	**연구기획 서비스/공공 서비스 설계**
3단계	**시의성 중심 발신체계**	**분석 플랫폼 지능화 시도**	**공공 서비스 시범 사업**

2018년 연구목표

- 환경 빅데이터 연구 : 비정형 대용량 자료 분석 및 연구영역 확대
  - 비정형, 대용량 자료 분석 : 화상(Image) 분석, 건강보험 자료, SNS 자료 분석
  - 연구영역 확대: 수질오염 예측 및 환경위험에 대한 수용체 반응 분석
- 환경 빅데이터 플랫폼 설계: 대용량 자료 수집-분석 환경 구축
  - 자료 수집 : 온라인 데이터 목록 및 Link 제공 환경 데이터 안내지도(Open Data Map) 구축
  - 자료분석: 대용량 자료 분석 및 알고리듬 개발 환경 제공
- 환경 빅데이터 서비스 개발 : 연구동향 서비스 개발
  - 연구동향 서비스 : 연구주제 동향 파악, 연구 키워드 동향 파악
    - 2017년 연구 '텍스트마이닝을 이용한 KEI 연구동향 분석' 개발 LDA 토픽 클러스터링 알고리듬 및 키워드 분석 알고리듬 활용

환경 빅데이터 연구영역 확대



중간 자문회의 자문의견 반영

1. 감성분류기 성능 제고 및 활용 방안 : 전처리를 강화하여 80% 이상 정확도 확보
2. 정책적 활용 방안 고민 : 결론 부분에 명기
3. 과제 별 제안 : 선별적으로 수용

	자문의견	반영내역
이성호	- 감성분류 분석 시 정확도를 높이는 방안으로 4명이 크로스 체크한 결과를 바탕으로 점수를 부여하여 훈련시키면 좋을 것 같음	- 전처리 강화 후 정확도가 80% 이상으로 제고
오세영	- 기존 분석방법과 딥러닝을 이용한 연구의 차별점을 설명할 필요 있음 - 분석결과에 대하여 정책적 활용 측면에서 고민할 필요 있음 - 감성분류기의 경우 정책의제 형성에 기여할 수 있음, 이에 대한 검토필요 - (CODP) 지역 낙인효과와 밀접한 연관이 있을 수 있으므로 이부분에 대한 설득 논거 개발 필요	- Reference Point로 기존 연구에서 자주 활용하는 통계적 방법론(Regression/ARIMA)과 결과를 비교 - 요약 및 시사점에서 정책활용 부분에 반영 - 요약 및 시사점에서 정책활용 부분에 반영 - 개인을 인구학적 특성으로 분류하여 분석한 결과에 초점을 맞출 예정
하종식	- COPD 사망영향 분석과 관련하여 분석 공간단위가 구체적인지 않음(7개 주요 광역시 또는 시군구 등 구체화 필요) - 분석하고자 하는 공간 및 시간단위를 결정하여 이를 고려한 설명변수 및 결과변수 자료 구축 필요 - HEAT package는 폭염 또는 고온 노출 관련한 건강영향 평가package이므로, 대기오염으로 인한 COPD 사망영향 분석 시 고려할 수 있는 혼란변수 선정 시 package의 활용 적절성 검토 필요	- 공간 단위가 아닌 개인 단위 분석이며 개인의 주거지(시군구)가 대기오염-기후 자료와 연계하는 단위 - 시간단위는 주단위이며 공간이 아닌 개인이 분석 대상 - 분석 방법으로는 logistic Regression을 사용하였으며 변수 선정 과정에서는 의료계 전문가들의 자문을 구함

중간 자문회의 자문의견 반영

	자문의견	반영내역
김호정	- 분석 대상 변수에서 부영양화 지수는 많이 이용하는 변수가 아님, 연구에 클로로필-a 변수를 이용하는 것이 좋을 것으로 판단됨 - 연구에서는 정확도가 많이 떨어져도 실시간 측정망 자료를 이용하는 것이 좋아 보임 - 독립변수에서 강우량보다 일조량이 더 영향이 있을 것으로 보임 - 분석 대상 지역을 좁히면, 더욱 의미있는 결과가 나올 것 같음, 예를 들어 주민들이 관심이 높은 지역을 분석	- 클로로필-a 를 분석 - 실시간 측정망 자료 이용 예정 - 독립변수 선정 시 반영 - 팔당(상수원), 노량진(도심), 가양(수변위락지역) 선정
신동원	- 보고서 서론에 본 연구에서 사용한 방법론들이 기존의 방법론과 어떠한 차이점이 있는지 설명할 필요 있음 - 측정소에서 측정한 수치가 비가 왔을때 신뢰도 문제가 생김, 따라서 time lag를 이용한 방법 등을 통해 신뢰도 문제를 해결하는 방법을 고민해야 함 - (ICNN) 고농도 사례의 경우를 분리하여 분석한다면, 재밌는 결과가 나올 것 같음 - (감성분류기) 국내 SNS 유저들의 사용패턴을 분석에 반영하면 더 정확한 분석이 될 것으로 보임	- 요약 및 시사점에서 언급하였으며 서론에 보완할 예정 - 센서 데이터 자료 축적이 충분하지 않아서 금년도 보고서의 부록에 수록하기로 잠정 결정 - 예측 시계가 6시간 이내일 경우 고농도 지역을 정확하게 분별 - 감성분류기 정확도 80% 달성

빨간색 : 미반영
파란색 : 자주 언급된 내용

연구 진행 상황 요약

- 90% 기준으로 1개 세부과제 진도 check 필요: 데이터 수집 과정에서 예정 이상 시간 소요
  - 2개 연구과제, 2개 서비스, 2개 플랫폼 구성요소 연구목표 100% 달성, 2개 연구과제 90% 달성

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. 환경 빅데이터 인프라 구축	1) Open Data Map									후속 조치	
	2) 빅데이터 분석 플랫폼									후속 조치	
3. 환경 빅데이터 연구	1) 컨벌루션 신경망(CNN)을 통한 미세먼지 예측									후속 조치	
	2) 데이터 기반 한강 수질 예측										
	3) 딥러닝 이용 국내 노인인구 호흡기 질환 사망 위험 추정										
	4) 기계학습 기반 환경이슈 감성분류기 개발 : 기후변화를 중심으로									후속 조치	
	5) 미세먼지 농도 및 예보가 서울 대중교통 이용에 미치는 영향										
4. 환경 빅데이터 서비스	연구동향 파악 서비스									후속 조치	
5. 결론	연구결과 요약 및 시사점										

2. 환경 빅데이터 분석

- 컨벌루션 신경망(CNN)을 통한 미세먼지 예측
- 기계학습 기반 환경이슈 감성분류기 개발 : 기후변화를 중심으로
- 데이터 기반 한강 수질 예측
- 딥러닝 이용 노인인구 호흡기 질환 사망 위험 추정
- 미세먼지 농도 및 예보가 서울 대중교통 이용에 미치는 영향

연구 진행 상황 : 4건 결과 도출, 1건 시범 분석

주제선정

자료수집

자료분석

결과발신

- ◆ 컨벌루션 신경망(CNN)을 통한 미세먼지 예측 [1.65 GB]
- ◆ 데이터 기반 한강 수질 예측 [0.344MB]
- ◆ 기계학습 기반 환경이슈 감성분류기 개발 : 기후변화를 중심으로 [3.0 MB]

- ◆ 미세먼지 농도 및 예보가 서울 대중교통 이용에 미치는 영향 [1.04GB]
- ◆ 딥러닝 이용 노인인구 호흡기 질환 사망 위험 추정 [1TB]

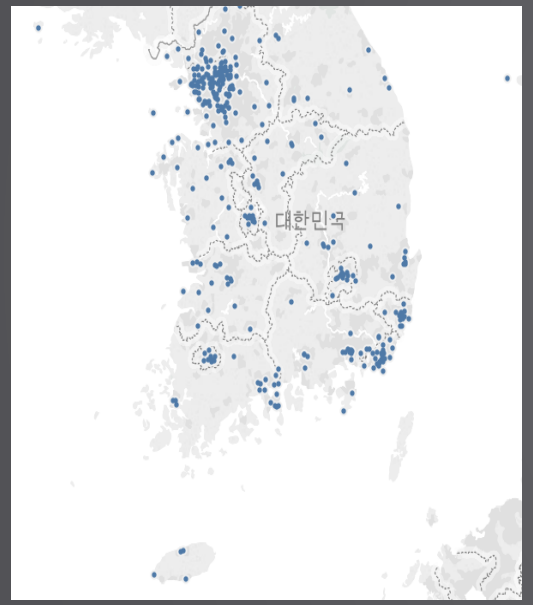
(1) 컨벌루션 신경망(CNN)을 통한 미세먼지 예측 [이동현]

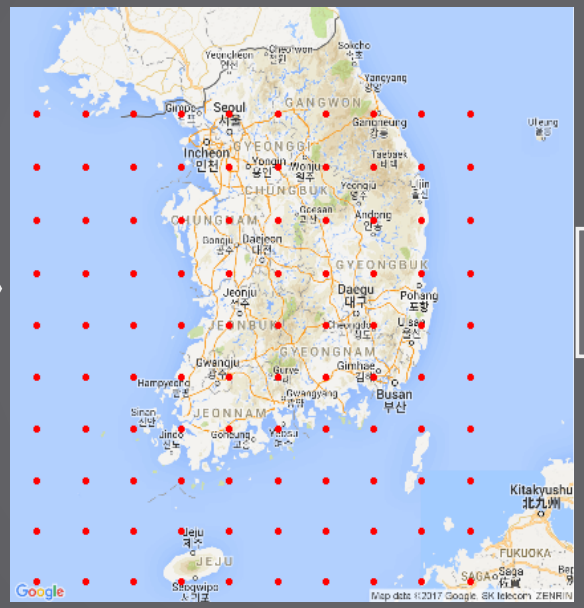
- 주변 지역 정보를 반영하여 대기오염을 예측하는 컨벌루션 신경망 모형
  - 전국 측정소 별 미세먼지를 거리를 반영하여 10 x 10 격자로 보간(IDW)하고 컨벌루션 신경망(CNN: Convolution Neural Network)을 적용
  - 위도, 경도, 대기 및 기상자료, 미세먼지 오염도 → 미래 미세먼지 오염도 추정

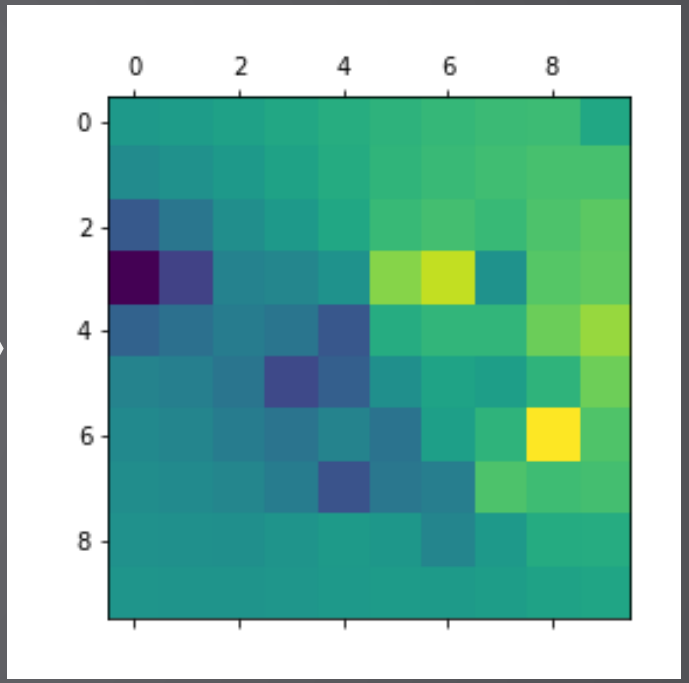
- 연구 내용 : 자체 개발 ICNN(Interpolation CNN) 활용 미세먼지 오염도 예측
  - CNN Algorithm : 동일지점 과거정보 집적 + 주변공간 정보 반영 + 여타 변수 정보 반영
    - 첫 번째 아키텍처 (Arc1) : 이전 7시간의 PM10으로 다음시간의 PM10을 예측
    - 두 번째 아키텍처 (Arc2) : 이전 1시간의 요인들로 다음 1시간의 PM10을 예측
    - **세 번째 아키텍처 (Arc3) : 이전 7시간의 요인들로 다음 1시간의 PM10을 예측**
    - 네 번째 아키텍처 (Arc4) : 이전 7시간의 요인들로 다음 1시간의 지역별 PM10을 예측
    - 다섯 번째 아키텍처 (Arc5) : Inception 모듈 응용, 이전 7 시간의 요인들로 다음 1시간의 PM10을 예측
  - Stochastic Gradient Descent (SGD) + Adaptive Moment Estimation (ADAM) 최적화 기법 이용

- 연구 성과 : 1시간 예측치 평균제곱근오차(RMSE) 2.07 $\mu g/m^{3,}$ 8 시간 예측치 9.09 $\mu g/m^{3,}$

데이터 변환 : 측정소 자료를 격자형 자료로 보간





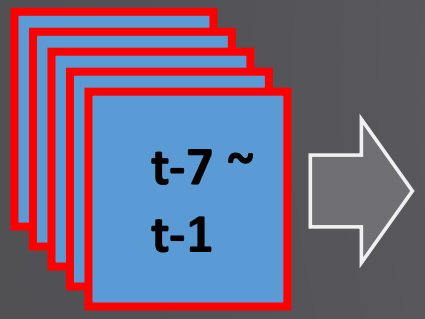


데이터 격자 보간

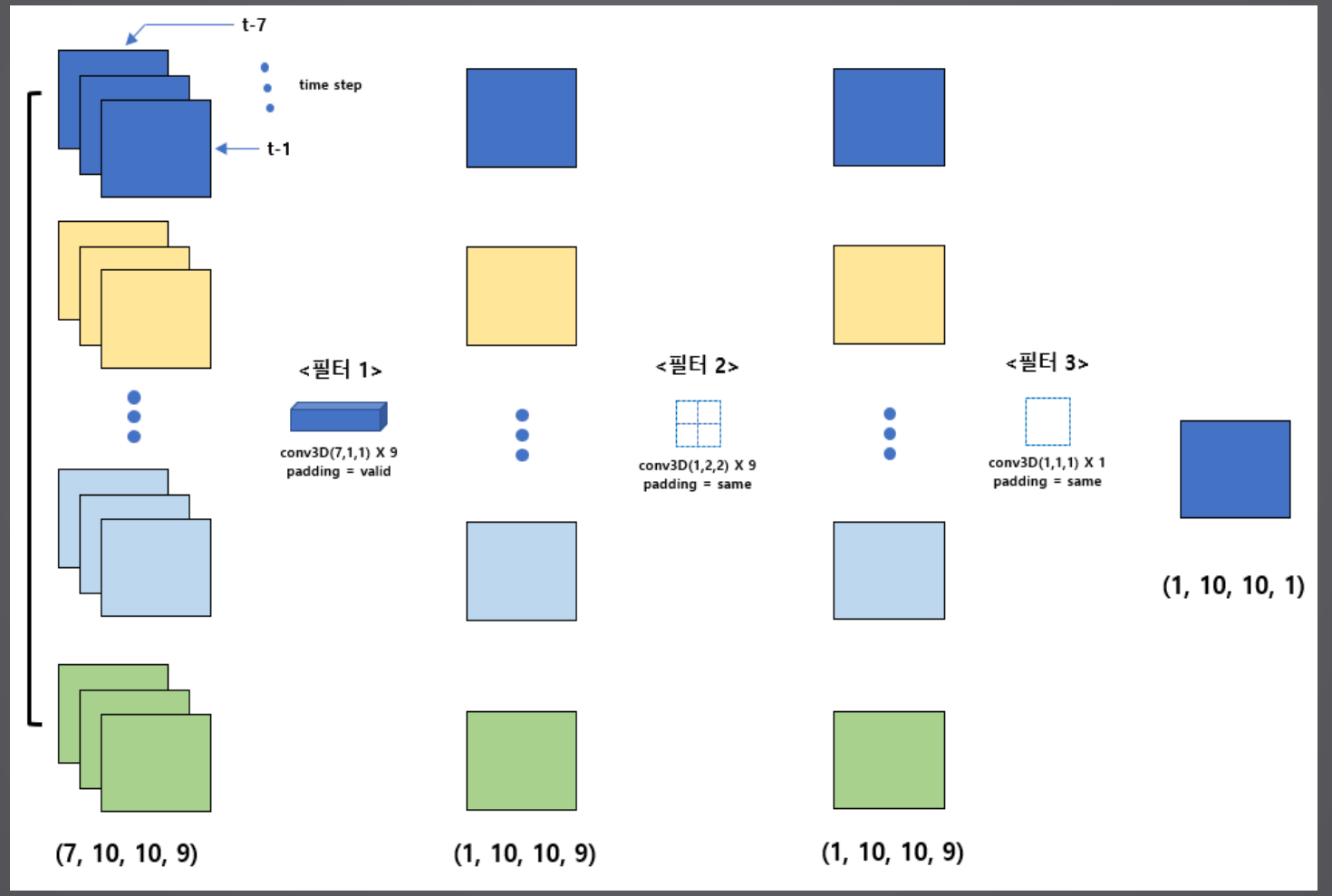
(IDW: Inverse Distance Weighted)

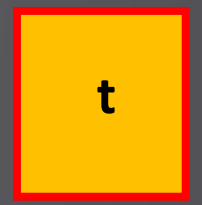
ICNN (Interpolation CNN) 을 활용한 예측

CNN 알고리듬 (Arc 3)



$x_1 \sim x_9$:
SO2, CO, O3, NO2,
Temp,
Precipitation,
Wind_Speed,
Wind_Direction,
PM10





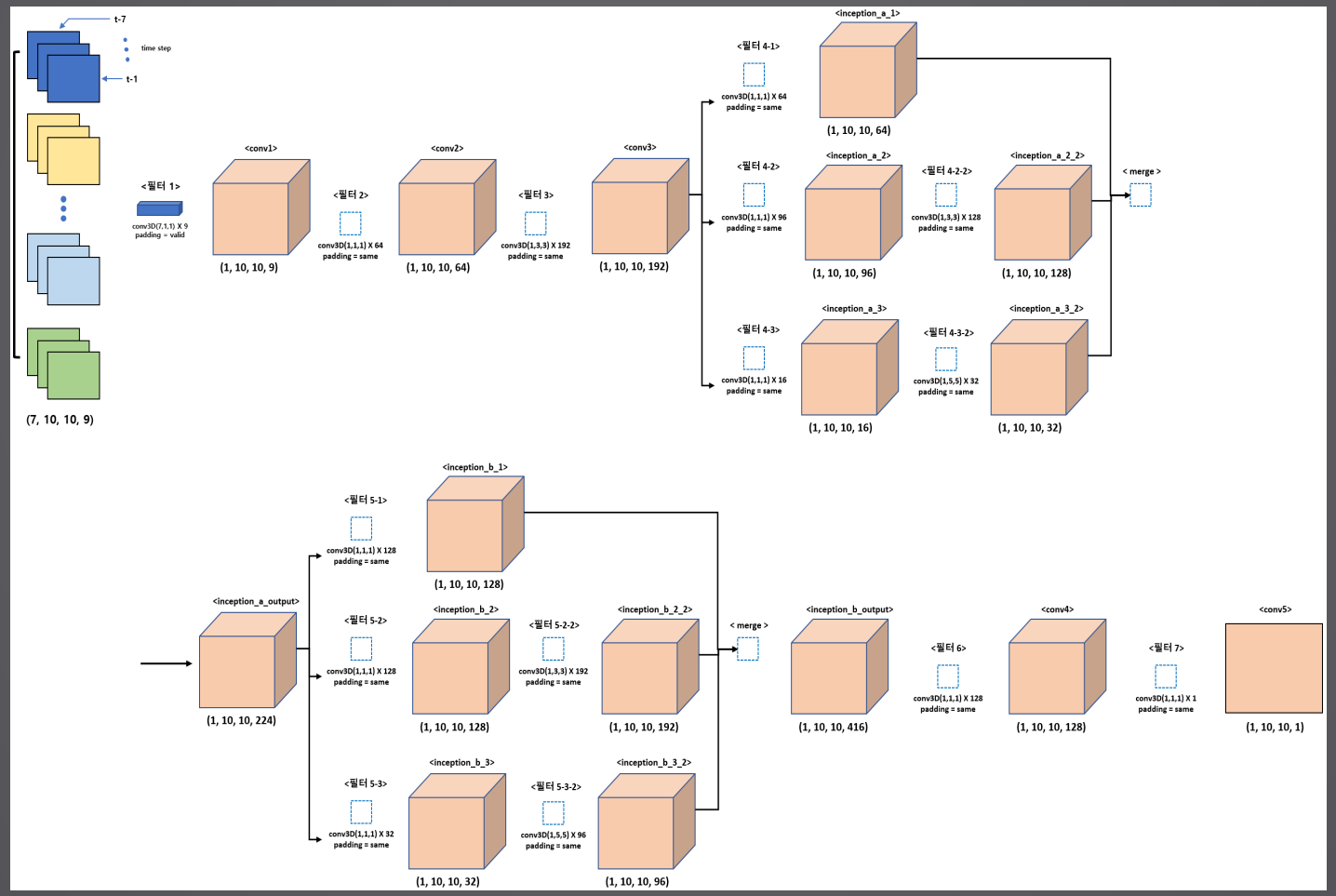
Y:
PM10

ICNN (Interpolation CNN) 을 활용한 예측 : Inception

CNN 알고리듬 (Arc 5)

t-7 ~ t-1

$x_1 \sim x_9$:
SO2, CO, O3, NO2, Temp, Precipitation, Wind_Speed, Wind_Direction, PM10



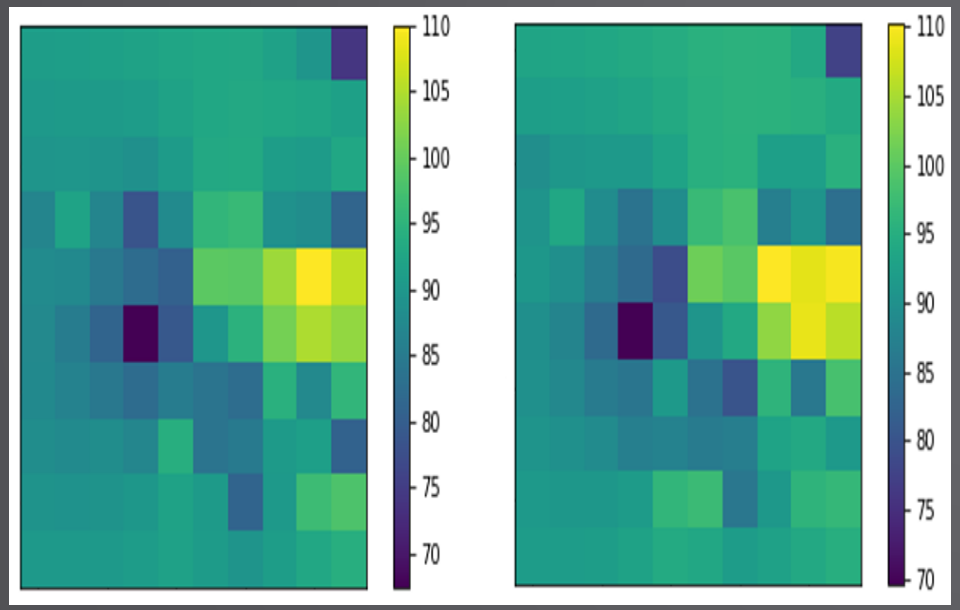
작은 필터를 여러 번 다양한 경로로 사용하여 다각도로 정보를 추출

t

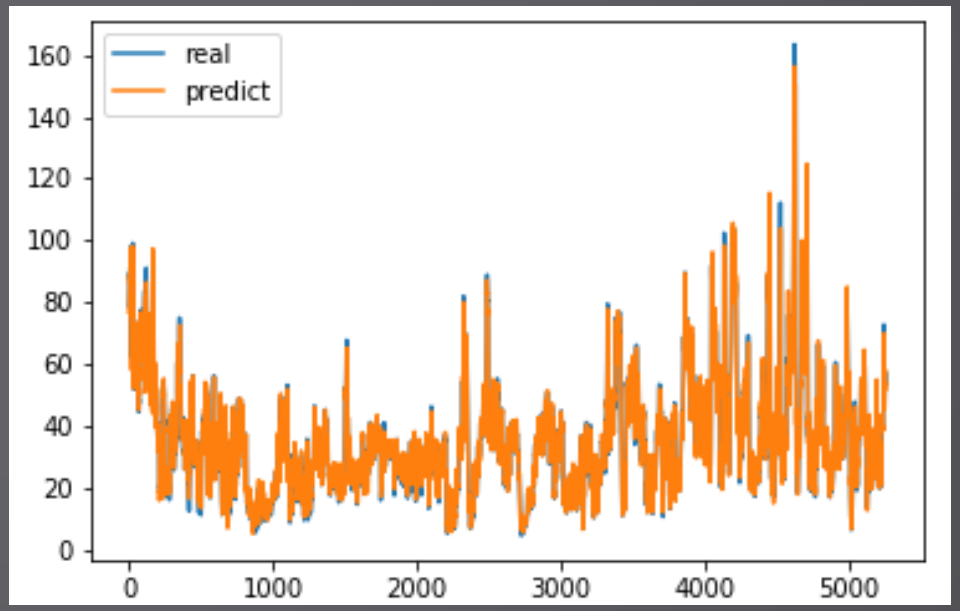
Y:
PM10

미세먼지오염도 예측 결과 (Arc3: 7시간 전 모든 변수)

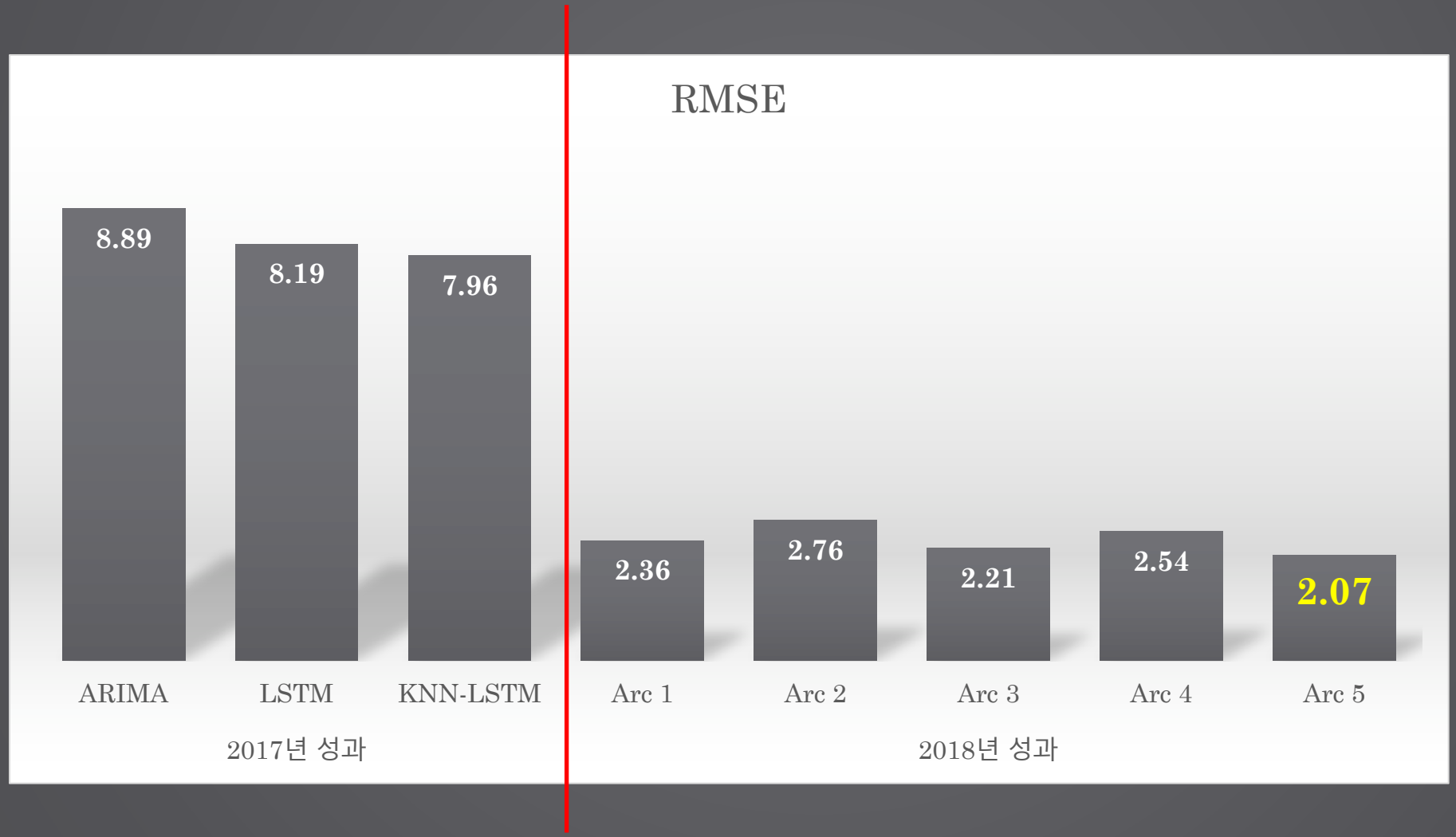
예측 이미지(좌측) vs. 실측 이미지(우측)



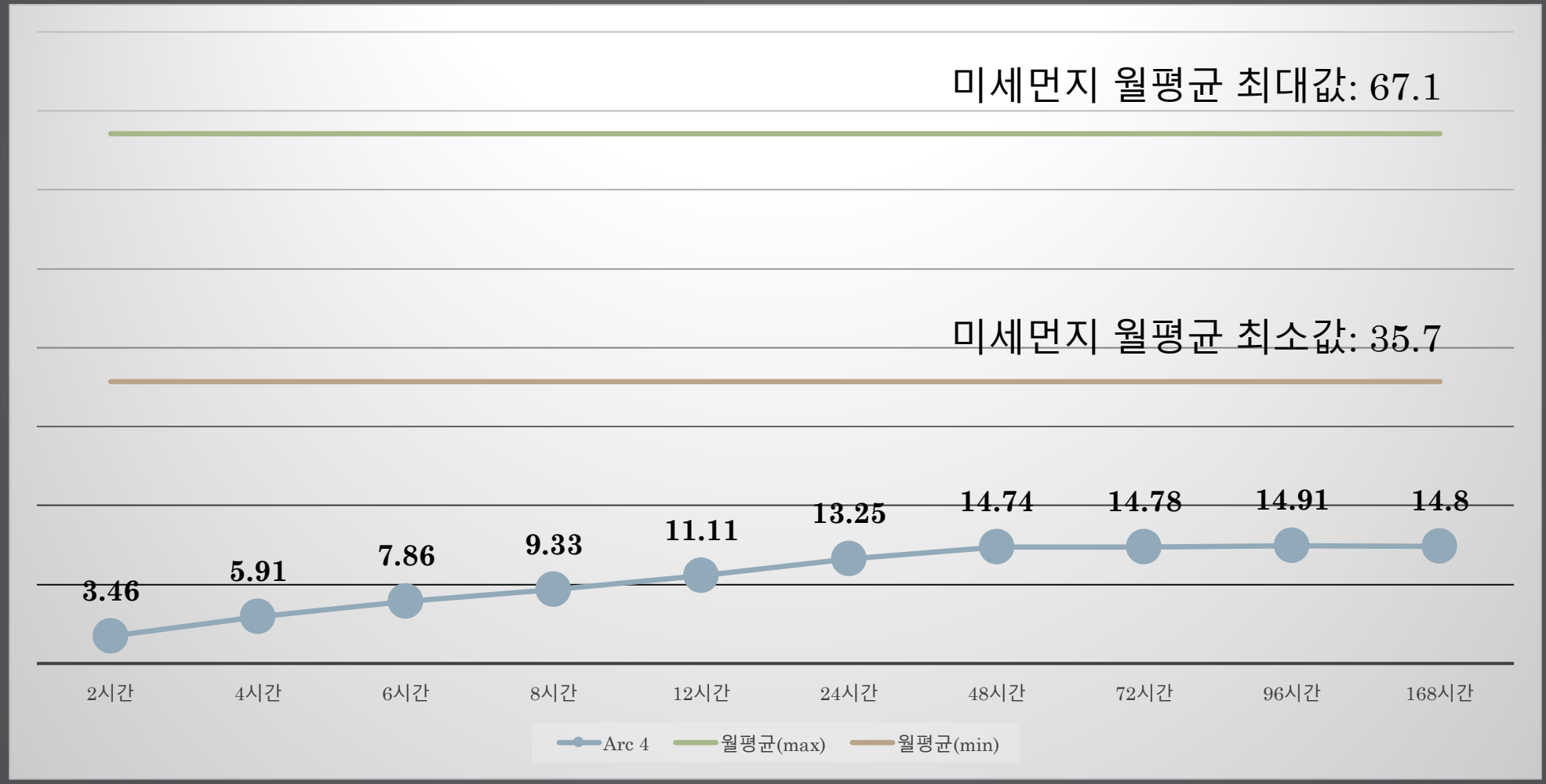
미세먼지 농도 예측-실측치 (4,4 지역)



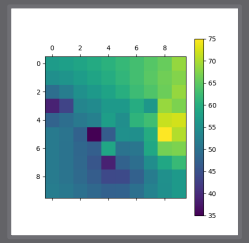
미세먼지 예측 평균제곱근오차(RMSE) : 1시간



미세먼지 예측 평균제곱근오차(RMSE) : 2시간 이상

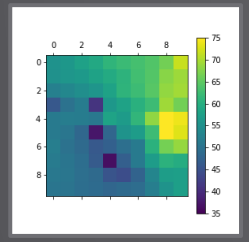


미세먼지 예측 이미지 : 2시간 이상

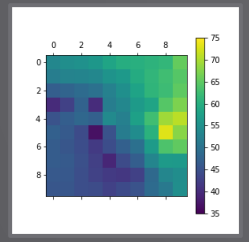


<Real>
2016.12.31 23:00

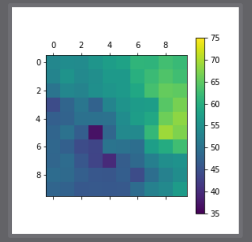
<Lag-2>



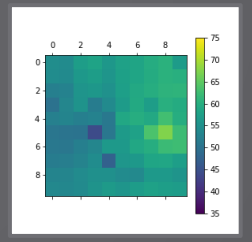
<Lag-4>



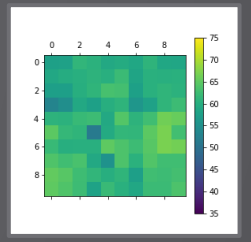
<Lag-6>



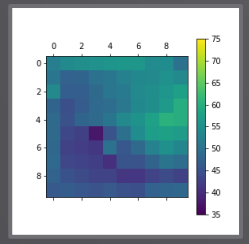
<Lag-8>



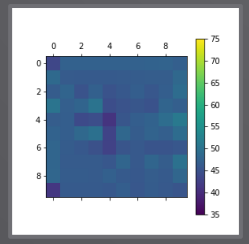
<Lag-12>



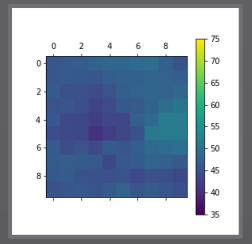
<Lag-24>



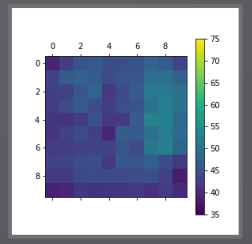
<Lag-48>



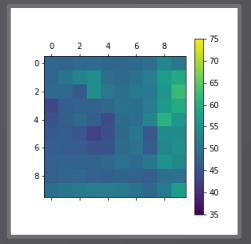
<Lag-72>



<Lag-96>



<Lag-168>



(2) 데이터 기반 한강 수질 예측 [홍한움]

◆ 인공지능 및 통계모형을 이용한 데이터 기반 수질 예측 알고리듬 개발

◆ 연구내용: 측정소 및 인근지역 수질, 수위, 기상자료에 RNN 적용 클로로필-a(Chl-a) 농도 예측

- 분석 대상: 수도권 수질 측정 지역 중 3개[가양, 노량진, 팔당] 지역 2008-17 주간 자료
  - 수질 일반측정망 자료(물환경정보시스템), 기상자료(기상자료개방포털) , 수위자료(한강홍수통제소)수집

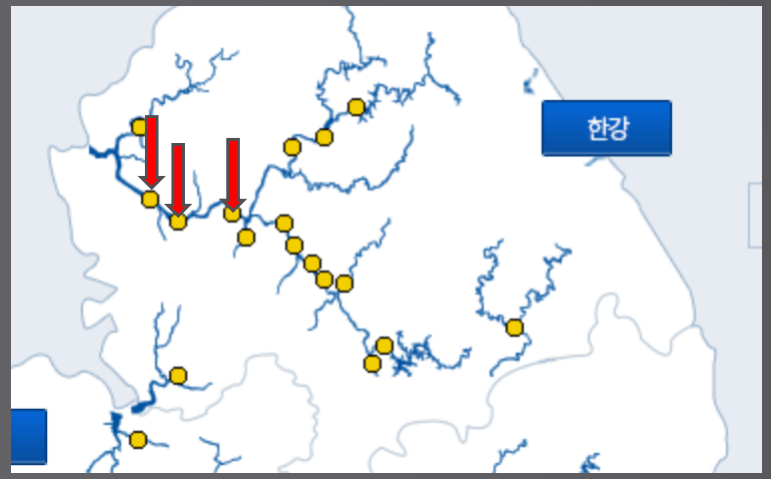
- 인근지역 정보를 시차를 두고 반영하여 예측 정확도 제고
  - 과거 정보 반영: RNN(Recurrent Neural Network)
  - Long Memory 특성 반영: LSTM(Long Short Tem Memory) GRU(Gated Recurrent Unit)

◆연구결과 : GRU 알고리듬을 사용하여 선형회귀분석 및 VAR 대비 35.9% 평균제곱근 오차 향상

- 극한값 예측에서 선형회귀분석 및 VAR의 지연예측 현상 개선

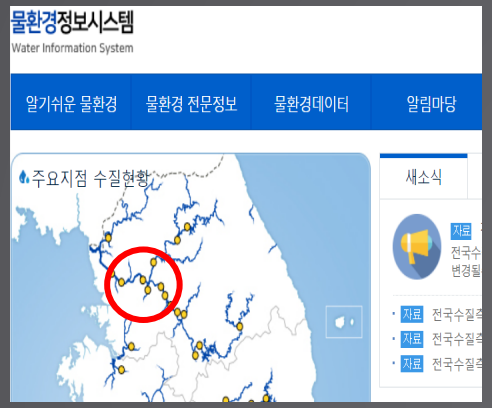
분석 대상 관측소

- 3개 한강 수질측정 지점 : 가양, 노량진, 팔당
  - 가양: 수변공간 활용지역
  - 노량진: 서울 중심지역
  - 팔당: 한강 상수원
  - 여름철 한강 물놀이 지점 포괄
- 자료 범위: 2008년 1월 – 2017년 12월
  주간
  - Training : 2008.01.10-2017. 08. 05 400개
  - Validation : 2013.05.18-2017. 07. 29 120개
  - Test set: : 2017.08.05-2018. 12. 30 120개



수질 관측지점: 왼쪽부터 순서대로 가양, 노량진, 팔당댐
(출처: 물환경정보시스템 메인페이지)
http://water.nier.go.kr/main/mainContent.do

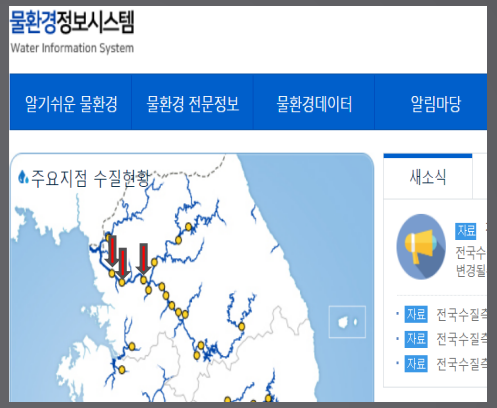
Input변수



현재 및 과거 Chl-a

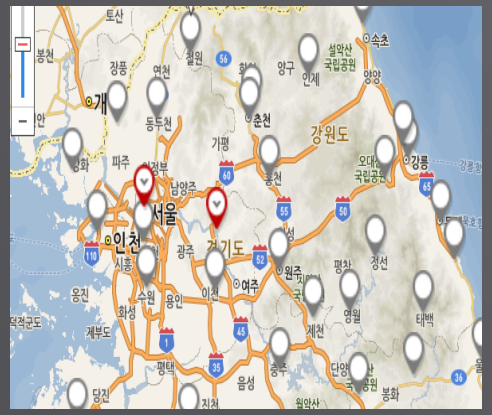
예측&상류 지점 과거 Chl-a

(가양, 노량진, 팔당 & 삼봉리, 경안천5, 강상, 강천)



예측지점 현재 및 과거 수질

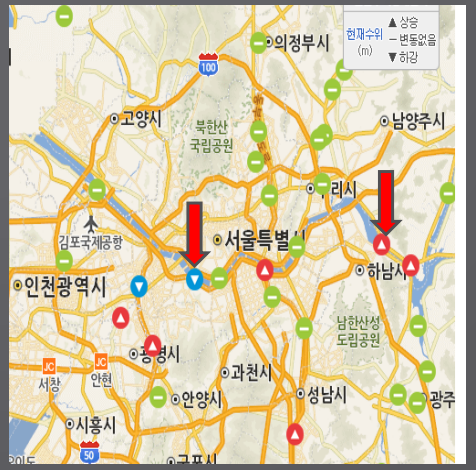
pH, DO, BOD, TN, TP, 수온, …, 전기전도도 총대장균군수 분원성대장균군수



기상 자료

기온, 강수량, 풍속, 습도, 기압, 일사량

(서울, 양평 관측소)



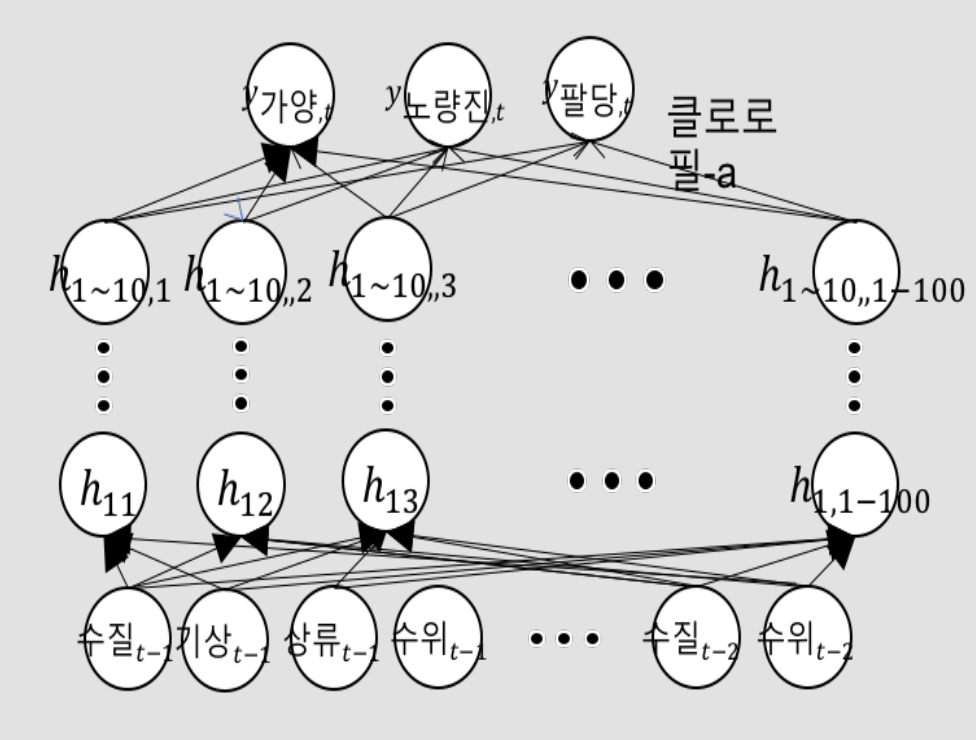
수위, 유량

팔당대교: 수위, 유량

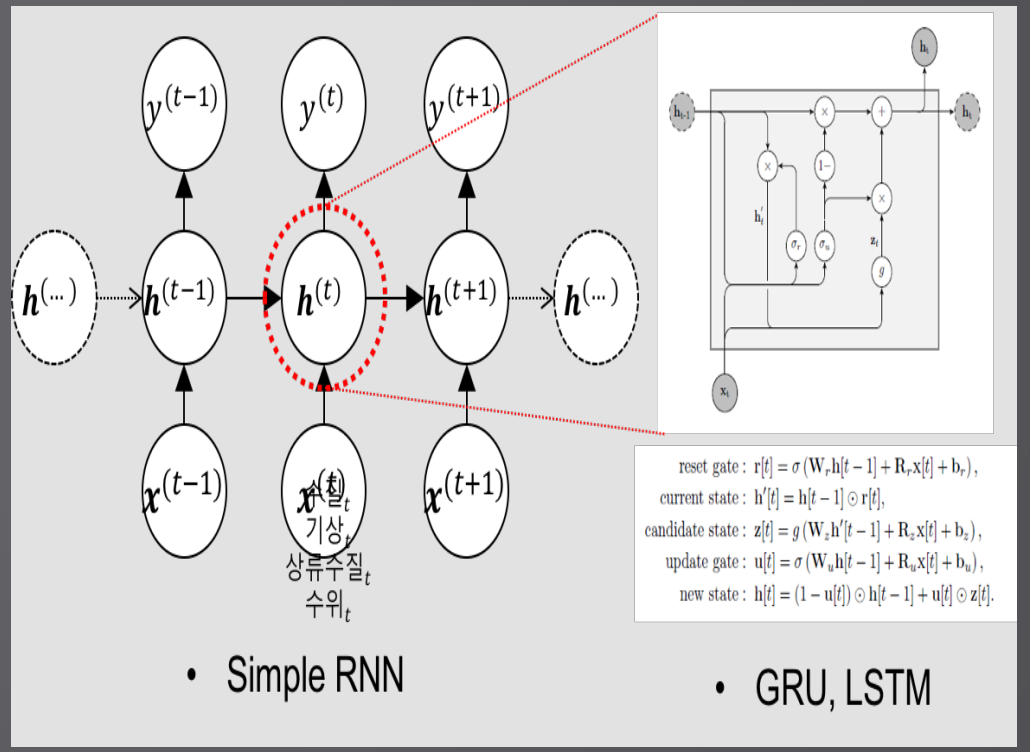
한강대교: 수위, (유량은 2008 년 자료가 없어 제외)

분석 알고리듬 : 선형회귀분석, MLP, RNN

MLP



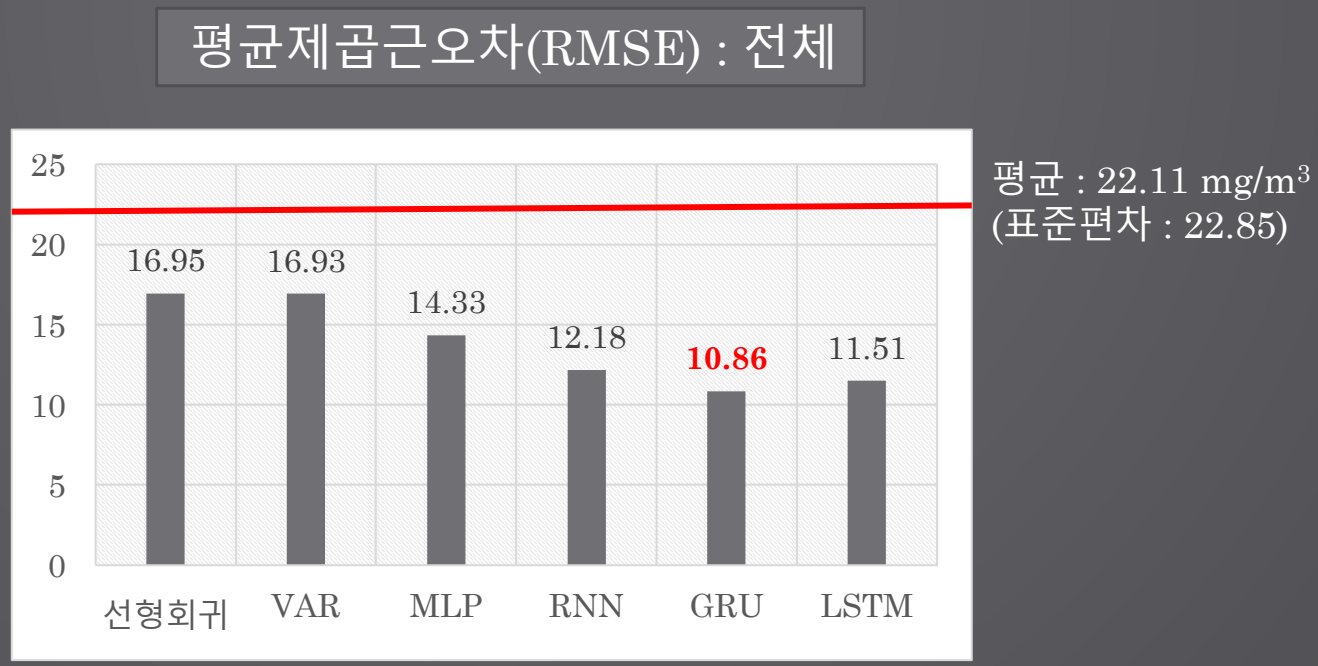
RNN, GRU, LSTM

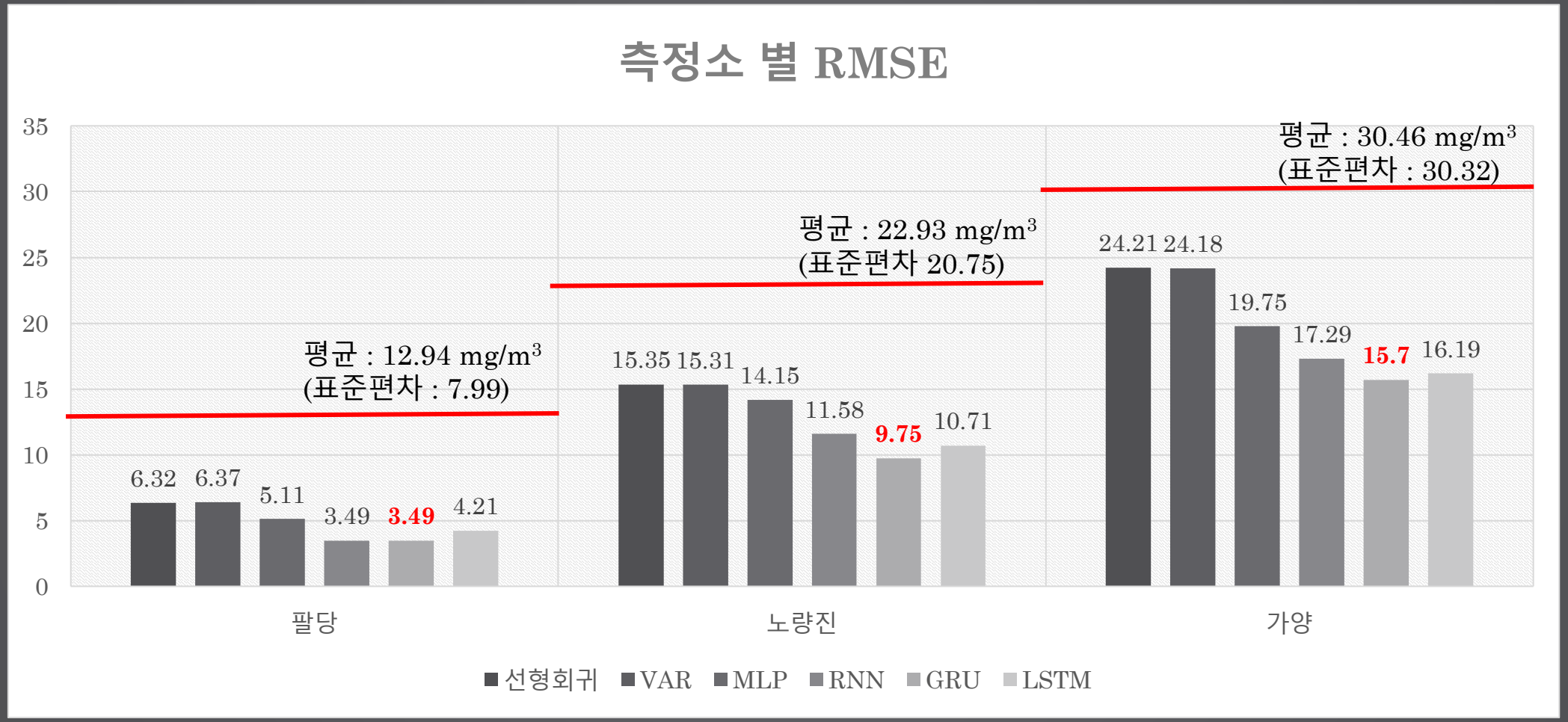


- Simple RNN
- GRU, LSTM

예측 결과 : RNN 계열 예측오차 축소 효과 탁월

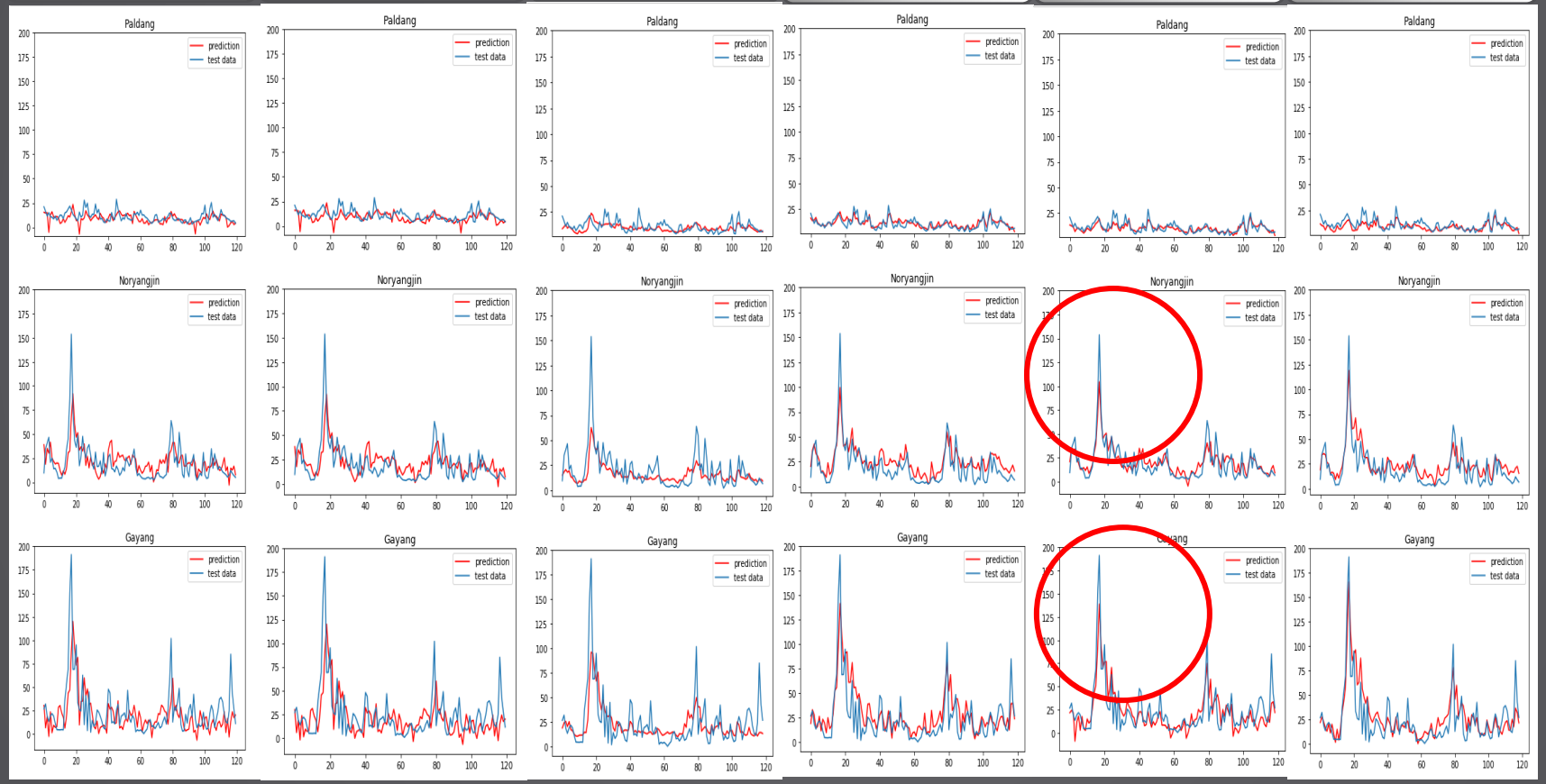
◆ Chl-a 의 변화 추세와 큰 값을 잘 예측함
- 회귀분석, VAR(Vector Autoregression)의 지연예측과 비교됨
- 데이터가 더 모일 경우 RMSE를 더 줄일 수 있을 것으로 기대
  - 학습을 진행하며 Validation set의 비율을 줄이면서 training set 의 크기를 조금씩 늘릴 때마다RMSE가 크게 축소
- 큰 값의 예측 성능이 좋으므로 녹조 예측에 활용 가능




측정소 별 RMSE
35
30
25
20
15
10
5
0
평균 : 12.94 mg/m3
(표준편차 : 7.99)
평균 : 22.93 mg/m3
(표준편차 20.75)
평균 : 30.46 mg/m3
(표준편차 : 30.32)
6.32
6.37
5.11
3.49
3.49
4.21
15.35
15.31
14.15
11.58
9.75
10.71
24.21
24.18
19.75
17.29
15.7
16.19
팔당
노량진
가양
선형회귀
VAR
MLP
RNN
GRU
LSTM

수도권 수질예측 결과

회귀분석	VAR	MLP	RNN	GRU	LSTM
16.95	16.93	14.33	12.18	10.86	11.51

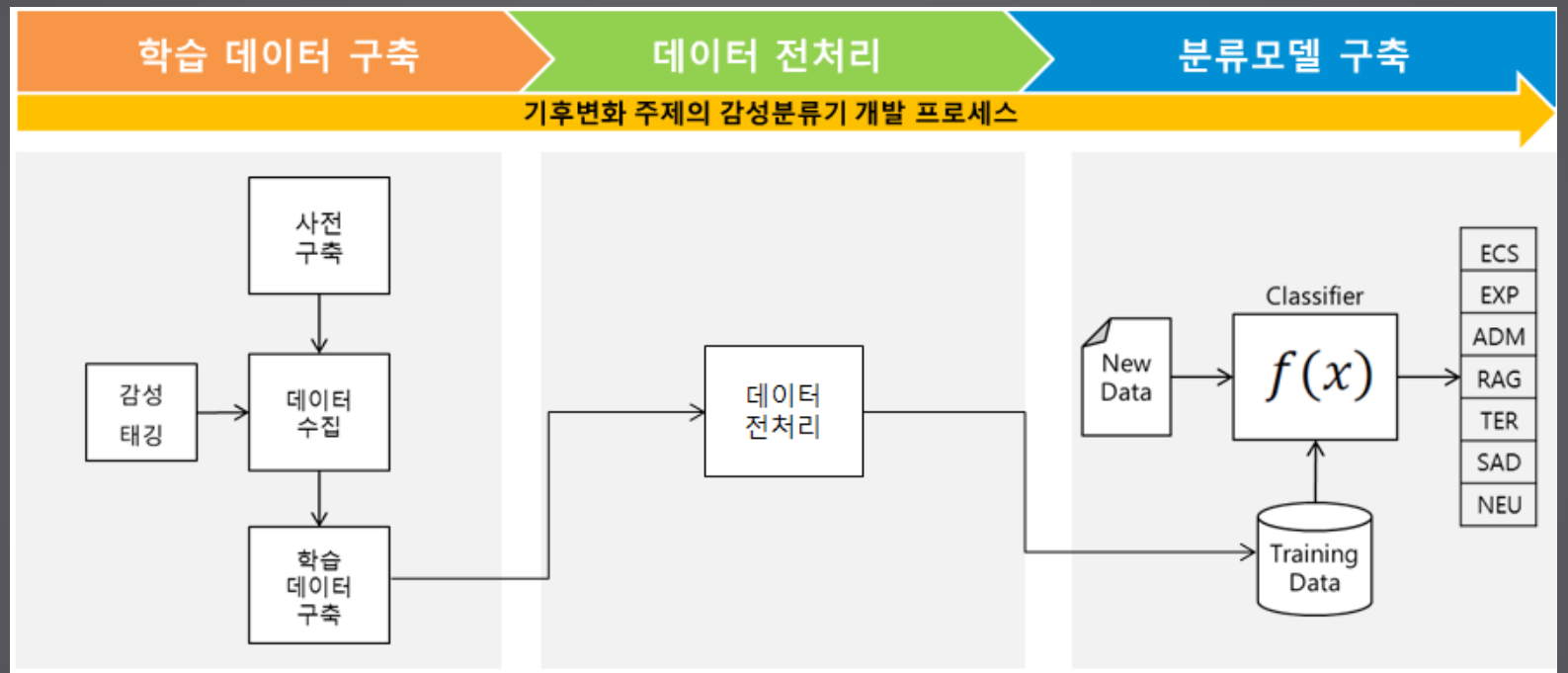


RMSE

(3) 기계학습 기반 환경이슈 감성분류기 개발 : 기후변화 중심으로 [김도연]

- 연구 목적: 기후변화 주제의 SNS 및 뉴스 댓글 데이터 기반 감성분류 알고리듬 개발
- 연구 내용: 기후변화 사전 구축, 감성 분류 학습 데이터 구축, 감성분류 알고리듬 개발
  - **기후변화 사전** : 기후변화에 따른 현상을 4개의 범주(온도, 강수, 토지, 해양) 분류 후 구축
    - 환경관련 문서에 워드 임베딩 방법(LDA, Word2Vec) 적용 후보군 추출
    - 전문가(최희선,명수정) 및 SNS 이용자 의견 반영
  - **감성분류 기준표** : 기후변화 현상에 자주 나타나는 7개 감성 클래스 구축
    - 7개 감성 카테고리 : 황홀/기쁨, 기대/관심, 감탄/존경, 분노/짜증, 두려움/공포, 슬픔/수심, 중립
  - **감성분류 학습데이터** : 5만 건 단문 데이터에 감성을 수작업으로 파악
    - 기후변화 사전 기준 5만건을 수집하여 7개의 감성 클래스 태깅
  - **감성분류 알고리즘** : 기계학습 기반 분류 알고리듬 구축
    - Naïve Bayes, SVM, CNN, Bi-directional LSTM
- 연구 성과 : 7개 감성 Category 대상 정확도 85.10% Bidirectional LSTM 알고리듬 구축
  - Accuracy : (True Positive + True Negative)/ALL sample

연구 범위 및 흐름도



- 기후변화 사전구축
- SNS 및 댓글 데이터 수집
- 7가지 감성 클래스 구축
- 학습데이터에 감성 태깅
- 감성 태깅 크로스 체크

- 이모지 한글화
- 형태소 분석
- 정규화
- 말뭉치 생성
- Sparse Terms 및 low TF-IDF 삭제

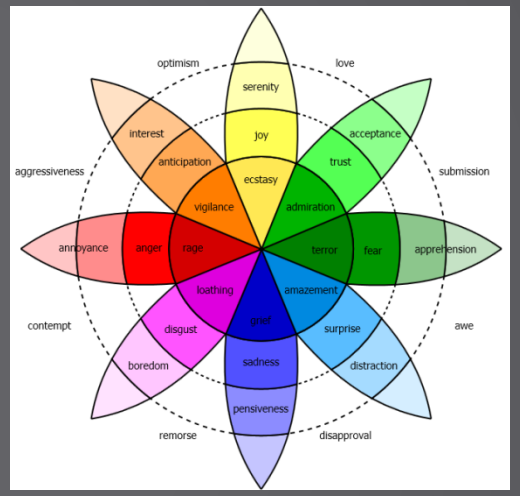
- 학습 데이터 및 검증 데이터 분류
- 기계학습 기반의 분류분석
  - SVM, Naive Bayes
  - CNN, BI-LSTM
- 분류모델 성능 비교 분석

기후변화 사전 및 감성 클래스

기후변화에 따른 현상 사전

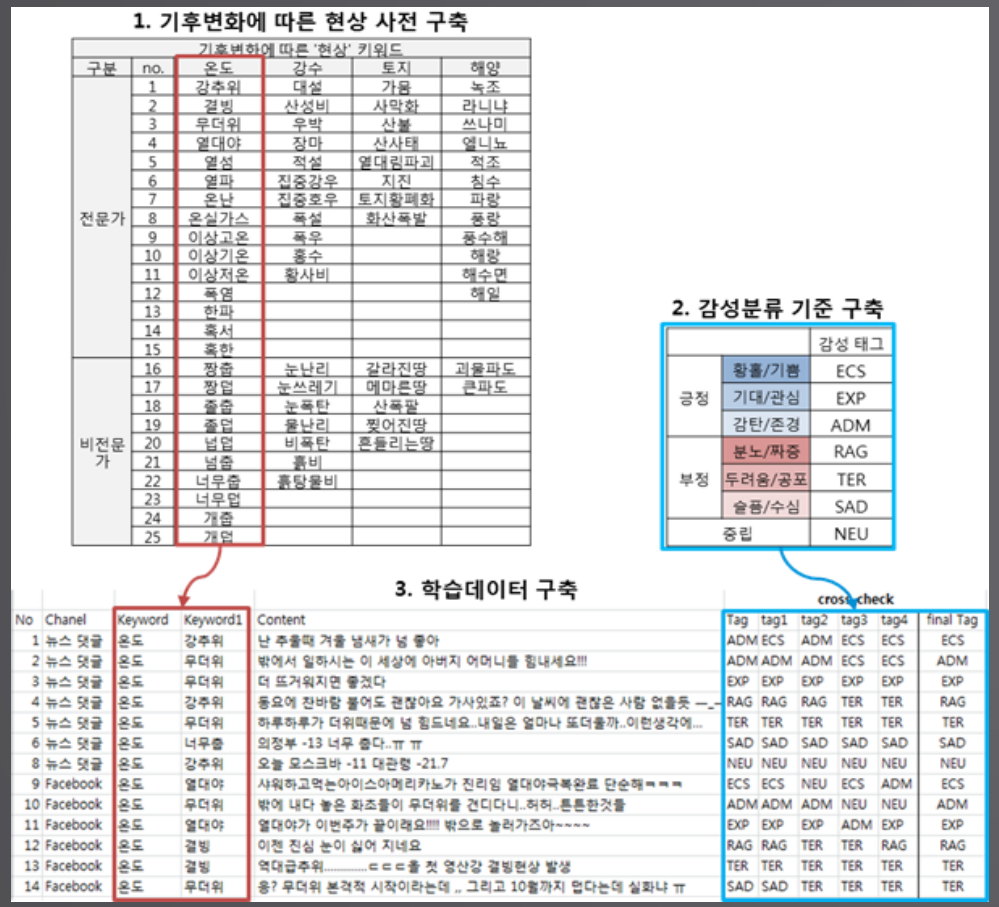
구분	순번	온도	강수	토지	해양
전문가	1	강추위	대설	가뭄	녹조
	2	결빙	산성비	사막화	라니냐
	3	무더위	우박	산불	쓰나미
	4	열대야	장마	산사태	엘니뇨
	5	열섬	적설	열대림파괴	적조
	6	열파	집중강우	지진	침수
	7	온난	집중호우	토지황폐화	파랑
	8	온실가스	폭설	화산폭발	풍랑
	9	이상고온	폭우	-	풍수해
	10	이상기온	홍수	-	해랑
	11	이상저온	황사비	-	해수면
	12	폭염	-	-	해일
	13	한파	-	-	-
	14	혹서	-	-	-
	15	혹한	-	-	-
비전문가	16	짧춥	눈난리	갈라진땅	괴물파도
	17	짧덥	눈쓰레기	메마른땅	큰파도
	18	졸춥	눈폭탄	산폭팔	-
	19	졸덥	물난리	찢어진땅	-
	20	넙덥	비폭탄	흔들리는땅	-
	21	넙춥	홈비	-	-
	22	너무춥	홈탕물비	-	-
	23	너무덥	-	-	-
	24	개춥	-	-	-
	25	개덥	-	-	-

감성분류 기준표



감성 구분		감성 태그
긍정	황홀/기쁨	ECS
	기대/관심	EXP
	감탄/존경	ADM
부정	분노/짜증	RAG
	두려움/공포	TER
	슬픔/수심	SAD
중립		NEU

학습 데이터 구축 및 전처리



1. 기후변화에 따른 현상 사전 구축

기후변화에 따른 '현상' 키워드

구분	no.	온도	강수	토지	해양
전문가	1	강추위	대설	가뭄	녹조
	2	결빙	산성비	사막화	라니냐
	3	무더위	우박	산불	쓰나미
	4	열대야	장마	산사태	엘니뇨
	5	열섬	적설	열대림파괴	적조
	6	열파	집중강우	지진	침수
	7	온난	집중호우	토지황폐화	파랑
	8	온실가스	폭설	화산폭발	풍랑
	9	이상고온	폭우		풍수해
	10	이상기온	홍수		해랑
	11	이상저온	황사비		해수면
	12	폭염			해일
	13	한파			
	14	혹서			
	15	혹한			
비전문가	16	짱춥	눈난리	갈라진땅	괴물파도
	17	짱덥	눈쓰레기	메마른땅	큰파도
	18	졸춥	눈폭탄	산폭팔	
	19	졸덥	물난리	찢어진땅	
	20	넙덥	비폭탄	흔들리는땅	
	21	넙춥	홍비		
	22	너무춥	흙탕물비		
	23	너무덥			
	24	개춥			
	25	개덥			

2. 감성분류 기준 구축

		감성 태그
긍정	황홀/기쁨	ECS
	기대/관심	EXP
	감탄/존경	ADM
부정	분노/짜증	RAG
	두려움/공포	TER
	슬픔/수심	SAD
중립		NEU

3. 학습데이터 구축

No	Chanel	Keyword	Keyword1	Content	cross-check Tag	tag1	tag2	tag3	tag4	final Tag
1	뉴스 댓글	온도	강추위	난 추울때 겨울 남새가 넘 좋아	ADM	ECS	ADM	ECS	ECS	ECS
2	뉴스 댓글	온도	무더위	밖에서 일하시는 이 세상에 아버지 어머니를 힘내세요!!!	ADM	ADM	ADM	ECS	ECS	ADM
3	뉴스 댓글	온도	무더위	더 뜨거워지면 좋겠다	EXP	EXP	EXP	EXP	EXP	EXP
4	뉴스 댓글	온도	강추위	동요에 잔바람 불어도 괜찮아요 가사있죠? 이 날씨에 괜찮은 사람 없을듯 ㅡ_	RAG	RAG	RAG	TER	TER	RAG
5	뉴스 댓글	온도	무더위	하루하루가 더워때문에 넘 힘드네요..내일은 얼마나 또더울까..이런생각에...	TER	TER	TER	TER	TER	TER
6	뉴스 댓글	온도	너무춥	의정부 -13 너무 춥다..ㅠㅠ	SAD	SAD	SAD	SAD	SAD	SAD
8	뉴스 댓글	온도	강추위	오늘 모스크바 -11 대관령 -21.7	NEU	NEU	NEU	NEU	NEU	NEU
9	Facebook	온도	열대야	사워하고먹는아이스아메리카노가 진리임 열대야극복완료 단순해ㅋㅋㅋ	ECS	ECS	NEU	ECS	ADM	ECS
10	Facebook	온도	무더위	밖에 내다 놓은 화초들이 무더위를 견디다니..허허..튼튼한것들	ADM	ADM	ADM	NEU	NEU	ADM
11	Facebook	온도	열대야	열대야가 이번주가 끝이래요!!!! 밖으로 놀러가즈아~~~~	EXP	EXP	EXP	ADM	EXP	EXP
12	Facebook	온도	결빙	이젠 진심 눈이 싫어 지네요	RAG	RAG	TER	TER	RAG	RAG
13	Facebook	온도	결빙	역대급추위.....ㄷㄷㄷ올 첫 영산강 결빙현상 발생	TER	TER	TER	TER	TER	TER
14	Facebook	온도	무더위	응? 무더위 본격적 시작이라는데 ,, 그리고 10월까지 덥다는데 실화냐 ㅠㅠ	SAD	SAD	TER	TER	TER	TER

전처리 단계	전처리 내용
1) 이모지 한글로 변환	- SNS 특성을 반영한 전처리 단계
2) 이모티콘(특수문자) 전처리	- 형태소 분석기 : 은전한닢-Mecab
3) 형태소 분석	- 이모지 전처리: 약 1,200개 이모지 한글로 변환
4) ID 삭제	
5) 정규화: 함축어, 신조어, 은어	- 예) 😍 😱
6) DTM 생성	- DTM(Document Term Matrix)
7) 말뭉치(Corpus) 생성	
8) Sparse Terms 삭제	
9) Low TF-IDF 삭제	

	Term1	Term2	…	TermM
Doc1	2	1	…	0
Doc2	0	4	…	2
…				…
DocN	3	1	…	1

SVM Classifier

- 커널(Kernel) 트릭을 이용한 비선형 데이터 분류
- 커널 종류 및 파라미터

  1) 선형 커널(Linear Kernel) : Cost, Gamma

  2) RBF 커널(Radial Basis Function Kernel : Cost, Gamma

  3) 시그모이드 커널(Sigmoid Kernel) : Cost, Gamma, Coefficient

  4) 다항식 커널(Polynomial Kernel) : Cost, Gamma, Coefficient, Degree

$$Linear\ Kernel: K(x_n, x_i) = (x_n, x_i)$$

$$RBF\ Kernel: K(x_n, x_i) = \exp(-\gamma \parallel x_n - x_i \parallel^2 + C)$$

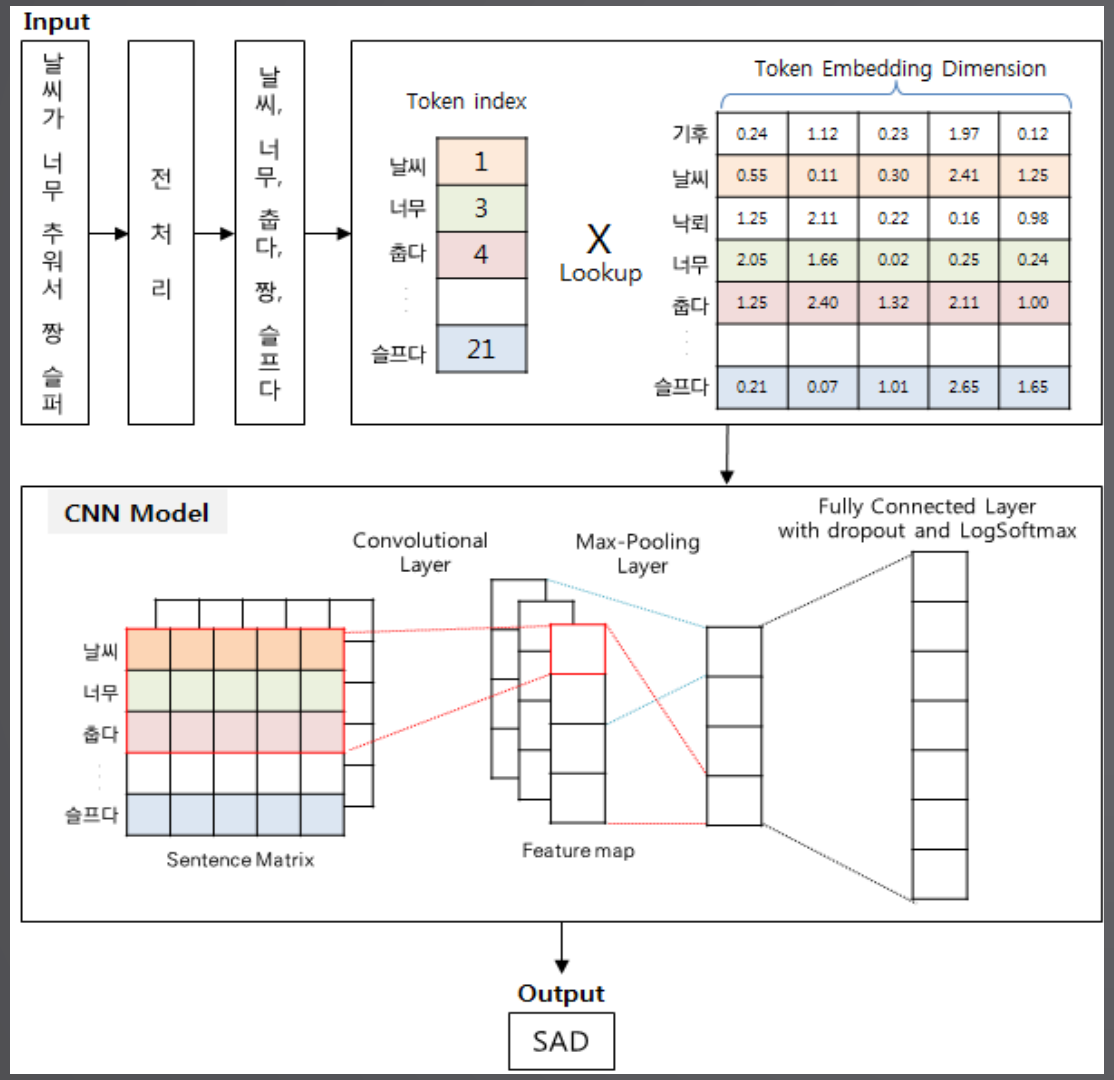
$$Sigmoid\ \ Kernel: K(x_n, x_i) = \tanh(\gamma(x_n, x_i) + r)$$

$$Polynomial\ Kernel: K(x_n, x_i) = (\gamma(x_n, x_i) + r)^d$$

- SVM Architecture

Parameter	Model			
	SVM_Model_1	SVM_Model_2	SVM_Model_3	SVM_Model_4
Kernel	Linear	RBF	Sigmoid	Polynomial
Cost	1	0.8	1	0.9
Gamma	0.0048	0.0040	0.0047	0.00047
Coefficient	-	-	0.001	0.001
Degree	-	-	-	3

CNN Classifier



CNN Classifier(

Embedding: dimension(13360, 128)

CNN-3-100 : Con2d(1, 100, kernel_size=(3, 128), stride=(1,1))

Activation : ReLU()

CNN-4-100 : Con2d(1, 100, kernel_size=(4, 128), stride=(1,1))

Activation : ReLU()

CNN-5-100 : Con2d(1, 100, kernel_size=(5, 128), stride=(1,1))

Activation : ReLU()

Generator : Linear(in_features=300, out_features=7)

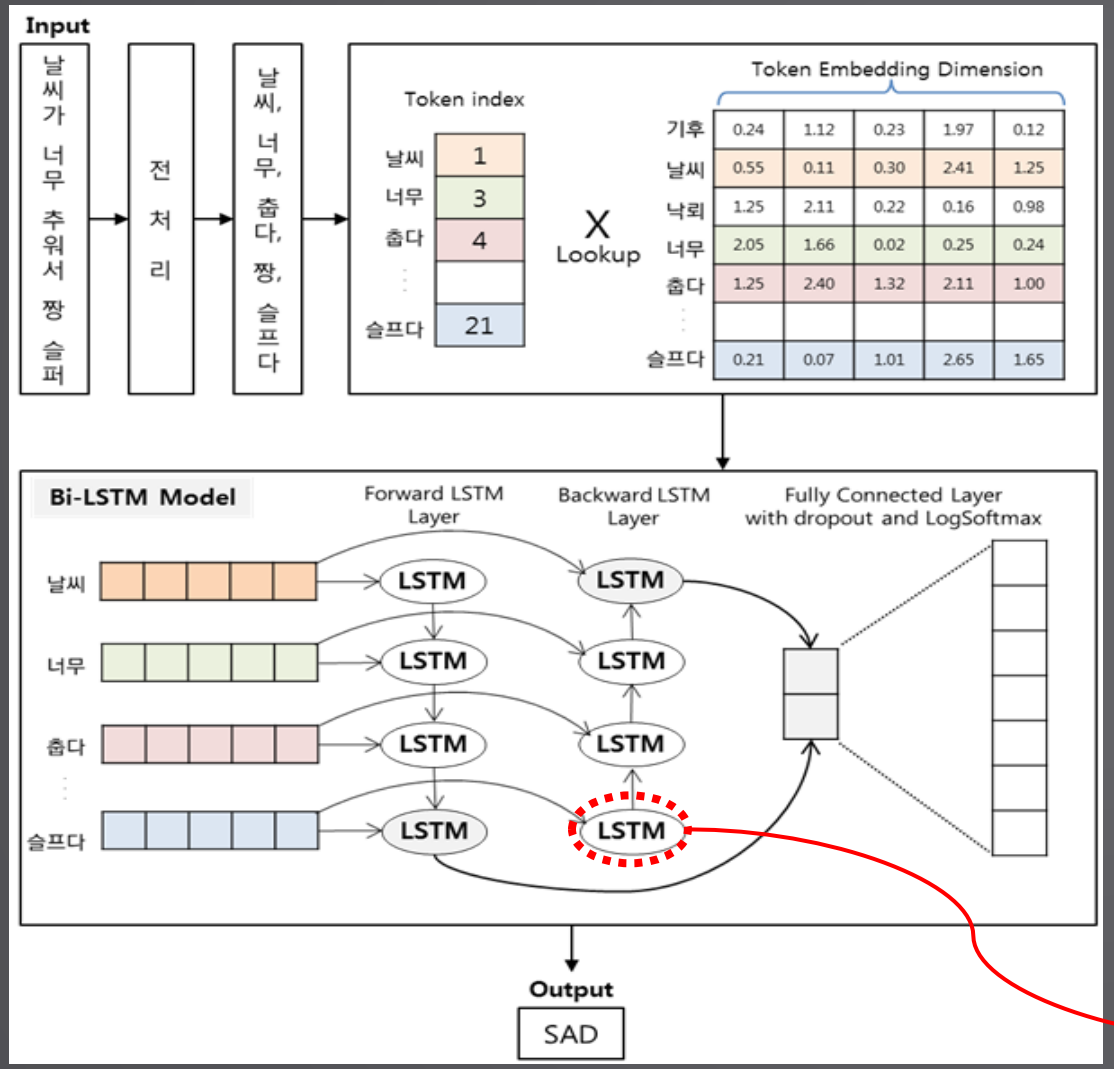
Pooling: Max-Pooling

Dropout : Dropout(p=0.3)

Activation: LogSoftmax()

)

Bidirectional LSTM Classifier



Bi-LSTM Classifier(

Embedding : dimension(13360, 128)

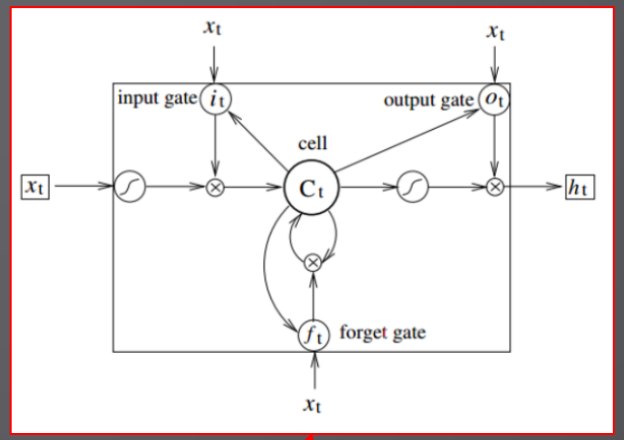
RNN: LSTM(num_layers = 4, bidirectional = True)

Generator : Linear(in_features=512, out_features=7)

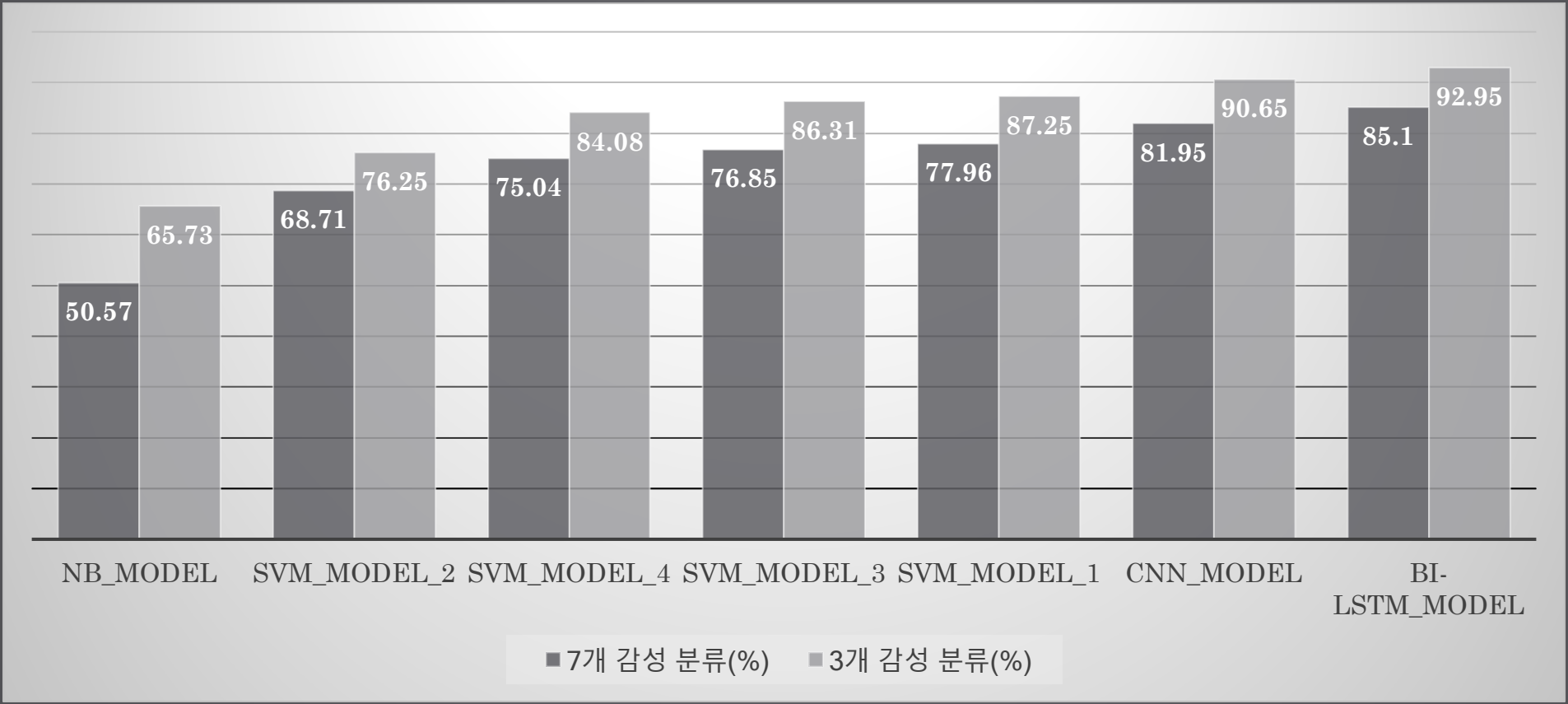
Dropout : Dropout(p=0.3)

Activation: LogSoftmax()

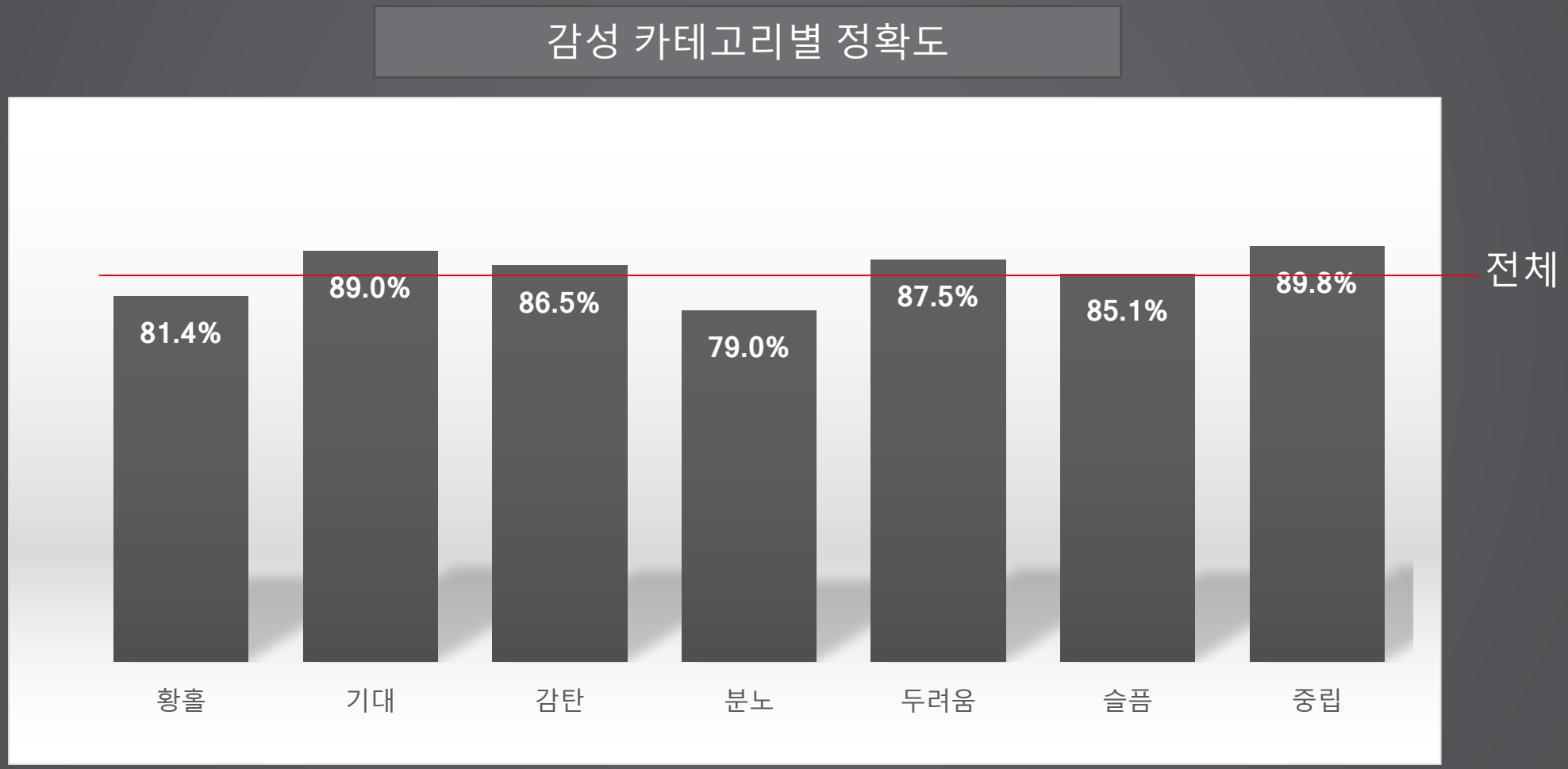
)



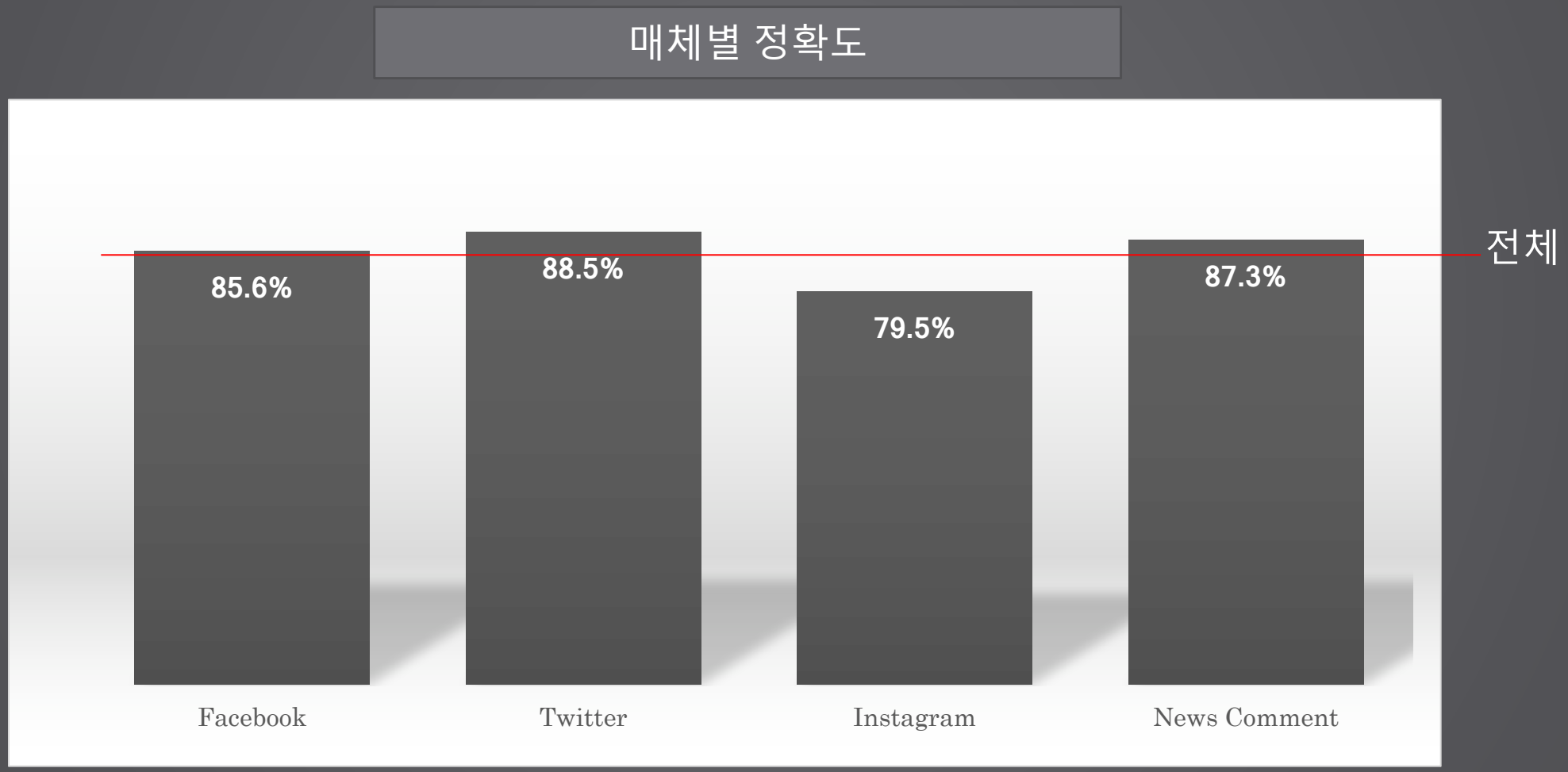
감성 분류 정확도(Accuracy) 비교



7 개 감정 class별 분류 성과 : Bi-LSTM 85.10%



매체별 분류 정확도: Bi-LSTM 85.10%



인스타그램: 신조어, 은어, 함축어, 띄어쓰기 문제가 다른 매체보다 심각

(4) 미세먼지 농도 및 예보가 서울 대중교통 이용에 미치는 영향 [김진형]

- 미세먼지 농도 및 예보가 사회적 행위(대중교통 이용)에 미치는 영향 파악
  - 미세먼지로 인한 외부활동의 감소 및 대중교통 수요 증가 현상 등을 정량적으로 파악

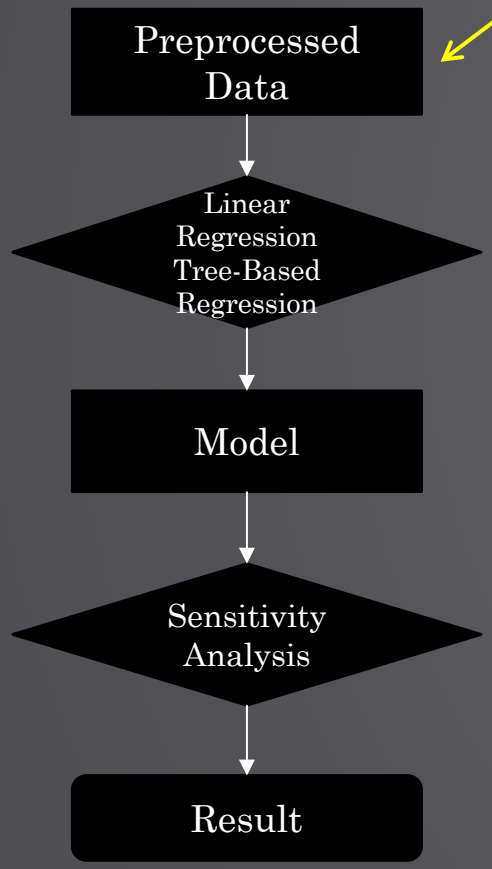
- 연구 내용: 미세먼지 농도가 지하철 이용에 미치는 영향 분석 및 지하철 이용 패턴 예측
  - 자료: 서울시 지하철 승하차 정보(서울 열린데이터 광장, 공공데이터 포털) , 기상기후 데이터(기상자료개방포털), 미세먼지 데이터(에어코리아)
  - 미세먼지 농도 변화에 따른 지하철 이용의 변화를 의사결정 나무 방법론을 적용하여 분석
    - 방법론 후보군: 회귀분석, SVM(Supporting Vector Mechanism), Boosted Tree
  - 실시간으로 변화하는 자료의 특성을 반영하여 추정 결과를 상시적으로 갱신하는 발신 방식을 고민

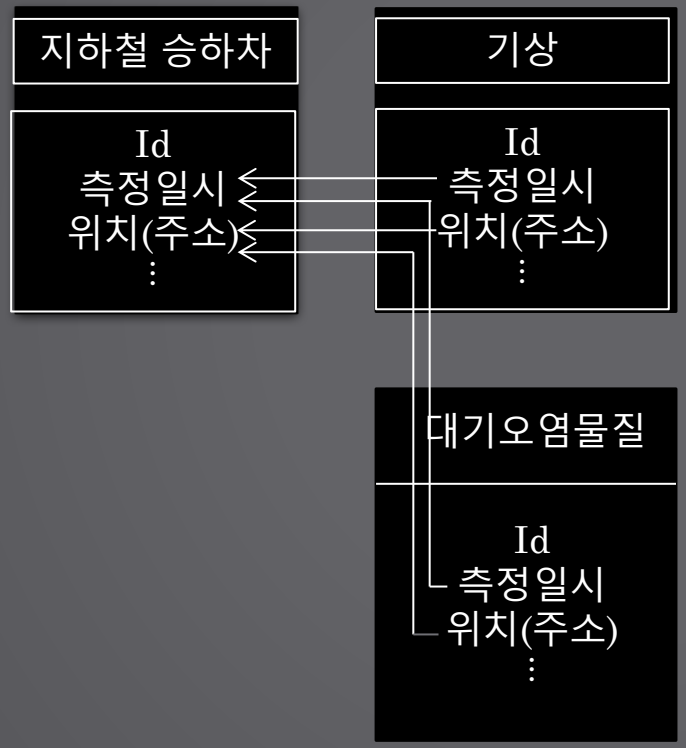
- 연구성과: 미세먼지 농도가 증가하면 지하철 승하차 인원이 감소하는 추세 발견
  - Boosted Decision Tree 알고리듬 사용 선형회귀분석 대비 예측오차 축소
  - 미세먼지 오염도와 승하차 인원 간 양의 상관관계가 존재하나 승하차 인원 증감 폭은 작은 현상 발견
  - 비대칭성 존재 : 오염도 심화 시 승하차 인원 감소 < 오염도 완화 시 승하차 인원 증가

- 향후 승하차 인구 절대값, 계절성 등을 조정하여 추정 및 민감도 분석을 진행할 예정

분석 알고리듬 및 데이터 전처리 (1)

분석 알고리듬 | 데이터 전처리 | 사용한 데이터



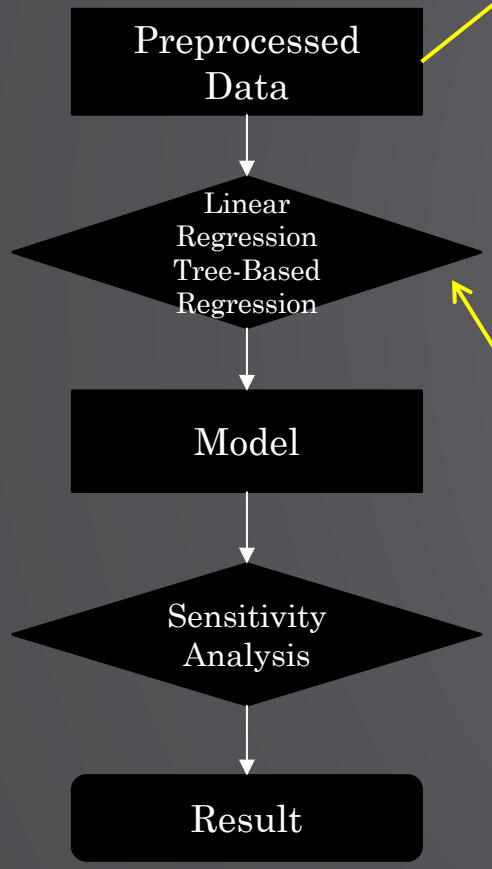


데이터명	출처	속성	비고
서울시 지하철 승하차 인구	서울교통공사 (공공데이터포털)	Id, 측정일시, 위치(주소), 승/하차, 인원	서울시 227개 역
기상 데이터	기상자료개방포털	Id, 측정일시, 위치(주소), 기온, 강수량 풍속, 습도, 적설	서울시 측정정보
대기오염물질농도 데이터	에어코리아	Id, 측정일시, 위치(주소), PM10, SO2, CO, O3, NO2	서울시 46개 측정소

- 측정일시 및 위치 정보(구 단위)를 기반으로 데이터 매칭
- 2015.01~2017.09 기간 동안 1시간 단위의 데이터 활용

분석 알고리듬 및 데이터 전처리 (2)

분석 알고리듬



데이터 전처리 결과

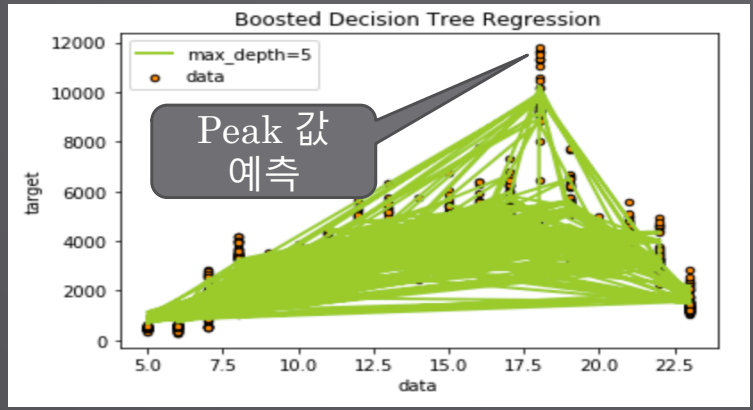
지하철 승하차 인원 | 기상 데이터 | 대기오염물질 데이터

	date	id	name	category	location	time	value	dow	지점	기온 (°C)	...	풍속 (m/s)	습도 (%)	적설 (cm)	시군구	SO2	CO	O3	NO2	PM10	PM25
0	2015-01-01	150	서울역	0	중구	5	441	Thu	108	-9.1	...	5.7	35.0	0.0	용산구	0.0045	0.25	0.0185	0.0080	111.0	19.0
1	2015-01-01	151	시청	0	중구				08	-9.1	...	5.7	35.0	0.0	중구	0.0055	0.40	0.0205	0.0075	97.5	12.0
2	2015-01-01	152	종각	0	종로구	5	898	Thu	108	-9.1	...	5.7	35.0	0.0	종로구	0.0045	0.20	0.0200	0.0055	118.5	5.0

데이터 조인

- 반응변수: 지하철 하차 인원
- 설명변수: 월, 시간, 요일 및 기상, 대기오염물질 데이터의 속성 데이터

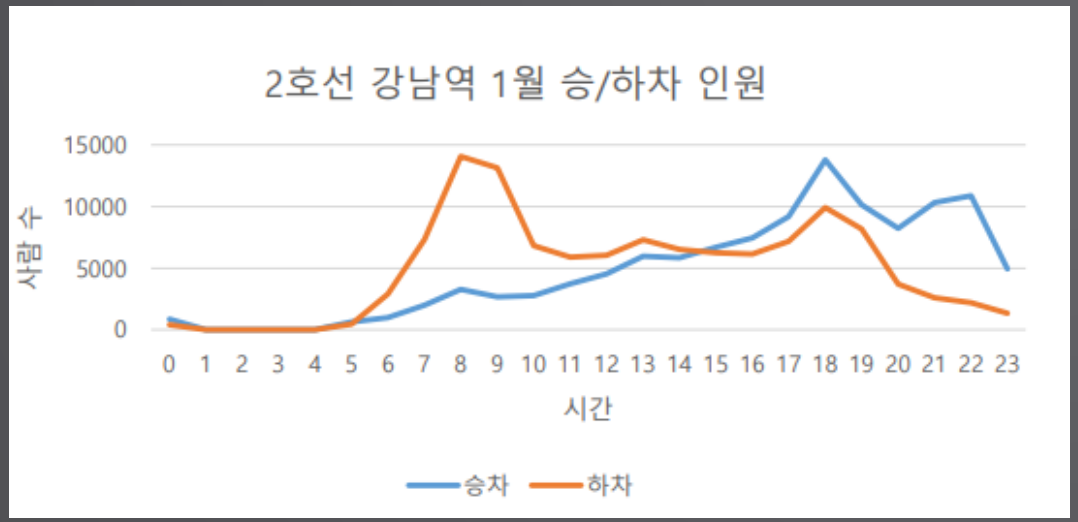
Decision Tree Regression 테스트 결과 : 비선형성 모델링에 유리함

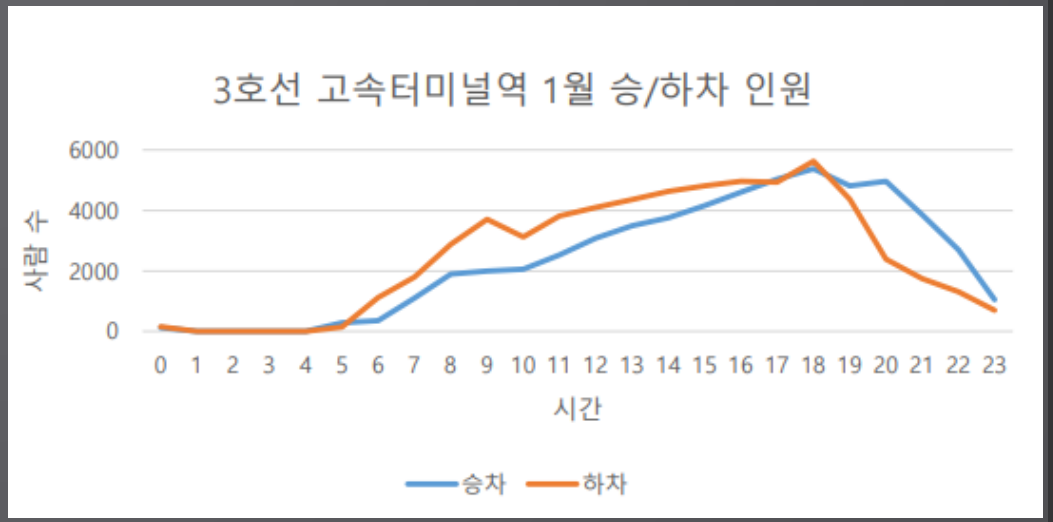


- 지하철 승하차 인구 패턴은 선형 모델만으로 설명할 수 없음
- Tree-Based 모델은 비선형성 모델링이 가능함

모델 구축 및 정확도 분석 (1)

- 선형회귀, 의사결정나무, 랜덤포레스트, 부스팅(Gradient Descent Boosting) 기법 사용
  - 선형회귀는 Tree-based 모델의 성능을 비교하기 위한 기준모델로서 사용함
- 모델 구축 시 지하철역별로 4개의 모델을 구축
  - 각 지하철역마다 서로 다른 승하차 인원 패턴을 보임 (아래 그림 참조)
  - 월, 시간, 요일, 기상, 대기오염 데이터를 설명변수로 사용하는 본 연구에서 한 가지 모델로 서로 다른 패턴의 전체 지하철역 승하차 패턴을 예측하기는 어려움

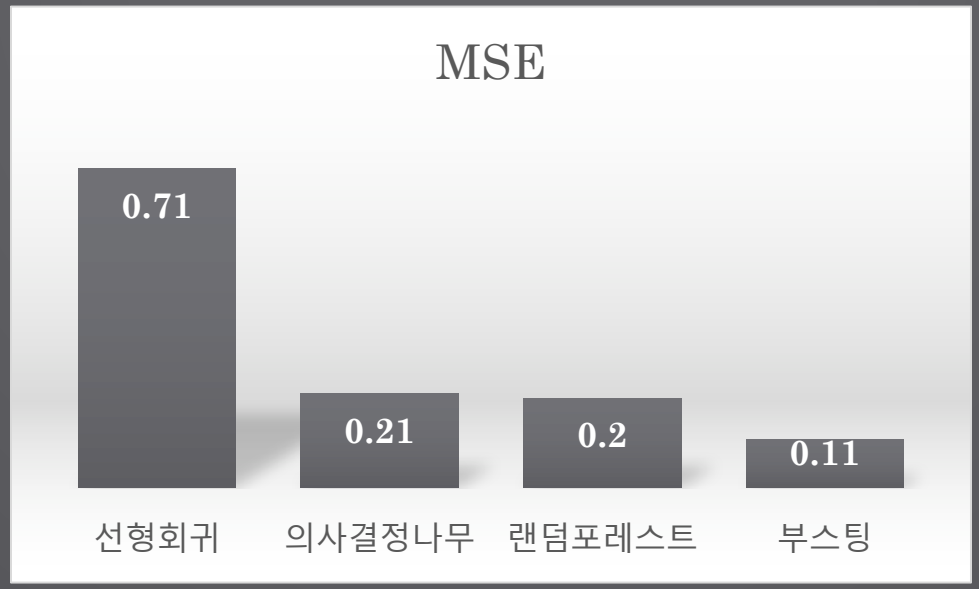




모델 구축 및 정확도 분석 (2)

- 각 지하철역의 승하차 인원을 표준화하여 분석 수행하여 정확도를 비교
- 네 가지 모델의 정확도 비교 결과 부스팅 기법의 정확도가 가장 높음
  - Tree-based 모델은 선형회귀보다 높은 정확도를 보였음
    - 0.71 × 승하차 인구 표본 분산 → 0.11 × 승하차 인구 표본 분산

구축된 모델의 정확도 분석 결과: MSE

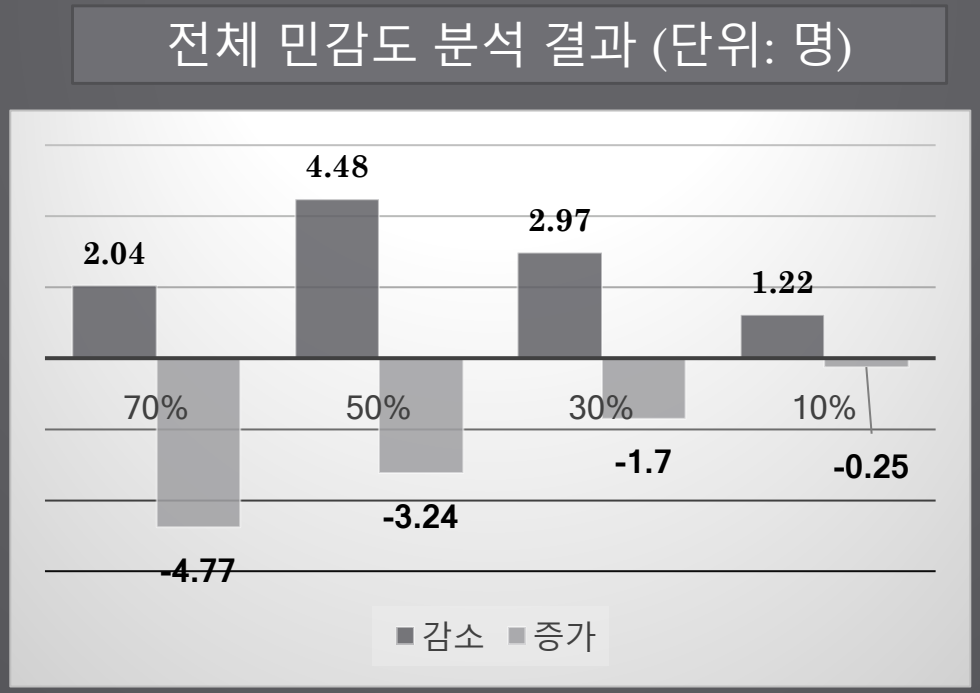


부스팅 모델 대상 민감도 분석 (1)

- 가장 정확도가 높은 부스팅 모델을 대상으로 민감도 분석 수행
  - PM10과 PM2.5의 농도를 10%, 30%, 50%, 70% 증가 혹은 감소시켜 하차 인원 예측
  - 가장 정확도가 높은 부스팅 모델을 대상으로 민감도 분석 수행
  - 각 지하철역의 민감도 분석 결과 평균값
- 지하철역의 특성에 따른 민감도 분석 수행
  - 공간적 특성 :지하철역 주변의 토지이용현황 분류 별 민감도 분석 결과 평균값
  - 승하차 인원 : 1시간 평균 지하철역 승하차인원에 따른 민감도 분석 결과 평균값
  - 시간대 :평일/주말, 출퇴근시간대에 따른 민감도 분석 수행 (예정)

부스팅 모델 대상 민감도 분석 결과 : 전체 표본

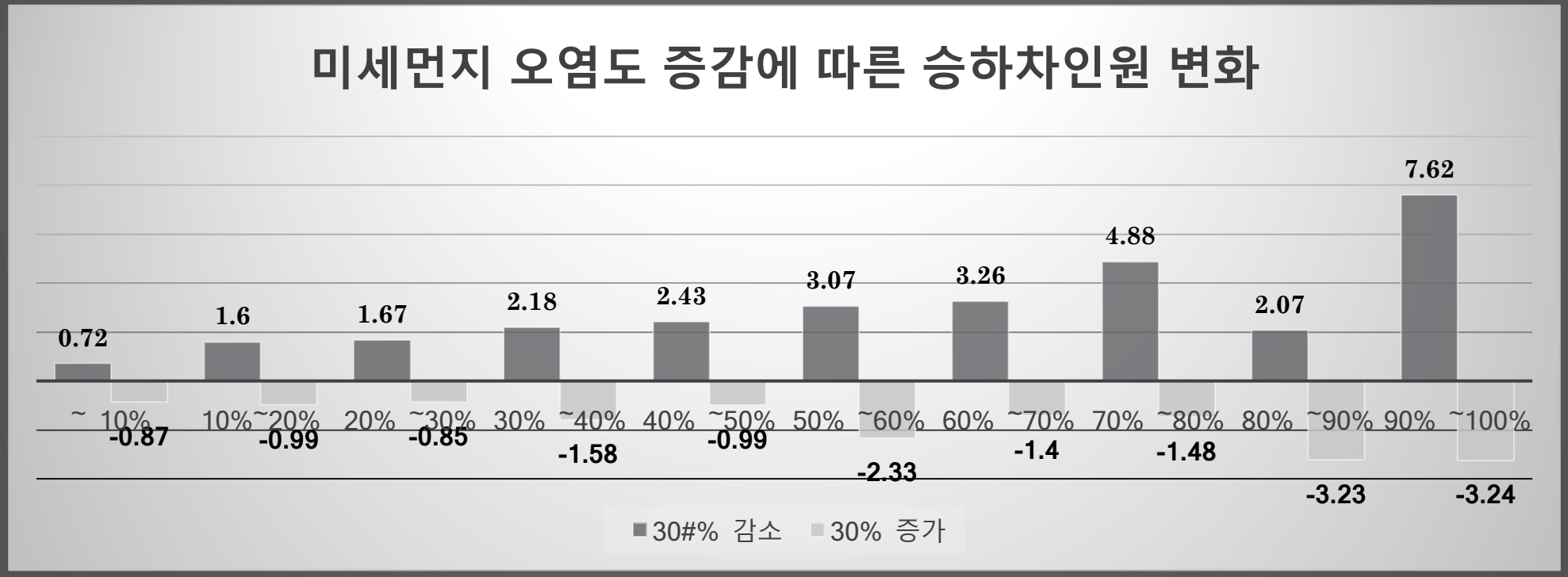
- 민감도 분석 결과 미세먼지가 증가할 때 지하철역 하차 인원이 감소함
  - 미세먼지가 70% 증가할 때, 서울시 내 지하철역의 하차인원이 4.77명 감소함
    - 평균 승하차 인원 946명, 승하차 인원 중위값 709명에 비하면 매우 작은 값
- 비대칭성 존재 : 미세먼지 감소 시 승하차 인원 증가 > 미세먼지 증가 시 승하차 인원 감소



부스팅 모델 대상 민감도 분석: 평균 승하차 인원의 영향

◆ 승하차 인원이 많을 수록 미세먼지 오염도에 따른 승하차 인원 변화가 큼

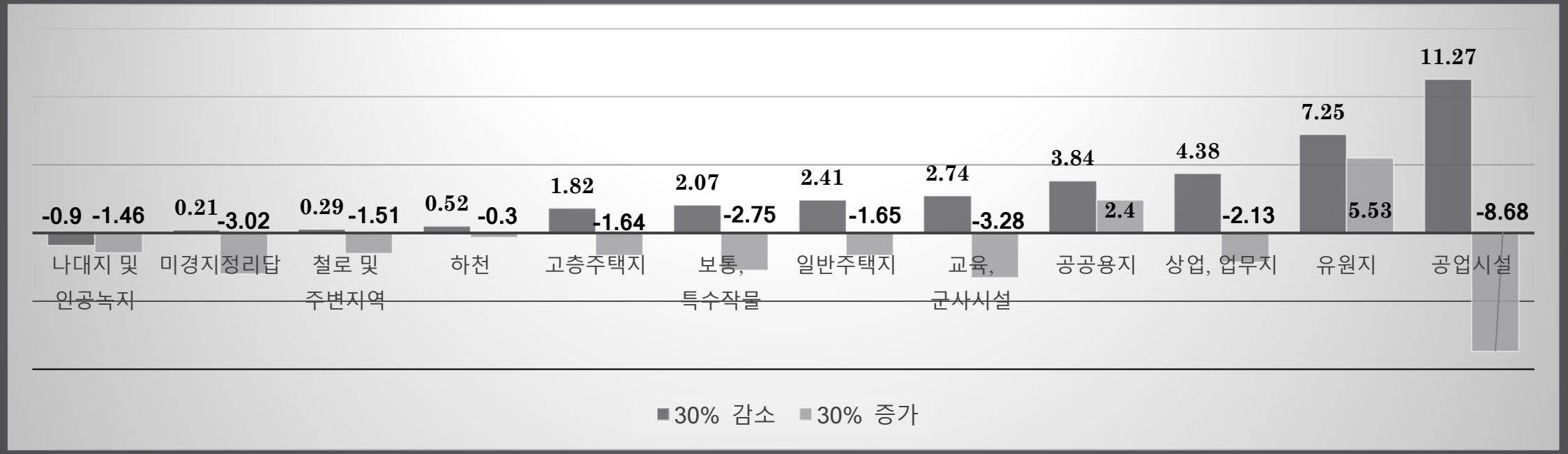
평균 승하차 인원에 따른 민감도 분석 결과 (단위: 명)



부스팅 모델 대상 민감도 분석 (3) : 지역적 특성

- 토지이용과 관계 없이 미세먼지가 증가함에 따라 유동인구가 감소
  - 공업시설 지역의 민감도가 가장 크고, 하천의 민감도가 가장 작음
  - 기타-유원지, 공공시설물-공공용지 지역에서는 미세먼지가 증가해도 유동인구가 증가
    - (봄,가을철 유동인구 증가로 인한 현상으로 추정)

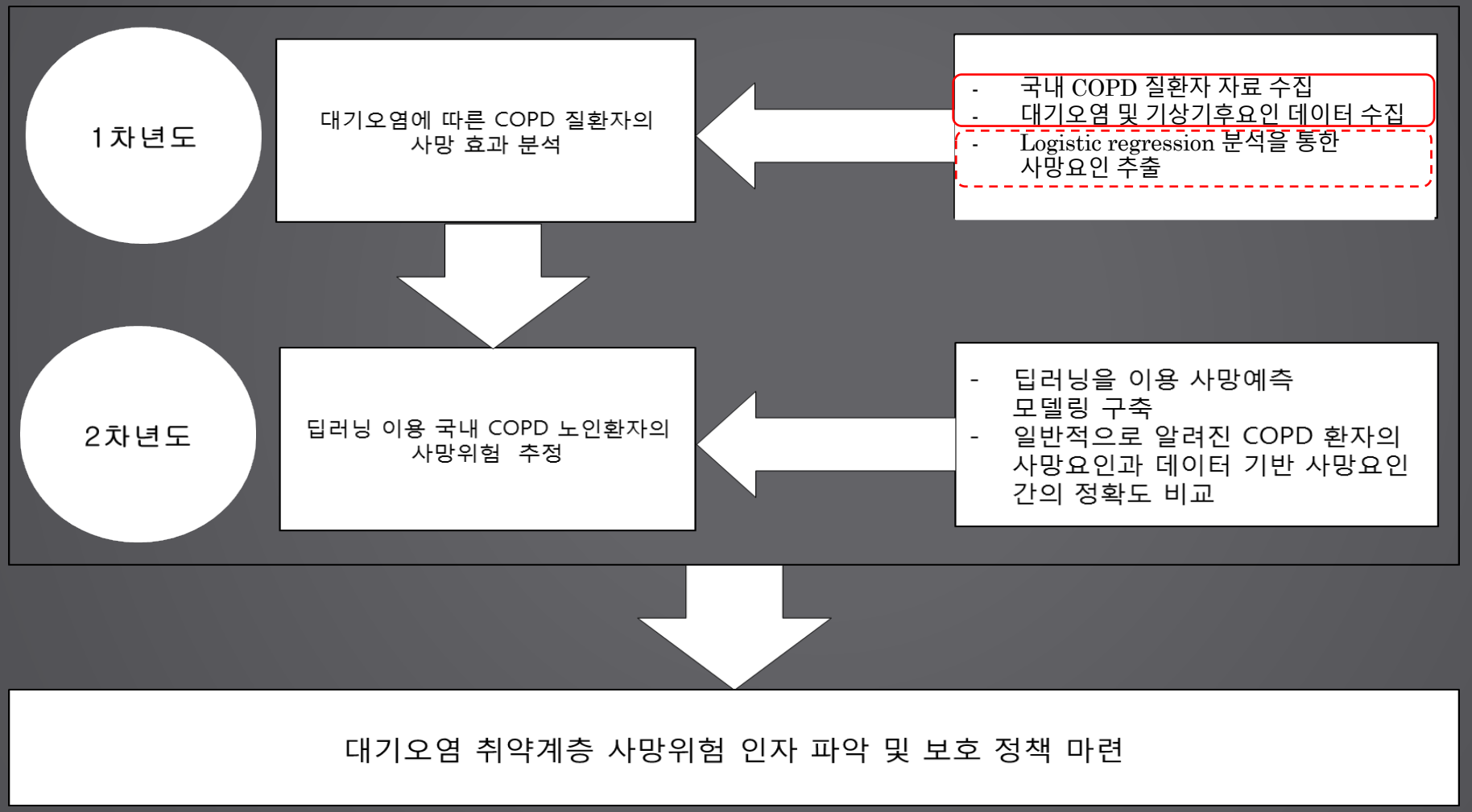
토지이용에 따른 민감도 분석 결과 (단위: 명)



(5) 딥러닝 이용 노인인구 호흡기 질환 사망 위험 추정 [강선아]

- 만성폐쇄성 폐질환 사망 위험을 딥러닝을 이용하여 추정
  - 연구 대상 :65세 이상 만성폐쇄성폐질환(COPD) 환자
- 연구내용 : 1단계 사망요인 파악 연구 , 2단계 사망확률 추정 연구 2년차로 확대
  - 자료 : 건강보험 맞춤형연구 DB , 인구, 기후, 대기오염도 및 대기오염물질 배출량 자료를 연계
  - 1차년도 (2018) : 대기오염에 따른 COPD 질환자 사망요인 분석(맞춤형 연구 DB)
    - GAM 분석을 통해 사망에 영향을 미치는 주요 변인을 발견하고 사망 효과를 분석
  - 2차년도 (2019) : COPD 노인 질환자 사망위험 추정
    - 딥러닝 이용 사망예측 모델 구축
    - 1차년도 분석 시 발견된 COPD 환자의 사망요인 .vs 데이터 기반 파악 사망요인 : 사망위험 추정치 정확도 비교
- 연구 진행 상황 : 사망요인 분석 중
  - 환경 및 기후인자, 진단 데이터가 사망에 미치는 요인 분석
  - 사망요인 선정 과정에서 의료 전문가 인터뷰 진행 (서울삼성병원 호흡기내과 박혜윤박사/서울 강남 카톨릭 병원 이진국박사)

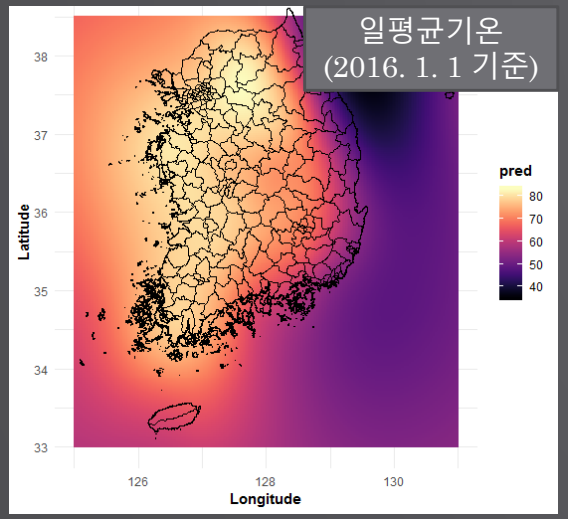
연구 프레임워크

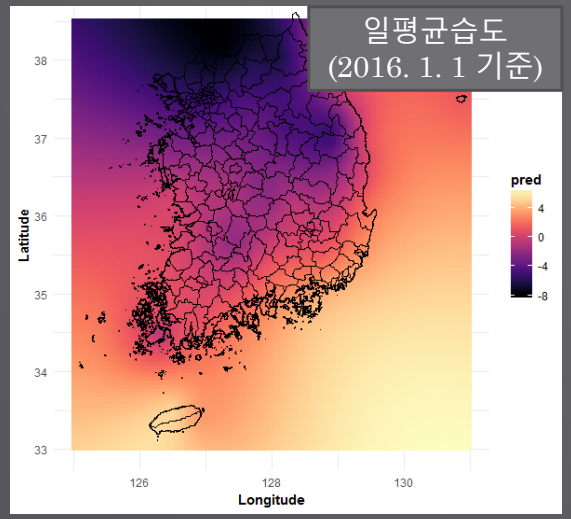


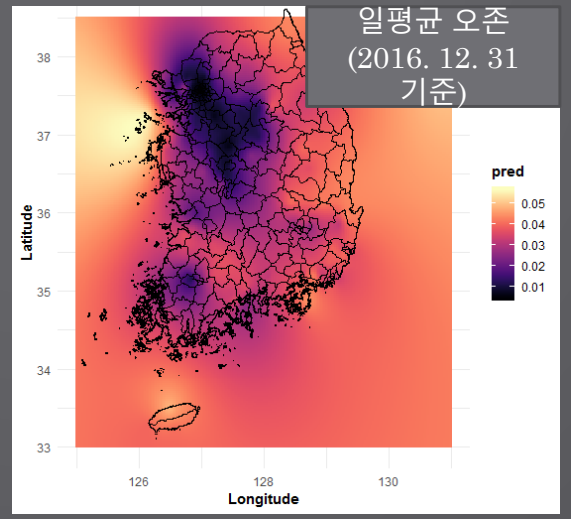
완료 | 진행 중

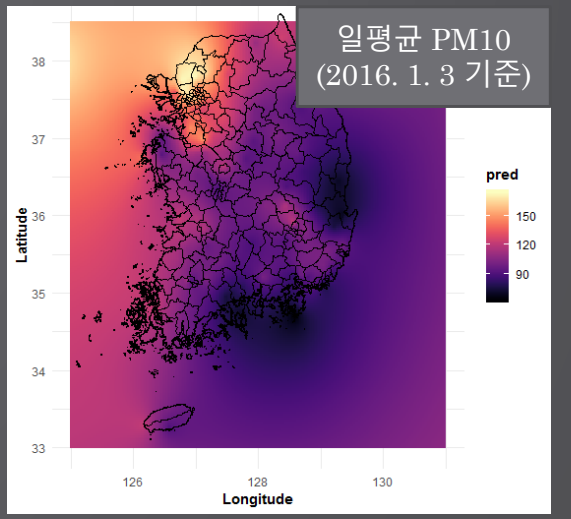
분석 자료 및 전처리 과정(대기오염 및 기상자료)

- 설명변수: 국민건강보험공단 맞춤형 DB 진료 기록, 기상기후 데이터, 대기오염 데이터
- 대기 및 기상 자료: 건강보험 DB 와 연계 목적 시공간 해상도 조정
  - GIS 기법 중 Kriging을 이용하여 점(point)데이터인 측정소 자료를 면 데이터인 시군구 자료로 변형
  - 시간단위 대기오염 데이터를 일간 평균을 취하여 일 단위 건강보험 DB와 연계







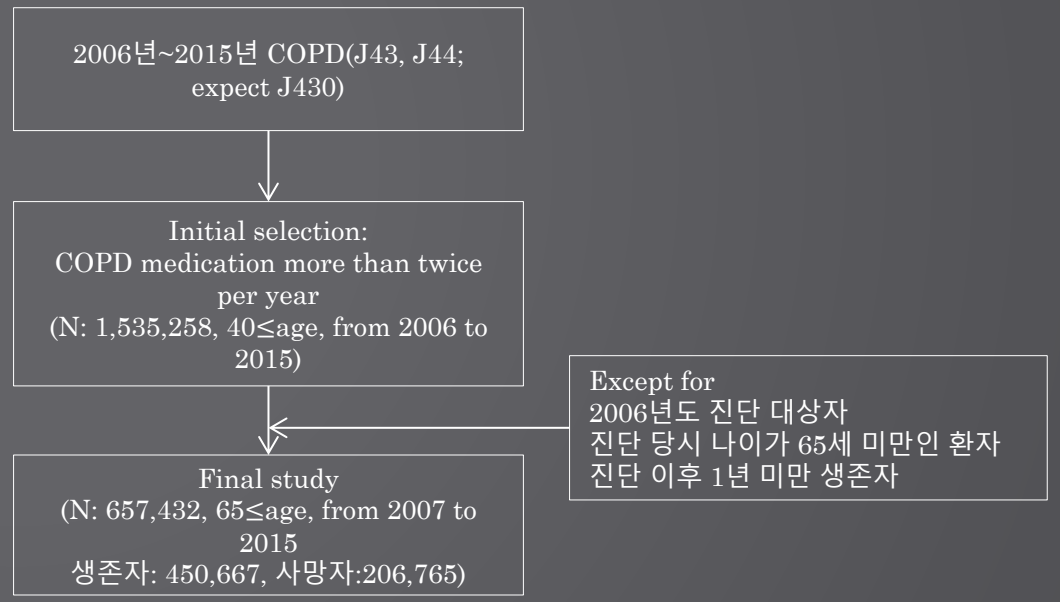


건강보험 맞춤형 DB 신청항목 및 전처리 내용

DB 신청내용

조건	내용
연도	2006년~2015년
상병코드	J43, J44(except J430)
주상병/부상병	주상병 및 모든 부상병
산정특례 특정기호구분	없음
의과/한방/치과/약국	의과
입원/외래	입원, 외래
행위수가코드	전체자료
약제주성분코드	전체자료
기타	없음

전처리 내용

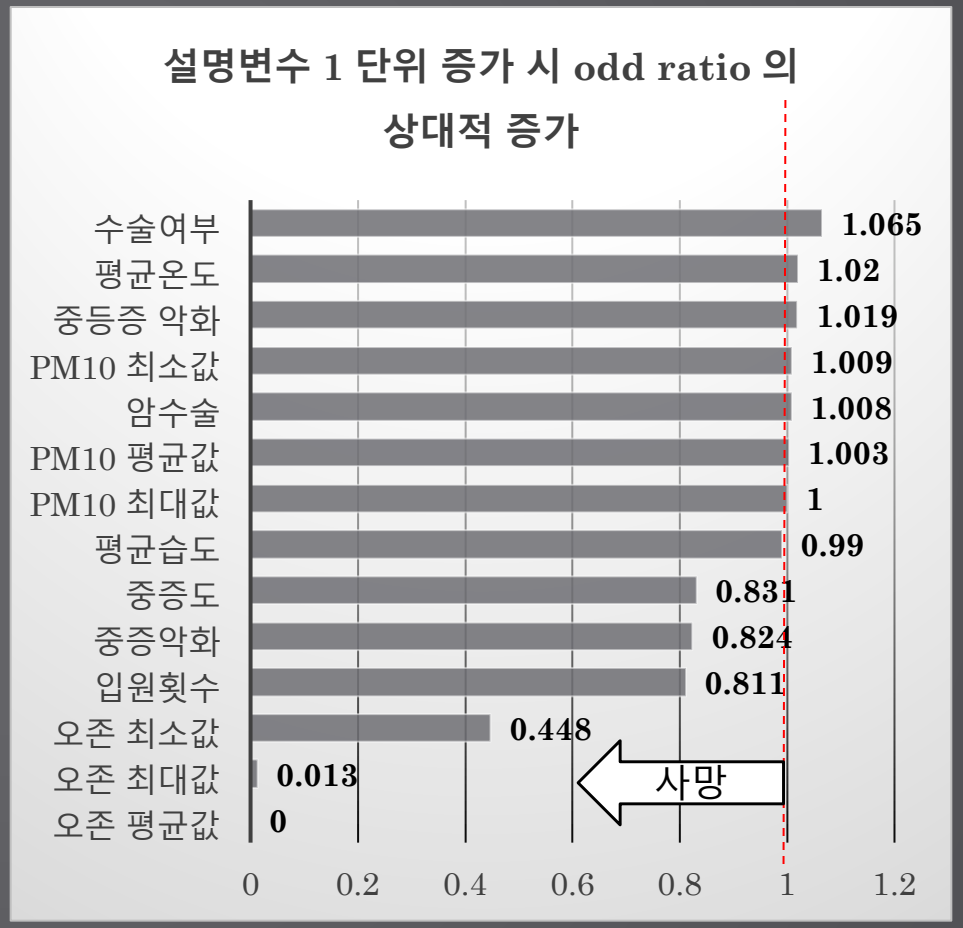


설명변수

구분		변수명	설명
설명 변수	개인정보 데이터	나이	연속형 변수
		성별	연속형 변수
		거주지	범주형변수
		보험료 20분위	범주형 변수
		직종	범주형 변수
		장애등급	범주형 변수
	진단 데이터	Charlson 동반질환지수	범주형 변수
		입원횟수	연속형 변수
		수술여부(암/비암)	범주형 변수
		COPD 중증도	범주형 변수
		외래 경증 악화-중등증 악화	범주형 변수
		외래 경증 악화-중증 악화	범주형 변수
		산소 처방전	범주형 변수
	기후 데이터	평균 기온	연속형 변수
		평균 습도	연속형 변수
	대기오염 데이터	일평균 오존, 일최대 오존, 일최소 오존	연속형 변수
		일평균 PM10, 일최대 PM10, 일최소 PM10	연속형 변수
		일평균 PM25, 일최대 PM25, 일최소 PM25	연속형 변수

COPD 사망요인 분석 : Logistic Regression

- ◆ 시군구별, 성별, 연령별 층화추출(sample size = 131,478명, 20% sampling)
- ◆ 조정변수 : 성별, 연령, 소득(보험료 20 구간), 직종, 거주지(대도시, 중소도시)
- ◆ 미세먼지 농도는 사망확률에 거의 영향을 주지 않음
  - 사망일자 11월~2월 집중 : 미세먼지 오염도가 낮은 기간
- ◆ 오존 농도는 사망확률을 대폭 증가시키는 경향
- ◆ 향후 계획
  - 통계청 사망원인통계자료 연계
  - 계절요인 통제 등 fine tuning 작업
  - 딥러닝을 이용한 사망위험 예측 작업



$$\exp(b_i) = \frac{\frac{\Pr[Y=1]}{\Pr[Y=0]}\Big|_{x_i=x_{i,0}}}{\frac{\Pr[Y=1]}{\Pr[Y=0]}\Big|_{x_i=x_{i,0+1}}}$$

COPD 사망요인 분석

Variable	Level of CLASS Variable	Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Estimation Type	exp(beta)
	1 for Variable						
Intercept		0.4824	0.2753	3.071	0.0797	MLE	1.620
성별	남성	0.3339	0.00045	551690.874	<.0001	MLE	1.396
나이(60대)	60대	-1.1236	0.00126	790988.901	<.0001	MLE	0.325
나이(70대)	70대	-0.5814	0.000828	493167.007	<.0001	MLE	0.559
나이(80대)	80대	0.3842	0.000884	188930.107	<.0001	MLE	1.468
보험료 20분위	0	0.345	0.00128	72439.9833	<.0001	MLE	1.412
보험료 20분위	1	0.1443	0.00179	6503.0957	<.0001	MLE	1.155
보험료 20분위	2	-0.1059	0.00241	1935.53	<.0001	MLE	0.900
보험료 20분위	3	0.0624	0.00278	503.4426	<.0001	MLE	1.064
보험료 20분위	4	-0.0821	0.00277	876.9987	<.0001	MLE	0.921
보험료 20분위	5	0.0457	0.00287	253.8497	<.0001	MLE	1.047
보험료 20분위	6	0.0165	0.00285	33.6005	<.0001	MLE	1.017
보험료 20분위	7	0.0939	0.00272	1193.1046	<.0001	MLE	1.098
보험료 20분위	8	-0.00995	0.00256	15.0609	0.0001	MLE	0.990
보험료 20분위	9	0.0637	0.00249	655.1108	<.0001	MLE	1.066
보험료 20분위	10	0.0629	0.00238	699.0907	<.0001	MLE	1.065
보험료 20분위	12	-0.0477	0.00231	425.4586	<.0001	MLE	0.953
보험료 20분위	13	0.00526	0.00216	5.9501	0.0147	MLE	1.005
보험료 20분위	14	-0.0145	0.00208	48.3566	<.0001	MLE	0.986
보험료 20분위	15	-0.0583	0.00196	882.7175	<.0001	MLE	0.943
보험료 20분위	16	-0.0758	0.00181	1749.4707	<.0001	MLE	0.927
보험료 20분위	17	-0.1056	0.00168	3931.5976	<.0001	MLE	0.900
보험료 20분위	18	-0.0779	0.00159	2399.0146	<.0001	MLE	0.925
보험료 20분위	19	-0.0654	0.00149	1932.2537	<.0001	MLE	0.937

COPD 사망요인 분석

직종	기능직		0.5597	0.2746	4.1541	0.0415	MLE	1.750
직종	1,2종 고용직		1.0858	0.2766	15.4116	<.0001	MLE	2.962
직종	경노무 고용직		0.8461	0.2763	9.3763	0.0022	MLE	2.331
직종	법관, 검사		0.6381	0.2763	5.333	0.0209	MLE	1.893
직종	일용직		0.5977	0.2747	4.7339	0.0296	MLE	1.818
직종	공중보건의		2.0399	0.2879	50.1919	<.0001	MLE	7.690
직종	기능직		2.6633	0.3133	72.243	<.0001	MLE	14.344
직종	원어민 영어교사		1.5123	0.3123	23.4426	<.0001	MLE	4.537
도시	대도시		-0.0638	0.000659	9378.4359	<.0001	MLE	0.938
도시	중소도시		-0.0428	0.000649	4346.9869	<.0001	MLE	0.958
입원횟수			-0.2098	0.00057	135258.097	<.0001	MLE	0.811
중증악화		0	-0.1935	0.000627	95286.3292	<.0001	MLE	0.824
중등증 악화		0	0.0192	0.000448	1828.2317	<.0001	MLE	1.019
중증도		0	-0.1849	0.00222	6929.3357	<.0001	MLE	0.831
암수술		0	0.00793	0.00179	19.6828	<.0001	MLE	1.008
수술여부		0	0.0629	0.000487	16677.2772	<.0001	MLE	1.065
평균습도			-0.0102	0.000055	34200.7216	<.0001	MLE	0.990
평균온도			0.0202	0.000086	54853.6957	<.0001	MLE	1.020
오존 최대값			-4.3565	0.0493	7818.8761	<.0001	MLE	0.013
오존 평균값			-10.7752	0.0895	14490.2895	<.0001	MLE	0.000
오존 최소값			-0.8038	0.101	63.3386	<.0001	MLE	0.448
pm10 최대값			0.000381	5.51E-06	4780.8784	<.0001	MLE	1.000
pm10 평균값			0.00291	0.00004	5327.909	<.0001	MLE	1.003
pm10 최소값			0.00914	0.000084	11725.8266	<.0001	MLE	1.009

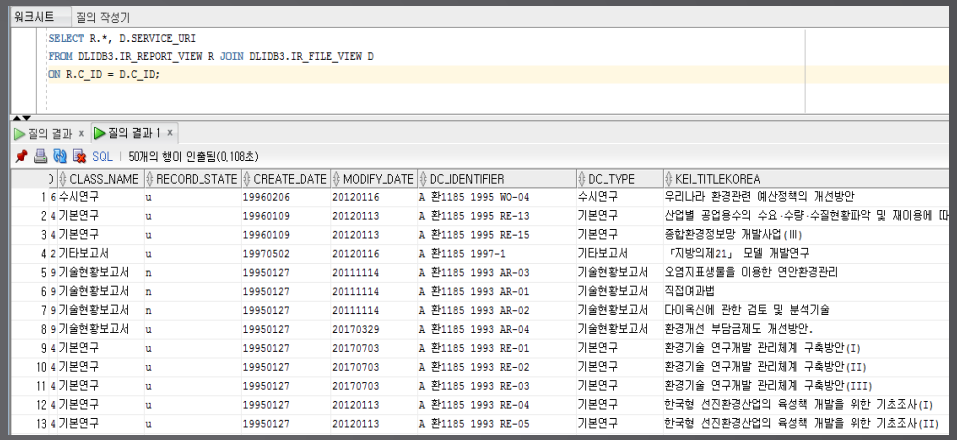
3. 환경 빅데이터 플랫폼

1. Open Data Map

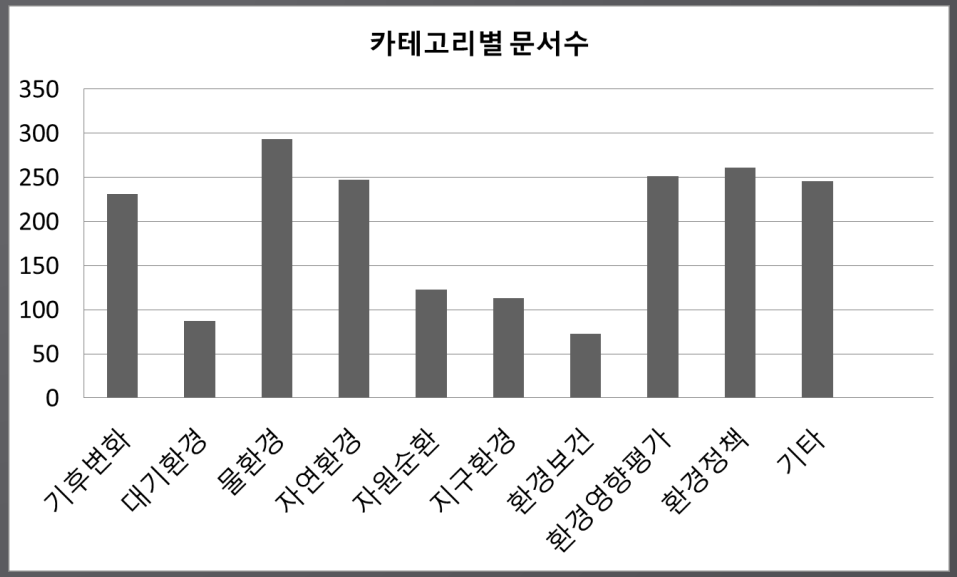
2. 분석 플랫폼 설계

(1) Open Data Map

- Open Data Map : 환경연구 활용 빈도가 높은 온라인 데이터 hub
  - 카테고리 별 데이터 소스/링크/ 내용소개 / 키워드를 제공
  - KEI 보고서에서 자주 인용된 데이터 소스 우선 파악
- KEI 도서관 DB로 부터 KEI에서 발간된 문서정보 추출
  - 문서 경로, 문서 제목, 키워드, 카테고리 등
- 문서 수집 및 변환
  - PDF → TXT
  - 총 1967개 문서 중 1925개 문서 활용
    - 미분류 문서카테고리 재분류 (277개)
    - 복수 카테고리 문서 제외 (12개)
    - 다운로드가 안되거나 txt 문서로 전환이 안되는 문서 제외 (30개)
- 문서 카테고리 재분류
  - 총 22가지 카테고리 → 10개 카테고리



	CLASS_NAME	RECORD_STATE	CREATE_DATE	MODIFY_DATE	DC_IDENTIFIER	DC_TYPE	KEI_TITLEKOREA
1	6 수시연구	u	19960206	20120116	A 환1185 1995 WO-04	수시연구	우리나라 환경관련 예산정책의 개선방안
2	4 기본연구	u	19960109	20120113	A 환1185 1995 RE-13	기본연구	산업별 공업용수의 수요·수량·수질현황파악 및 재이용에 따
3	4 기본연구	u	19960109	20120113	A 환1185 1995 RE-15	기본연구	종합환경정보망 개발사업(III)
4	2 기타보고서	u	19970502	20120116	A 환1185 1997-1	기타보고서	「지방의제21」 모델 개발연구
5	9 기술현황보고서	n	19950127	20111114	A 환1185 1993 AR-03	기술현황보고서	오염지표생물을 이용한 연안환경관리
6	9 기술현황보고서	n	19950127	20111114	A 환1185 1993 AR-01	기술현황보고서	직접여과법
7	9 기술현황보고서	n	19950127	20111114	A 환1185 1993 AR-02	기술현황보고서	다이옥신에 관한 검토 및 분석기술
8	9 기술현황보고서	u	19950127	20170329	A 환1185 1993 AR-04	기술현황보고서	환경개선 부담금제도 개선방안.
9	4 기본연구	u	19950127	20170703	A 환1185 1993 RE-01	기본연구	환경기술 연구개발 관리체계 구축방안(I)
10	4 기본연구	u	19950127	20170703	A 환1185 1993 RE-02	기본연구	환경기술 연구개발 관리체계 구축방안(II)
11	4 기본연구	u	19950127	20170703	A 환1185 1993 RE-03	기본연구	환경기술 연구개발 관리체계 구축방안(III)
12	4 기본연구	u	19950127	20120113	A 환1185 1993 RE-04	기본연구	한국형 선진환경산업의 육성책 개발을 위한 기초조사(I)
13	4 기본연구	u	19950127	20120113	A 환1185 1993 RE-05	기본연구	한국형 선진환경산업의 육성책 개발을 위한 기초조사(II)



환경연구 활용 빈도가 높은 온라인 데이터 Source 파악

◆ KEI 발간된 문서 DB의 문서들로부터 데이터 리스트 추출

– 전처리

→ 정규표현식 활용 데이터 리스트 추출

→ 추가규칙 적용

→ 문서 카테고리별 카운트

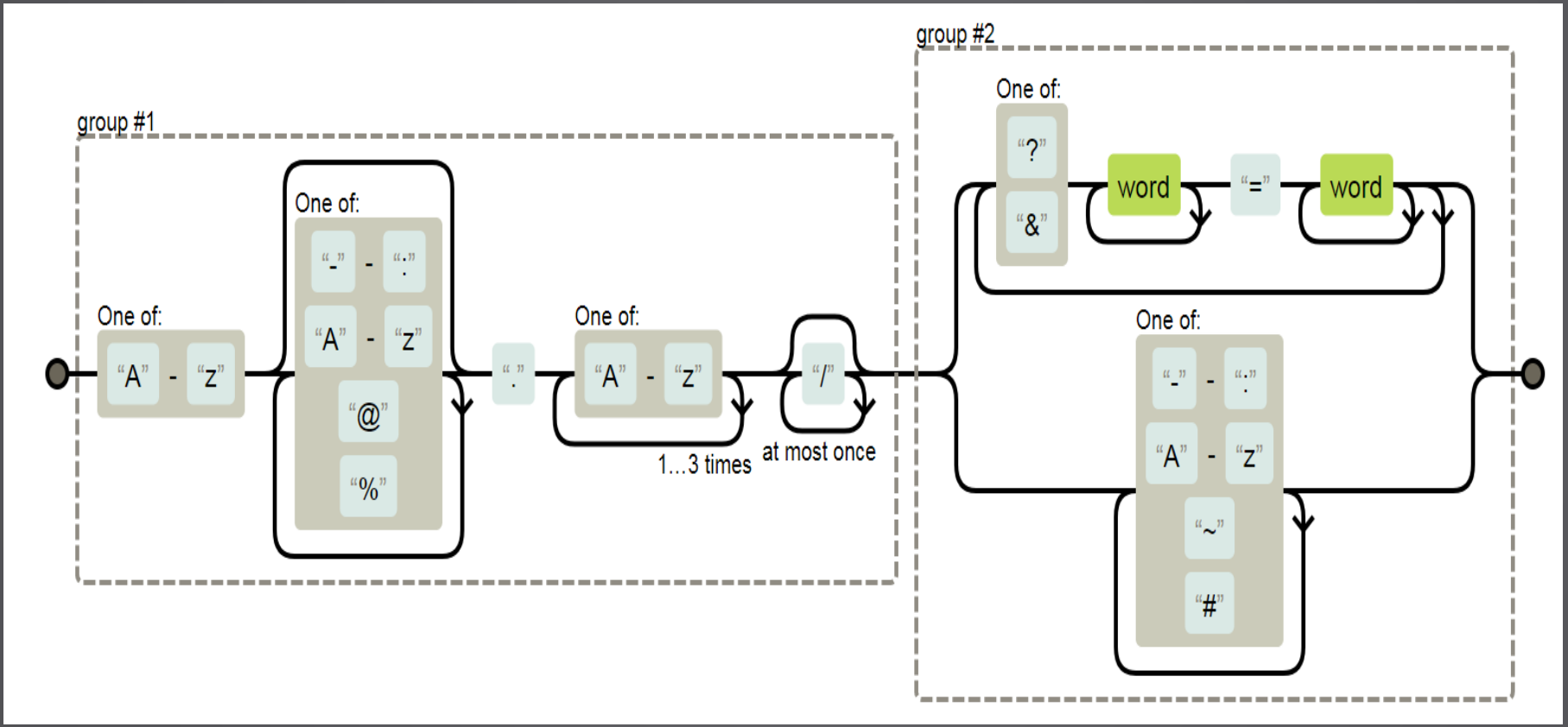
1단계: 정규표현식

- 정규표현식 핵심규칙
  - 영어로 시작
  - 콤마(.) 이후 1글자 – 3글자의 영어로 구성된 부분
    예) .kr .com 등
  - 그 이후에 내용이 있을경우 반드시 '/'

2단계: 추가 규칙

- 추가 규칙
  1. 프로토콜이 없는 경우 -> http:// 추가
  2. https://, http:// -> https://로 통일
  3. .kr, .com, .gov 등 국가기관 코드 없는 경우 제외

STEP 1 : 정규표현식


group #1
One of:
"A" - "z"
One of:
"-" - "."
"A" - "z"
"@"
"%"
"."
One of:
"A" - "z"
1...3 times
"/"
at most once
group #2
One of:
"?"
"&"
word
"="
word
One of:
"-" - "."
"A" - "z"
"~"
"#"

Open Data Map 카테고리 분석

각 카테고리 별 데이터가 사용된 문서의 갯수

순위	데이터	환경정책	환경보건	물환경	대기환경	자원순환	환경영향평가	기후변화	자연환경	지구환경	기타	합	분류	연구자 검증
1	https://www.me.go.kr	20	4	20	9	11	15	19	26	5	20	149	종합	종합
2	http://www.law.go.kr	2	3	9	0	7	11	11	18	4	14	81	종합	종합
3	https://www.epa.gov	7	4	16	4	8	5	4	12	0	10	70	종합	종합
4	https://kosis.kr	1	5	10	1	5	1	18	2	1	20	64	종합	종합
5	http://water.nier.go.kr	2	0	25	0	2	8	6	5	1	5	55	물환경	물환경
6	https://www.wamis.go.kr	4	1	20	0	0	2	12	11	1	2	53	물환경	물환경
7	http://www.kma.go.kr	2	5	9	2	2	2	20	5	4	1	52	기후변화	기후변화,대기환경
8	https://www.eiass.go.kr	0	1	1	0	3	21	6	6	2	9	50	환경영향평가	환경영향평가
9	http://www.kosis.kr	1	4	5	0	1	3	13	3	1	11	42	기후변화, 기타	종합
10	https://www.index.go.kr	0	0	5	0	3	2	14	7	2	7	40	기후변화	종합
11	https://www.airkorea.or.kr	3	8	2	1	2	1	8	3	3	5	37	기후변화, 환경보건	기후변화, 대기환경
12	https://egis.me.go.kr	0	0	2	0	1	6	8	14	2	4	37	자연환경, 기후변화	종합
13	http://www.env.go.jp	4	2	1	4	3	4	2	3	4	7	35	종합	종합
14	http://www.moleg.go.kr	4	0	4	1	3	6	3	10	0	3	34	종합	종합
15	http://kosis.nso.go.kr	10	1	2	4	1	2	1	8	2	1	32	환경정책	종합
16	https://ecos.bok.or.kr	3	1	1	0	4	0	8	0	1	14	32	기타	기타
17	http://www.nier.go.kr	2	0	4	1	1	2	9	5	0	7	31	종합	종합
18	https://www.gims.go.kr	1	1	10	0	0	3	4	7	1	0	27	물환경, 자연환경	물환경
19	http://www.nso.go.kr	17	1	0	0	2	0	2	2	2	1	27	환경정책	종합
20	http://airemiss.nier.go.kr	2	3	0	2	0	1	8	0	0	8	24	기후변화, 기타	기후변화, 대기환경

Open Data Map 웹 서비스 구축

◆ 카테고리별 데이터 소스(웹 사이트) 출력을 위한 랭킹 알고리즘 적용

- 데이터소스가 각 카테고리에 등장한 빈도수 활용
- 대부분의 카테고리에서 자주 등장한 데이터소스에 대해서는 페널티 부여
- $S_{nk} = D_{nk} / \log(\sum_{i=1}^{N} D_{ni})$: 점수 계산식
  - S_{nk} : n번째 데이터소스의 k번째 카테고리에서의 점수
  - D_{nk} : n번째 데이터소스의 k번째 카테고리 문서에서의 빈도수
  - $\log(\sum_{i=1}^{N} D_{ni})$: n번째 데이터소스의 각 카테고리의 문서에서 나타난 빈도수의 합 (페널티로 사용)

◆ Open Data Map 웹 서비스 구축 : R shiny 활용

- 카테고리 별 순위화된 데이터 소스 출력
- 데이터 소스에 대한 순위, 웹사이트 주소, 제목, 설명, 한국어 키워드, 영문 키워드 제공
- 사용자 편의를 위한 검색기능, 출력 문서 개수 설정 등 추가

Open Data Map 웹 인터페이스

환경 카테고리 선택

검색

기후변화 | 대기환경 | 물환경 | 자연환경 | 자원순환 | 지구환경 | 환경보건 | 환경영향평가 | 환경정책 | 기타

Show 10 entries 페이지당 출력 개수

Search:

Rank	Website	타이틀	설명	KR_Keywords	EN_Keywords
1	http://www.kma.go.kr	'하늘을 친구처럼, 국민을 하늘처럼' 기상청	대한민국의 날씨와 기후에 대해 조사를 하고 관찰을 하여 앞으로 어떤 날씨가 있을지 예측하는 기관	환경영향평가, 전략환경영향평가, GIS, 환경 자료 조사, 지리 정보,만경강, 용담댐, 외부유입수량, 하천유지유량, 증가방류,저탄소 사회, 소비행태, 구조분해분석(SDA), 다지역...	Urban ecology,Health,Socioeconomic valuation,Climatic changes,Urban poor,Water resources development...
2	https://kosis.kr	국가통계포털	국가승인통계 전체를 데이타베이스로 한 곳에 구축하여 국민들이 한 번의 접속으로 쉽고 편리하게 통계자료를 이용할 수 있도록 지원하는 이용자 중심의 One-Stop 국가통계포털 서비스	지역 기후경쟁력, 기후변화 산업 영향, 기후변화 리스크, 기후변화 적응대책,초미세먼지, 건강영향, 대기환경규제지역, 배출량 전망, 관리 방향,물환경목표, 오염원변화, 수질변화, 수...	Regional climate competitiveness, Climate change impacts on industry, Climate change risk, Climate c...
3	https://www.me.go.kr	환경부	자연환경, 생활환경의 보전, 환경오염방지, 수자원의 보전·이용 및 개발에 관한 사무를 관장하는 대한민국의 중앙행정기관	환경평가, 육상풍력발전, 수상태양광발전, 개발 잠재량, 신재생에너지,녹색생활, 교통, 건물, 여성, 녹색의식,사후환경영향조사, 사후모니터링, 전과정 환경영향평가, 사후환경관리, 협...	Environmental Assessment, Onshore Wind Power Generation, Floating Photovoltaic Power Generation, Ren...
4	https://www.climate.go.kr	기후정보포털	기후정보포털은 기후변화 과학정보를 중점적으로 소개하며, 영향평가와 정책에 관련된 정보를 동시에 제공하고, 기후변화 연구에 필요한 자료들의 데이터 베이스를 구축하여 기후변화 탐지 모...	적응보고제도, 사회기반시설 기후변화 영향, 기후변화 리스크평가,가뭄, 기후변화, 적응대책, 피해 유형, 가뭄 리스크 관리 유형,기후변화, 취약생태계, 구상나무, 환경안보, 기후변화...	Drought, Climate Change, Adaptation Policy, Types of Drought Effects, Types of Drought Risk Manageme...
5	https://www.index.go.kr	국가지표체계	정부기관에서 엄선한 740개 지표를 통하여 다양한 방면에서의 우리나라 현위치를 보여줍니다. e-나라지표 시스템에서 제공하는 지표들은 국가 공식 승인 통계자료뿐만 아니라, 현황이나 ...	갯벌매립사업, 환경영향평가, 환경평가 모니터링, 사후환경관리,국토개발정책, 공간환경정책, 국토계획, 환경계획,대심도 지하공간, 도심지역, 지속 가능성, 지반환경영향, 위해성 소통,...	Reclamation Project, Environmental Impact Assessment, Environmental Assessment Monitoring, Post-Envi...
6	http://kostat.go.kr	통계청	국가통계 발전을 선도하며, 신뢰받는 통계생산으로, 각 경제주체에게 유용한 통계정보 제공	기후변화 적응, 민간부문, 산업계, 리스크, 적응 역량,석탄화력, LNG 복합화력, 탄소중립 프로그램, 배출허용총량,기후변화 리스크 평가, 리스크 목록, 기후변화 영향, 기후변화 ...	Climate Change Adaptation, Private Sector, Industrial Sector, Risk, Adaptative Capacity,thermal powe...
7	http://www.safekorea.go.kr	국민재난안전포털	대국민 재난정보의 일관되고 통합적인 제공을 위해, 기존 국가재난정보센터, 재난심리 상담정보센터, 재난훈련시스템을 통합·제공	기후변화, 사회경제적 영향, 분류 체계, 주요 지표, 통계,기후변화, 사회경제적 영향, 분류 체계, 주요 지표, 통계,기후변화, 취약성, 평가지표, 영향분석,도시 회복력, 도시 기...	Climate Change, Socio-economic Impacts, Framework, Indicators, Dataset, Adaptation,Climate Change, S...
8	http://www.kosis.kr	KOSIS 국가통계포털	국가통계포털(KOSIS, Korean Statistical Information Service)은 국내·국제·북한의 주요 통계를 한 곳에 모아 이용자가 원하는 통계를 한 번에 찾을...	저탄소사회 정책, 소비행태 분석, 온실가스 유발배출량, 배출증감 요인분해, 다지역투입산출분석, 연산가능 일반균형모형,개발사업, 수질오염, 인식조사, 지방자치단체, 총량관리,환경가치...	ecology,Urban poor,Development Project, Local Government, Pollution Load, Staff Survey, Water Qualit...
9	http://www.wamis.go.kr	국가수자원관리 종합정보	물관련 정보를 대국민 서비스하기 위해 구축 운영되고 있는 인터넷 기반의 포털 시스템	지중 환경, 지중 자연환경, 지중 생활환경, 지중 환경 구성요소, 지중 환경 가치,지하수 오염, 관리, 정화, 제도적 지원, 사후조치,물-효율적 경제사회, 물이용 인식, 물이용 갈...	Subsurface Environment, Subsurface Natural Environment, Subsurface Living Environment, Subsurface En...
10	http://www.greengrowth.go.kr	녹색성장위원회	녹색성장위원회는 정부의 녹색성장 정책을 심의·조율하고, 사회 각계의 다양한 의견을 수렴하는 국무총리 소속 기구입니다.	녹색생활, 교통, 건물, 여성, 녹색의식,기후변화, 에너지, 온실가스, 감축, 국가 계획,환경협력, 중견국협의체, 중견국 외교, 한?ASEAN 환경협력, 거버넌스,기후변화 적응, ...	Climate Change, Greenhouse-gas, Energy, Emission Reduction, National Plans,Environmental Cooperation...

Showing 1 to 10 of 1,000 entries

Previous 1 2 3 4 5 … 100 Next

향후 계획

- 데이터 소스에 대한 '설명(annotation)' 작업 및 세부 조정
  - 데이터 소스 접근 가능 여부 업데이트 필요 (잘못되거나 없어진 데이터 소스 삭제/대체 필요)
  - 데이터 소스 제목, 설명 등에 대한 annotation 작업 필요
  - 랭킹 알고리즘 및 카테고리 세부 조정
- 원내 서비스를 위한 서비스 배포 관련 작업
  - 웹 서비스 테스트
  - 서비스 배포

Open Data Map 보완 : 데이터 직접 수집

◆ 데이터 수집-저장 소프트웨어(수집기) 강화

– [비정형_텍스트_웹] 네이버 뉴스 : 이 시각 주요 뉴스 2종(텍스트, 포토)

- 2009년 6월 7일(서비스 개시일) 이후 일별 수집

– [반정형_텍스트_파일] 한국언론진흥재단 빅카인즈 : 뉴스 매체의 뉴스 메타DB(엑셀 파일)

- 중앙지(8), 경제지(7), 지역종합지(27), 방송사(4), 전문지(2) 개별* / 전체 수집

– [비정형_텍스트_웹] 한국학술정보 KISS : 제목, 저자, 발행기관, 초록(국문, 영문) 등 메타 데이터

- 발행기관(1,747)의 간행물(4,508) 개별 / 전체* 수집

※ 깃허브 공개 예정(분석플랫폼 서버에는 탑재) / *확인 예정

※ 데이터 형태 구분

- 정형 : 제공되는 매뉴얼(또는 메타 데이터)을 통해 식별이 가능함(공공데이터포털의 OpenAPI)
- 반정형 : 제공되는 매뉴얼(또는 메타 데이터)은 없으나 데이터가 식별이 가능함(에어코리아의 미세먼지 엑셀파일)
- 비정형 : 데이터 식별이 불가능함(일반적인 인터넷 뉴스나 트위터, 페이스북과 같은 SNS 데이터)

(2) 환경 빅데이터 분석플랫폼 : 설계 및 안정화

- 연구 내용: 연구자가 대용량 데이터 분석을 수행할 수 있는 분석플랫폼 설계 및 시험가동
  - 기존 알고리듬을 이용하는 사용자(User) : 웹 개발환경 제공
  - 알고리듬 개발자 : 터미널(Command Line Interface) 접근, 개인환경 구성 허용
  - 2017년 도입한 서버 1기 (56-core, 192GB, 28TB) 에 구현

- 주기적 점검 및 SW upgrade로 오류 발생을 억제
  - 오류 처리 횟수: 1월 15회 → 7월 이후 1회

환경 빅데이터 분석플랫폼 시범운영 서비스입니다.

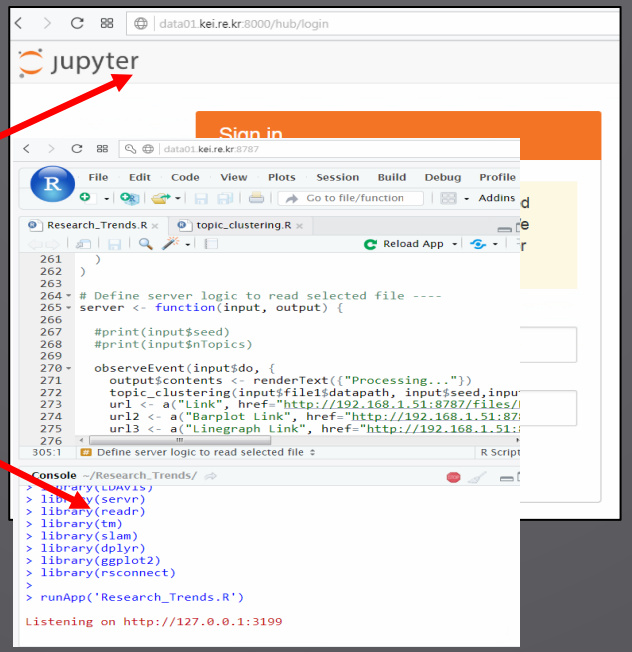
주피터 노트북 - http://data01.kei.re.kr:8000/
파이썬은 터미널(SSH) 사용을 권장드리며
virtualenv로 개인설정하셔서 주피터 노트북 구성하시기 바랍니다.

RStudio Server - http://data01.kei.re.kr:8787/

환경관련 데이터 서비스 - http://data01.kei.re.kr:8080/

한글 파일명 사용은 지양해 주시고
필요시 FTP 소프트웨어를 사용하시기 바랍니다. - Filezilla

문의는 dataq@kei.re.kr로 주시기 바랍니다.




```
b3nn9@DataLX01:~$ source nlpNaver/bin/activate
(nlpNaver) b3nn9@DataLX01:~$ pip list
Package            Version
------------------ ---------
backcall           0.1.0
bleach             2.1.3
boto               2.48.0
boto3              1.7.30
botocore           1.10.30
bz2file            0.98
certifi            2018.4.16
chardet            3.0.4
cycler             0.10.0
decorator          4.3.0
docutils           0.14
entrypoints        0.2.3
gensim             3.4.0
html5lib           1.0.1
idna               2.6
imread             0.6.1
ipykernel          4.8.2
ipyparallel        6.1.1
ipython            6.4.0
ipython-genutils   0.2.0
jedi               0.12.0
Jinja2             2.10
jmespath           0.9.3
JPype1             0.6.3
jsonschema         2.6.0
jupyter-client     5.2.3
jupyter-core       4.4.0
```

4. 환경 빅데이터 서비스

연구동향 파악 서비스

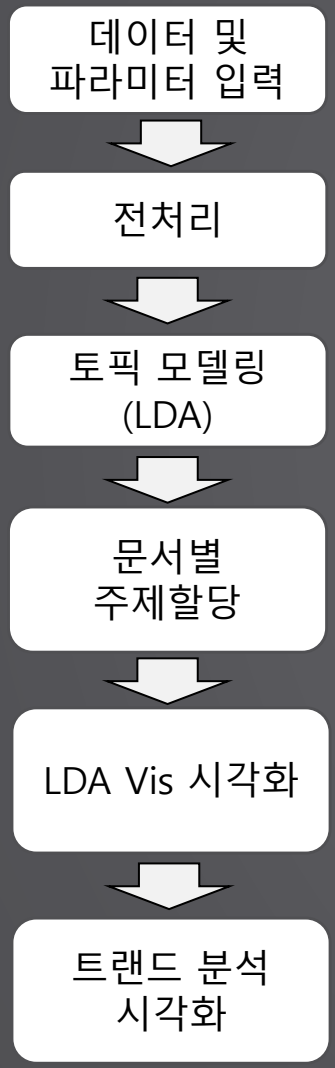
1. 연구 주제 동향 파악 서비스

2. 연구 키워드 동향 파악 서비스

(1) 연구주제 동향 파악 서비스: LDA 토픽 모델링

- LDA 토픽 모델링 및 연구동향 분석 서비스
  - 텍스트(문서)들의 집합에 포함된 주제를 분석하고, 이를 토대로 년도별 동향을 파악하는 서비스
  - 2017년 '텍스트마이닝을 이용한 KEI 연구동향 분석' 연구의 소스코드 이용
- 동일 포맷 다른 데이터 입력 시 토픽 모델링 결과를 출력할 수 있도록 기존 코드 일반화
  - 코드 간소화 : 파일과 인자를 입력 받으면 한번의 실행으로 모든 결과가 도출되게 하는 작업
  - 형태소 분석기 추가, 파라미터 지정 부분 분리, 시각화 관련 부분 수정 등
- 웹서비스 구축: R shiny를 활용한 웹 기반 GUI 인터페이스 구축
  - 파일 업로드 기능
  - 년도별 연구동향 분석 기능
  - 문서별 토픽 분석결과 및 문서들의 토픽 분포 추가
- 향후 계획 : 시각화 결과 정교화 후 beta 테스트를 거쳐서 원내 배포 추진

연구주제 동향 파악 분석 과정



CSV 파일 입력(업로드), 토픽의 개수, 시드값 입력 (사용자 파라미터)

형태소 분석(명사추출), 불용어/특수문자 제거, 2글자 이상 한글 키워드만 활용, 문서-단어 행렬 계산

토픽-단어, 문서별-토픽 행렬 계산

각 문서에 기준으로 최대 확률을 가지는 토픽으로 문서 주제 할당

시각화에 필요한 토픽-단어 행렬, 문서별-토픽 행렬, 문서수,
사전(Dictionary) 정보, PHI (단어별 문서 기여정도 설정) 입력

토픽 비율 분포 및 년도별 연구동향분석 시각화

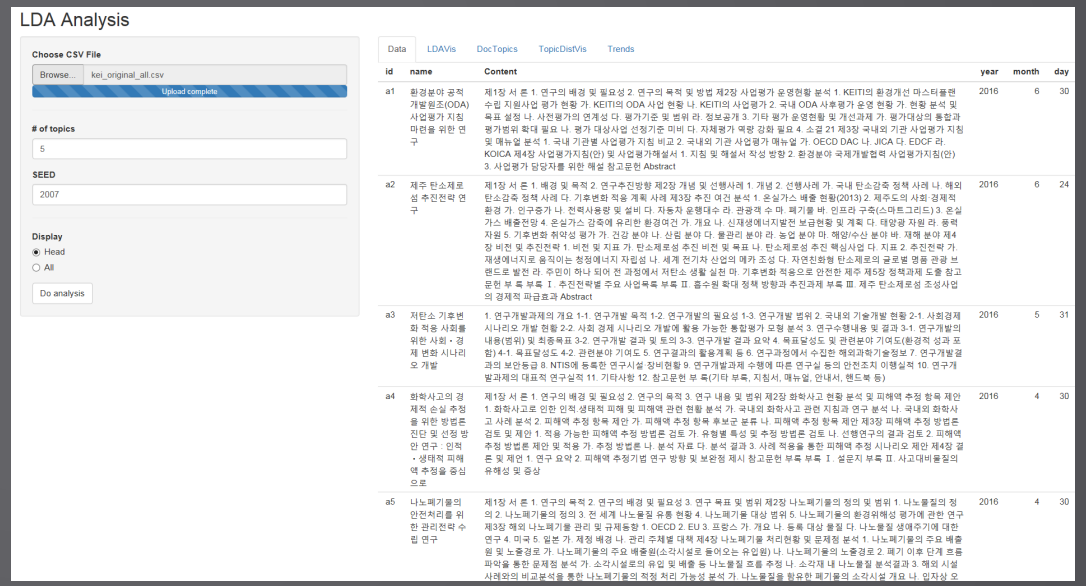
연구 주제 동향 서비스 GUI : 입력

입력 파일 포맷 (csv)

LDA 토픽 모델링 웹서비스

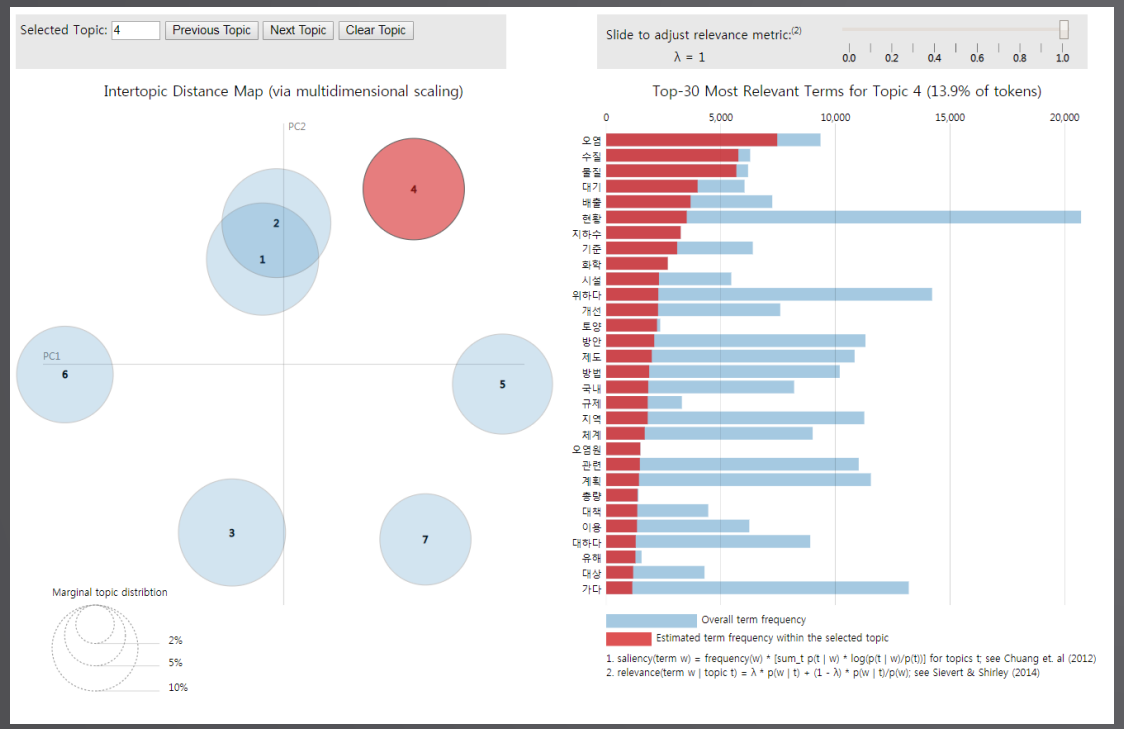
id	name	Content	year	month	day
a1	환경분야 공적개발원조(ODA) 사업평가 지침 마련을 위한 연구	제1장 서 론 1. 연구의 배경 및 필요성 2. 연구의 목적 및 방법 제2	2016	6	30
a2	제주 탄소제로섬 추진전략 연구	제1장 서 론 1. 배경 및 목적 2. 연구추진방향 제2장 개념 및 선행	2016	6	24
a3	저탄소 기후변화 적응 사회를 위한 사회・경제 변화 시나리오 개	1. 연구개발과제의 개요 1-1. 연구개발 목적 1-2. 연구개발의 필요	2016	5	31
a4	화학사고의 경제적 손실 추정을 위한 방법론 진단 및 선정 방안 (	제1장 서 론 1. 연구의 배경 및 필요성 2. 연구의 목적 3. 연구 내용	2016	4	30
a5	나노폐기물의 안전처리를 위한 관리전략 수립 연구	제1장 서 론 1. 연구의 목적 2. 연구의 배경 및 필요성 3. 연구 목표	2016	4	30
a6	가뭄 단계에 따른 적응형 가뭄관리정책 연구 : 지역 차원의 비구	제1장 서 론 1. 연구 배경 및 목적 2. 연구 내용 및 방법 제2장 201	2016	3	31
a7	국내 능선축 GIS기반 통합관리시스템 개발	1. 연구개발과제의 개요 1-1. 연구개발 목적 1-2. 연구개발의 필요	2016	3	31
a8	기후변화에 따른 건강영향 평가・적응 기술 및 정책지원 시스템	연구개발과제의 개요 1-1. 연구개발 목적 1-2. 연구개발의 필요성	2016	2	28
a9	Post-2020 신기후체계 협상 적응의제 대응방안 연구	제1장 서론 1. 연구의 배경 및 필요성 2. 연구방향 3. 연구수행 방	2015	12	31
a10	사물인터넷(IoT)을 활용한 스마트 물환경관리 방안 및 정책기반	제1장 서 론 1. 연구의 배경 및 필요성 가. 사물인터넷: 혁명적 핵	2016	10	31
a11	중국의 '일대일로(一帶一路)' 대응 유라시아 지역 환경전략 연구	제1장 서론 1. 연구의 필요성 및 목적 가. 연구의 배경 및 필요성	2016	10	31
a12	도시의 기후 회복력 확보를 위한 공간단위별 평가체계 및 모형 기	제1장 서론 1. 연구의 배경 및 목적 2. 연구의 대상 3. 연구의 구조	2016	10	31
a13	지역기반 환경보건정책 지원 방안 연구(II)	제1장 서 론 1. 연구의 배경 및 필요성 2. 연구의 범위 제2장 국내	2016	10	31
a14	신기후체제의 기후변화 적응 및 손실과 피해에 관한 대응방안	제1장 서 론 1. 연구의 배경 및 필요성 2. 연구의 목적 제2장 파리	2016	9	30
a15	대기환경비용을 고려한 친환경차 구매보조금 실효성 제고 연구 :	제1장 서 론 1. 연구의 필요성 및 목적 2. 연구의 범위 및 방법 3.	2016	9	30
a16	실도로에서 경유차의 대기오염물질 초과 배출에 따른 사회적 비	제1장 서 론 1. 연구의 필요성 및 목적 2. 연구 범위 3. 연구 내용	2016	9	22
a17	토양정화 곤란 부지의 최적 관리방안 연구	제1장 서 론 1. 연구의 배경 및 목적 2. 연구의 접근 및 방법 제2장	2016	8	31
a18	제3차 국가생물다양성전략 이행상황 점검	제 1 장 서 론 1. 과업 배경 및 필요성 2. 과업 내용 가. NBSAP 실	2016	8	30
a19	국가 지속가능성 평가 등에 관한 연구	지속가능발전 개념은 '87년 환경과 개발에 관한 세계위원회(WCE	2016	7	30
a20	유네스코 세계지질공원 운영 강화에 따른 국가지질공원제도의 기	제1장 서 론 1. 배경 2. 필요성 및 목적 3. 연구 범위 4. 연구 내용	2016	11	22
a21	시스템과 네트워크 이론을 활용한 미래 환경정책 방향 연구	제1장 서 론 1. 연구의 필요성 및 목적 2. 시스템과 네트워크 언어	2016	11	6
a22	국가 및 지역 미래성장동력에 대한 환경성 분석 및 환경영향평가	제1장 서 론 1. 연구의 배경 및 필요성 2. 연구의 목적 3. 연구의 내	2016	10	31
a23	지중환경관리를 위한 제도 개선방안 연구(II)	제1장 서 론 1. 연구의 배경 및 필요성 2. 연구의 목적 3. 연구의 내	2016	10	31
a24	정부3.0 기반 지역기피시설 주민수용성 평가 방안(II)	제1장 서 론 1. 연구의 배경 및 목적 2. 연구의 범위 및 방법 제2장	2016	10	31
a25	공간정보를 활용한 재해폐기물 성상별 최적 관리방안	제1장 서 론 1. 연구의 목적 2. 연구의 배경 및 필요성 3. 선행 연구	2016	10	31
a26	자원순환사회 전환 촉진을 위한 재활용산업 활성화 방안 : 재활용	제1장 서 론 1. 연구의 배경 및 목적 2. 연구의 범위 및 방법 제2장	2016	10	31
a27	폐자원흐름분석을 통한 전기·전자제품의 upcycling 활성화 방안	제1장 서 론 1. 연구의 필요성 및 목적 2. 연구의 범위 3. 연구의 내	2016	10	31
a28	사회적 투자수익률(SROI)을 고려한 물환경 인프라시설 투자방향	제1장 서 론 1. 연구의 배경 및 목적 2. 연구의 범위 및 구성 제2장	2016	10	31
a29	생태계서비스 기반의 자연자본 지속가능성 지수 개발 연구(Ⅰ)	제1장 서 론 1. 연구 배경 및 목적 2. 연구 범위 3. 연구 방법 제2장	2016	10	31
a30	근지표환경 임계영역(Critical Zones)의 환경적 중요성과 환경관리	제1장 서 론 1. 연구 배경 및 목적 2. 연구 내용 및 수행 체계 제2장	2016	12	6
a31	드론을 이용한 환경재난 사후대응 기술 및 연구동향 분석 연구	제1장 서 론 1. 연구 배경 및 목적 2. 연구 범위 및 방법 제2장 국내	2016	12	6
a32	건물부문의 환경부하 평가모형 개발을 위한 기초연구	제1장 서 론 1. 연구의 필요성 및 목적 2. 연구의 범위 제2장 건물	2016	12	6
a33	미래 고온환경 변화와 직종 간 임금격차 추정	제1장 서 론 제2장 선행연구 고찰 1. 기후변화와 노동자에 대한 영	2016	12	6
a34	자율주행 자동차의 친환경성 제고를 위한 기초 연구	제1장 서 론 1. 연구의 필요성 및 목적 2. 주요 내용 및 보고서 구성	2016	12	6

KEI 연구보고서 내용
1522건 (1993년~2016년)



연구 주제 동향 서비스 GUI : 토픽 추출 및 할당

LDA 토픽 모델링 결과



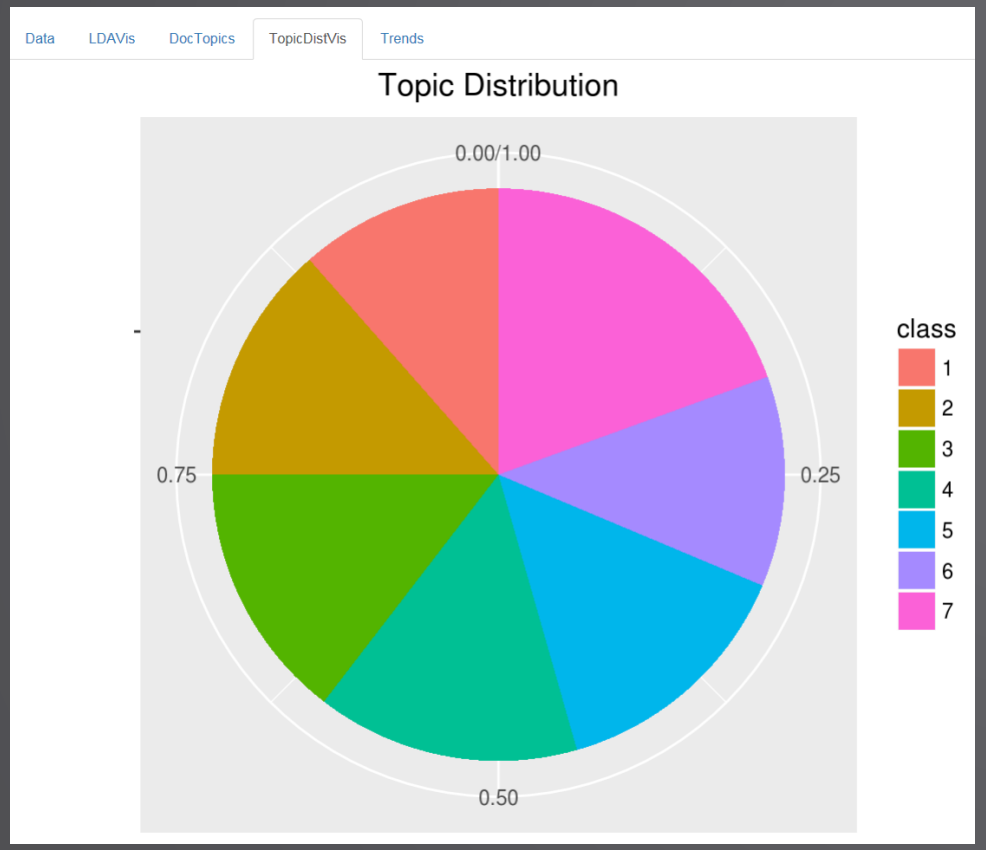
문서별 주제할당

Data | LDAVis | DocTopics | TopicDistVis | Trends

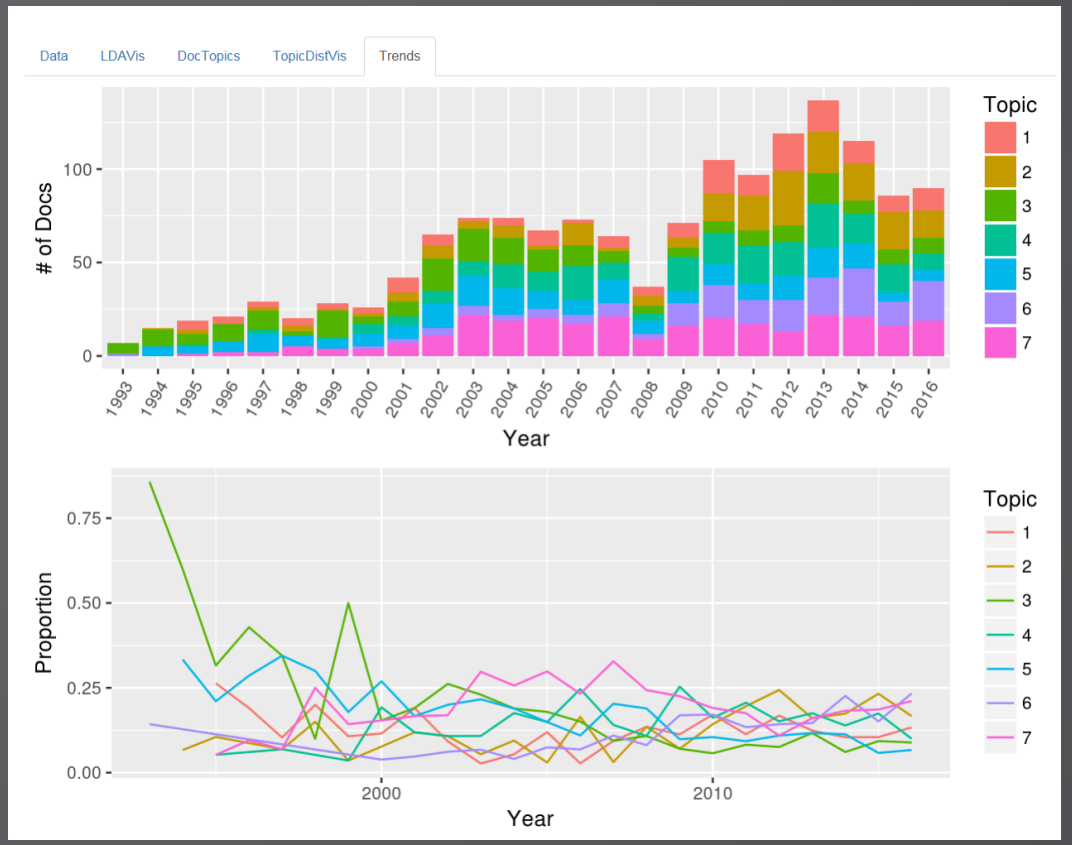
name	doc_topic
환경분야 공적개발원조(ODA) 사업평가 지침 마련을 위한 연구	4
제주 탄소제로섬 추진전략 연구	1
저탄소 기후변화 적응 사회를 위한 사회・경제 변화 시나리오 개발	2
화학사고의 경제적 손실 추정을 위한 방법론 진단 및 선정 방안 연구 : 인적・생태적 피해액 추정을 중심으로	4
나노폐기물의 안전처리를 위한 관리전략 수립 연구	5
가뭄 단계에 따른 적응형 가뭄관리정책 연구 : 지역 차원의 비구조적 가뭄대책을 중심으로	2
국내 능선죽 GIS기반 통합관리시스템 개발	2
기후변화에 따른 건강영향 평가・적응 기술 및 정책지원 시스템 개발(I-IV)	6
Post-2020 신기후체계 협상 적응의제 대응방안 연구	2
사물인터넷(IoT)을 활용한 스마트 물환경관리 방안 및 정책기반 마련 연구	7
중국의 '일대일로(一帶一路)' 대응 유라시아 지역 환경전략 연구	3
도시의 기후 회복력 확보를 위한 공간단위별 평가체계 및 모형 개발(II)	7
지역기반 환경보건정책 지원 방안 연구(II)	2
신기후체제의 기후변화 적응 및 손실과 피해에 관한 대응방안	2
대기환경비용을 고려한 친환경차 구매보조금 실효성 제고 연구 : 차종별 적정 보조금 수준 분석을 중심으로	1
실도로에서 경유차의 대기오염물질 초과 배출에 따른 사회적 비용 연구	1
토양정화 곤란 부지의 최적 관리방안 연구	5
제3차 국가생물다양성전략 이행상황 점검	2
국가 지속가능성 평가 등에 관한 연구	2
유네스코 세계지질공원 운영 강화에 따른 국가지질공원제도의 개선방안 연구	7
시스템과 네트워크 이론을 활용한 미래 환경정책 방향 연구	7
국가 및 지역 미래성장동력에 대한 환경성 분석 및 환경영향평가 대비방안 연구	1

연구 주제 동향 서비스 GUI : 토픽 비중

토픽 분포도



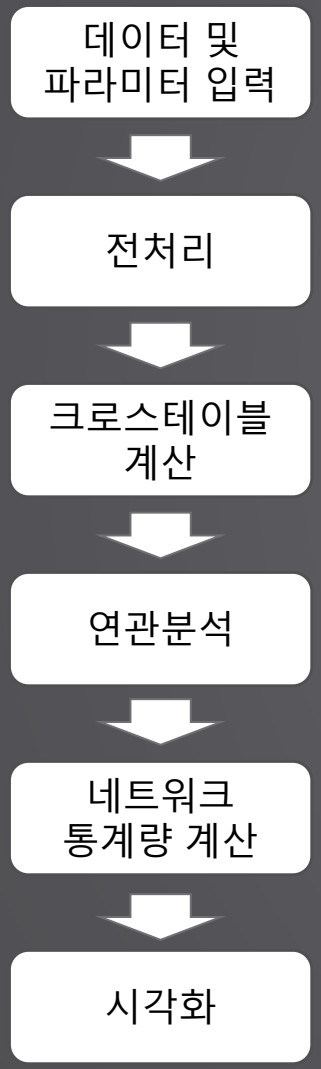
토픽 분포 동향



(2) 연구 키워드 동향 서비스: 키워드 네트워크 분석

- 키워드 네트워크 분석 서비스
  - 환경 텍스트의 키워드를 파악하고 키워드 사이의 연관 빈도가 높은 단어들의 네트워크를 도출
- 동일 포맷 다른 데이터 입력 시 연관분석 결과를 출력할 수 있도록 기존 소스코드 일반화
  - 코드 간소화 : 파일과 인자를 입력 받으면 한번의 실행으로 모든 결과가 도출되게 하는 작업
  - 연관분석 핵심코드 정리 및 함수화 및 각종 인자 파라미터화 (연관분석 및 시각화 부분 : 차수, 중심성, 근접중심성 등 고려 및 활용)
  - 규칙 수정 : 일부 오류 수정, 빈도수 규칙 추가, LIFT 기준으로 규칙 정렬
- 웹서비스 구축: R shiny를 활용한 웹 기반 GUI 인터페이스 구축
  - 파일 업로드 기능 추가
  - 지지도(support), 신뢰도(confidence), 사용할 규칙 수 입력 기능
  - 날짜 및 카테고리 필터링 기능 추가
- 향후 계획 : 시각화 결과 정교화 후 beta 테스트를 거쳐서 원내 배포 추진

연구 키워드 동향 분석 과정



CSV 파일 입력(업로드), 연관분석 파라미터(Support, Confidence)입력, 필터링 정보 (시작날짜, 끝날짜, 카테고리) 입력

형태소 분석(명사추출), 불용어/특수문자 제거, 2글자 이상 한글 키워드만 활용

단어-단어 동시발생(co-occurrence) 테이블 계산

Apriori 알고리즘 수행

시각화에 필요한 키워드 노드 차수(degree), 근접 중심성(closeness),
매개중심성(betweenness) 계산

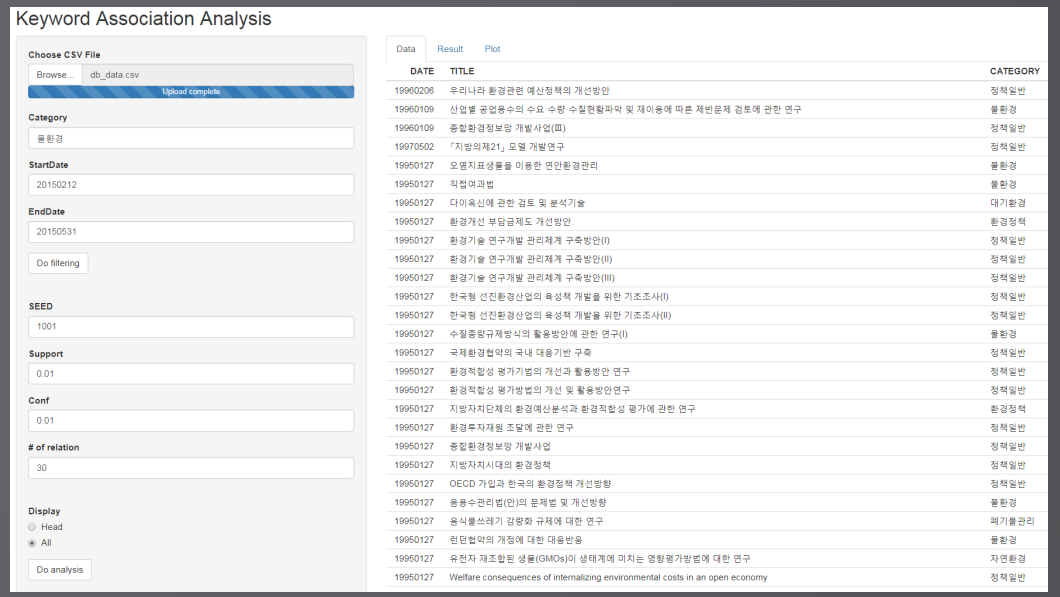
주제의 비율 분포 및 년도별 연구 동향 분석 시각화

연구 키워드 동향 서비스 GUI : 자료 입력

입력 파일 포맷 (csv)

DATE	TITLE	CATEGORY
19960206	우리나라 환경관련 예산정책의 개선방안	정책일반
19960109	산업별 공업용수의 수요·수량·수질현황파악 및 재이용에 따른 제반문제 검토에 관한 연구	물환경
19960109	종합환경정보망 개발사업(Ⅲ)	정책일반
19970502	「지방의제21」 모델 개발연구	정책일반
19950127	오염지표생물을 이용한 연안환경관리	물환경
19950127	직접여과법	물환경
19950127	다이옥신에 관한 검토 및 분석기술	대기환경
19950127	환경개선 부담금제도 개선방안.	환경정책
19950127	환경기술 연구개발 관리체계 구축방안(Ⅰ)	정책일반
19950127	환경기술 연구개발 관리체계 구축방안(Ⅱ)	정책일반
19950127	환경기술 연구개발 관리체계 구축방안(Ⅲ)	정책일반
19950127	한국형 선진환경산업의 육성책 개발을 위한 기초조사(Ⅰ)	정책일반
19950127	한국형 선진환경산업의 육성책 개발을 위한 기초조사(Ⅱ)	정책일반
19950127	수질총량규제방식의 활용방안에 관한 연구(Ⅰ)	물환경
19950127	국제환경협약의 국내 대응기반 구축	정책일반
19950127	환경적합성 평가기법의 개선과 활용방안 연구	정책일반
19950127	환경적합성 평가방법의 개선 및 활용방안연구	정책일반
19950127	지방자치단체의 환경예산분석과 환경적합성 평가에 관한 연구	환경정책
19950127	환경투자재원 조달에 관한 연구	정책일반
19950127	종합환경정보망 개발사업	정책일반
19950127	지방자치시대의 환경정책	정책일반
19950127	OECD 가입과 한국의 환경정책 개선방향	정책일반
19950127	음용수관리법(안)의 문제법 및 개선방향	물환경
19950127	음식물쓰레기 감량화 규제에 대한 연구	폐기물관리

입력 자료

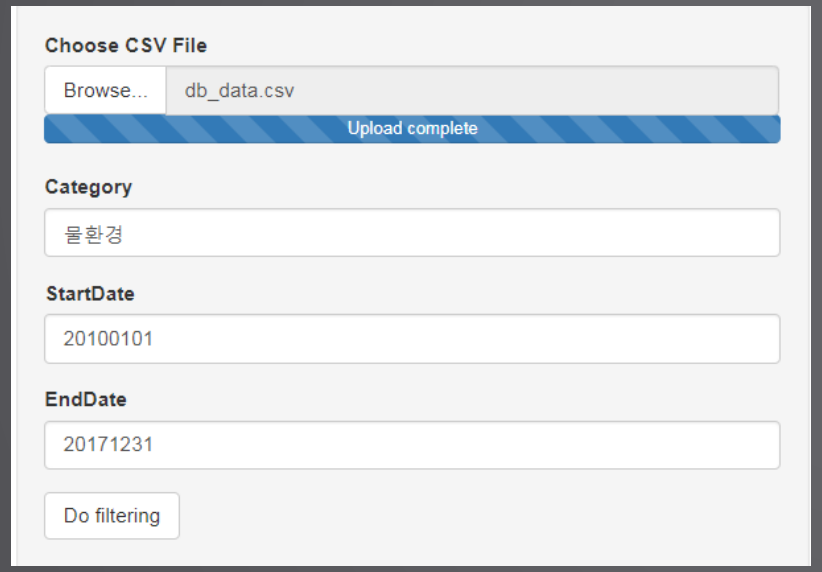


도서관 연구보고서 DB
연구제목 1967건
(1993년~2018년 5월)

연구 키워드 동향 서비스 GUI : 데이터 필터링

데이터 필터링

필터링 된 데이터



Data | Result | Plot

DATE	TITLE	CATEGORY
20111220	상하수도사업 민영화 기본계획 수립연구	물환경
20111220	낙동강수계 오염총량관리제 시행방안 연구	물환경
20100223	수질오염총량관리를 위한 배출거래제 적용방안 연구	물환경
20100223	기후변화 대응을 위한 물환경 관리전략 및 정책방향(I)	물환경
20100224	기후변화 연동 4대강 유역 지하수 함양 및 이용가능량 산정 기법 개발	물환경
20100224	셀렝게유역 통합물환경관리모델 개발 연구 II	물환경
20100224	합리적인 수리권 및 수자원에의 기여와 보상체계 연구	물환경
20100224	남·북한 공유하천의 관리 현황과 물안보 확보 방향	물환경
20100224	4대강 살리기 사업을 위한 필요 전문기술인력 추정	물환경
20100224	다목적댐 상류 폐광산 등 비점오염원 관리방안 연구	물환경
20100224	도서지역 용수공급 체계에 관한 고찰	물환경
20100226	물환경 기준의 통합적 관리방안에 관한 연구	물환경
20100226	4대강 관련 법률 및 제도의 현황분석과 효율적 개선방안	물환경
20100226	지역단위 하수재이용 활성화를 위한 기초연구	물환경
20100226	실시간 수질 모니터링 및 모델링 체계에 관한 고찰	물환경

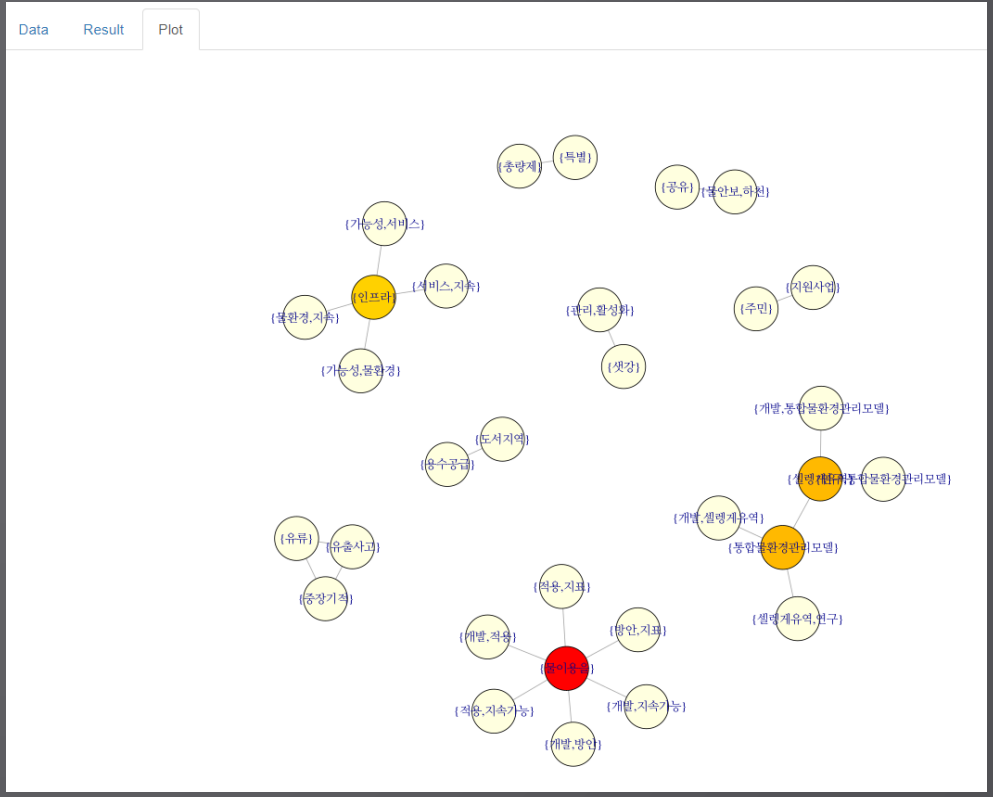
연구 키워드 동향 서비스 GUI : 네트워크 분석 결과

키워드 연관 분석 결과

Data | Result | Plot

lhs	V2	rhs	support	confidence	lift	count
{셀렝게유역}	=>	{통합물환경관리모델}	0.01398601	1.00000000	71.5000000	2
{통합물환경관리모델}	=>	{셀렝게유역}	0.01398601	1.00000000	71.5000000	2
{용수공급}	=>	{도서지역}	0.01398601	1.00000000	71.5000000	2
{도서지역}	=>	{용수공급}	0.01398601	1.00000000	71.5000000	2
{지원사업}	=>	{주민}	0.01398601	1.00000000	71.5000000	2
{주민}	=>	{지원사업}	0.01398601	1.00000000	71.5000000	2
{유류}	=>	{유출사고}	0.01398601	1.00000000	71.5000000	2
{유출사고}	=>	{유류}	0.01398601	1.00000000	71.5000000	2
{유류}	=>	{중장기적}	0.01398601	1.00000000	71.5000000	2
{중장기적}	=>	{유류}	0.01398601	1.00000000	71.5000000	2
{유출사고}	=>	{중장기적}	0.01398601	1.00000000	71.5000000	2
{중장기적}	=>	{유출사고}	0.01398601	1.00000000	71.5000000	2
{특별}	=>	{총량제}	0.01398601	1.00000000	71.5000000	2
{총량제}	=>	{특별}	0.01398601	1.00000000	71.5000000	2
{개발,셀렝게유역}	=>	{통합물환경관리모델}	0.01398601	1.00000000	71.5000000	2
{개발,통합물환경관리모델}	=>	{셀렝게유역}	0.01398601	1.00000000	71.5000000	2
{셀렝게유역,연구}	=>	{통합물환경관리모델}	0.01398601	1.00000000	71.5000000	2
{연구,통합물환경관리모델}	=>	{셀렝게유역}	0.01398601	1.00000000	71.5000000	2

키워드 연관 분석 시각화



5. 요약 및 시사점

연구결과 요약 : 4개 알고리듬, 플랫폼 설계, 2개 서비스

- 환경 빅데이터 연구: 통상적 방법론보다 예측오차가 작은 4개 알고리듬 신규 구축
  - 대용량/비정형 : 미세먼지 오염도 예측, 기후변화 SNS 감성분류기 → 예측오차 획기적 축소
  - 연구영역 확대 : 수질오염도 예측, 지하철 승하차 인원 예측
  - 기후변화 사전, 기후변화 SNS, 고령 COPD 환자 건강 Data 등 신규 자료 확충
- 환경 빅데이터 플랫폼 : 자료 수집 Gateway 구축 및 대용량 자료 분석 기능 확보
  - Open Data Map : 자료 수집 Gateway
  - 대용량 자료 분석 : 분석 플랫폼 Web 환경/CLI 환경
    - 2018년 기후변화 SNS 감성분류기 구축 과제 수행 시 활용
- 환경 빅데이터 서비스 : 환경 빅데이터 연구 성과 활용 연구정보 서비스 구축
  - 연구주제 동향 서비스: 임의의 입력 자료에 대한 LDA 토픽 클러스터 분석 수행
  - 연구 키워드 동향 서비스 : 임의의 입력 자료에 대한 키워드 분석 수행

연구성과 : 빅데이터 방법론의 정교함/재활용 가능성 확인

- 빅데이터 연구방법론의 장점: 예측의 정교함/미지의 패턴 파악/ 연구결과의 재활용 및 확산 가능성

- 정교함: 빅데이터 분석방법론으로 예측오차 축소가 가능함을 다양한 분야에서 확인
  - 대용량/ 비정형 데이터 분석에서 예측오차를 획기적으로 축소
    - 미세먼지 오염도 예측 오차 축소 및 정교한 기후변화 SNS 감성분류기 개발
  - 수질오염/수용체 반응 분석으로 빅데이터 방법론 확대
    - 적은 자료에서도 극한값이 존재하는 경우 기계학습 방법론으로 예측오차 축소 가능 확인
  - 재활용/범용성:연구결과의 재활용 및 확장 가능성을 활용하여 연구정보 서비스 구축
  - 일회성 연구에 사용한 2017년 개발 알고리듬을 실시간 분석 가능한 알고리듬으로 확대 개편

- 자료 수집 및 분석 기능을 갖춘 빅데이터 플랫폼을 설계하고 시범운영
  - 수집 : Open Data Map/ 분석 : 분석 플랫폼

- 기후변화 사전, 기후변화 SNS, 고령 COPD 환자 건강 Data 등 신규 자료 확충

정책적 활용 : 환경오염 조기경보/환경관련 국민감성 파악

◆ 환경오염 조기경보 : 미세먼지 예측 알고리듬 (8시간), 클로로필-a 예측 알고리듬 (1주일)은 대응이 가능한 시간 여유가 있는 시점의 예측치를 제공

- 민간 공급 : 환경위험에 직접 대비 / 예측치를 가공한 종합 정보 서비스 상품 개발
- 정부 활용 : 지역 별 자원 배분 우선 순위 결정 근거로 활용

◆ 국민인식 파악: 기후변화 SNS 감성분류기를 이용하여 국민 감성을 실시간 파악

- 기후변화 SNS 파악 → 수집 → 전처리 → 감성분석 일괄 처리에 필요한 인프라 구축
  - 파악 및 수집 : 기후변화 사전
  - 전처리 : 기후변화 SNS자료 전처리 경험이 축적되어 기간 축소 가능
  - 감성분석 : 감성분류기
- 감성분류 기후변화 SNS DB + 단어 embedding : 학습 데이터를 확대하여 정확도 제고 가능
  - Weakly supervised learning : SNS 를 수집하고 기존 DB 와 유사한 SNS에 기존 DB 감성 부여
  - Augmented Sampling : 단어 embedding에 미세한 변화를 가하여 유사 Data 생성

감사합니다

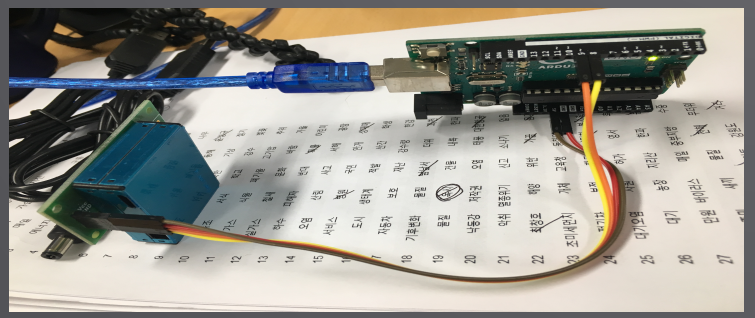
(3) 스마트센서 활용 Data 수집

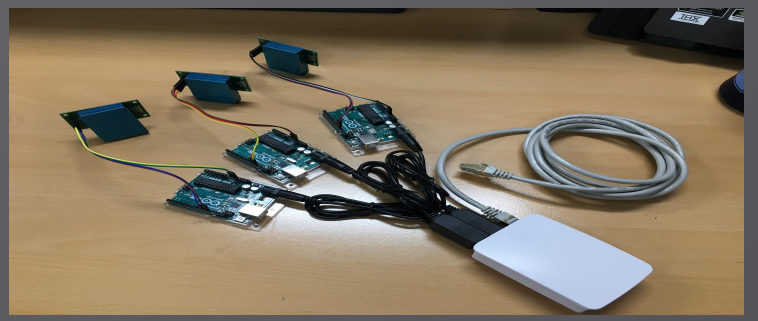
- 의의: 데이터 음영지역 및 소규모 연구대상 지역 데이터 수집 수단으로서의 가능성 모색
  - 미세먼지 측정기 미설치 지역 데이터 수집 : 데이터의 질을 평가하고 측정소 데이터와 비교
  - 저비용으로 자체 데이터 수집 및 수집-저장체계 마련 : 개별 연구/시민 참여형 연구 활용 가능성 점검

- 아두이노, 라스베리파이와 미세먼지 센서(PMS7003)를 활용한 스마트 센서를 구축하고 데이터를 수집 중
  - 스마트 센서 1기 : 센서 3기, 데이터 컨트롤러 3기(아두이노), 데이터 측정소 1기(라즈베리파이)
    - 미세먼지 센서[측정] - 아두이노 [데이터 전달] – 라즈베리파이 [임시저장]
  - 설치장소 : KEI(10층), 세종 새롬동(28층)
  - 수집된 Data 를 실시간으로 서버에 저장 : 10건 수집 Data 중 최대, 최소값 제외하고 평균값을 기록

- 향후 계획 : 실시간으로 수집 데이터 파일을 생성하여 플랫폼에 수록 및 수집자료의 신뢰성 점검
  - 신뢰성 점검 : 3개 센서의 측정치 상호 비교 등 관측 데이터의 정보를 활용하고 전문가 상담을 진행

스마트 센서 시스템 Architecture





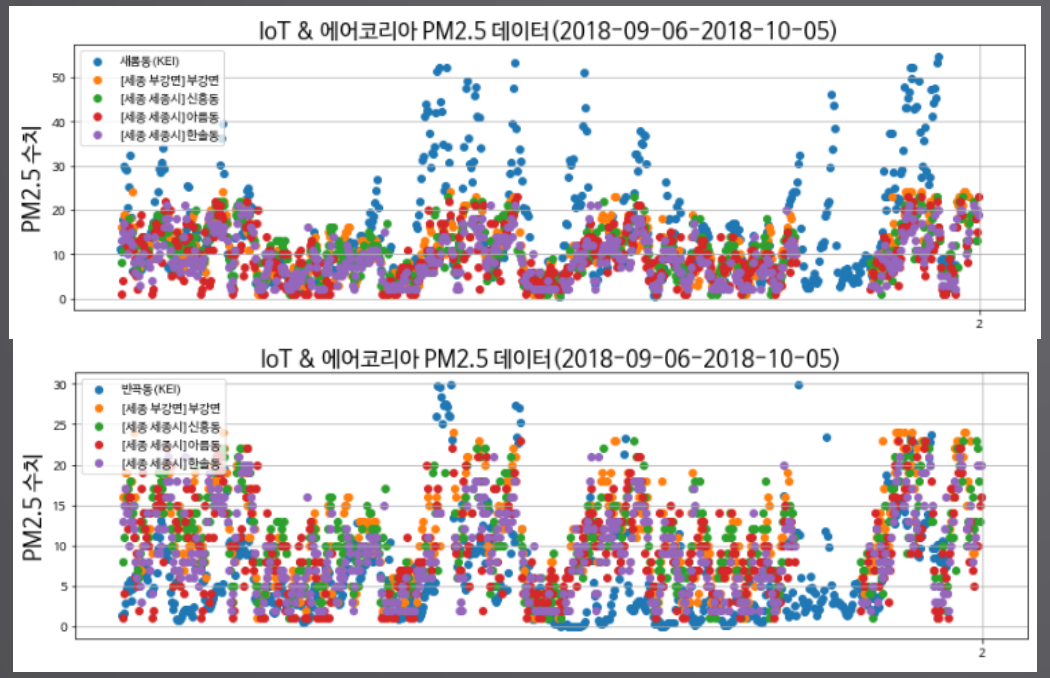
스마트센서 활용 데이터 수집 현황

◆ 월 단위 데이터 분석(에어코리아 세종시 4개 측정소)

- 설치 위치 차이 : 지도 참고
- 높이 차이 : 새롬동(아파트 32층 높이), 반곡동(오피스 빌딩 11층 높이)

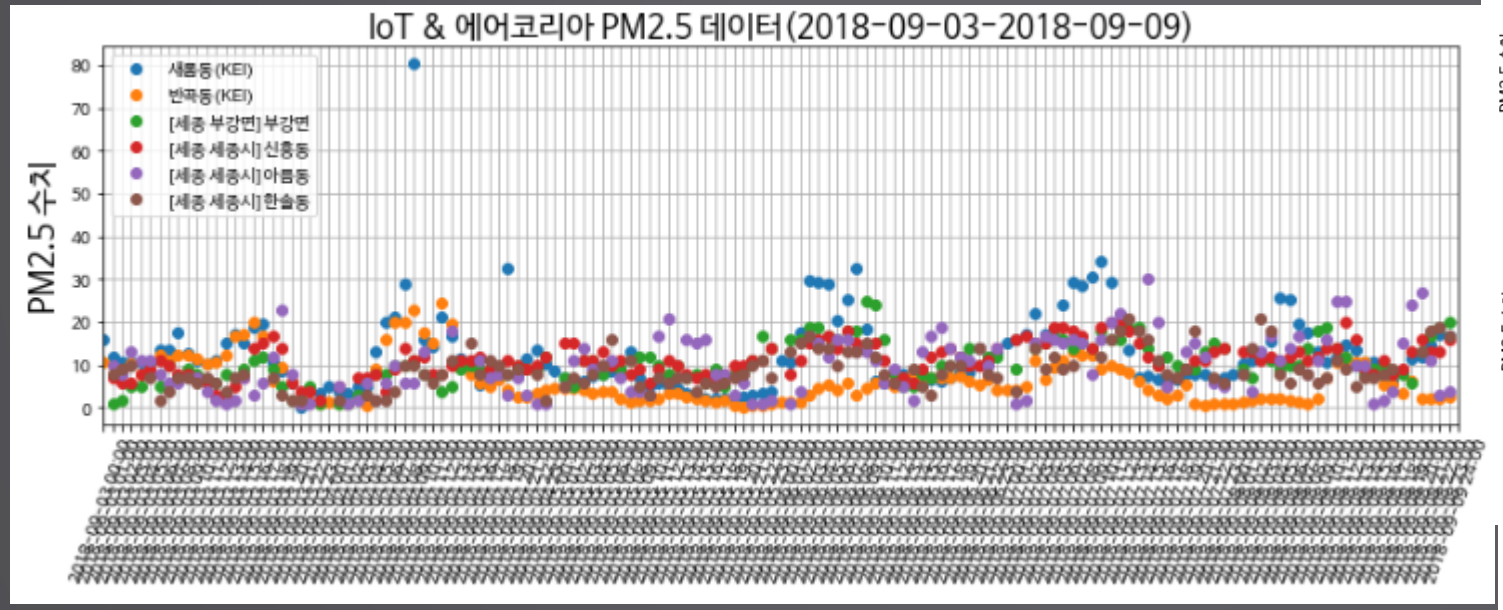
※ 지상 10m이내, 흡입구 이격 1.5m / 대기오염측정망 설치 운영지침(2016, 환경부)

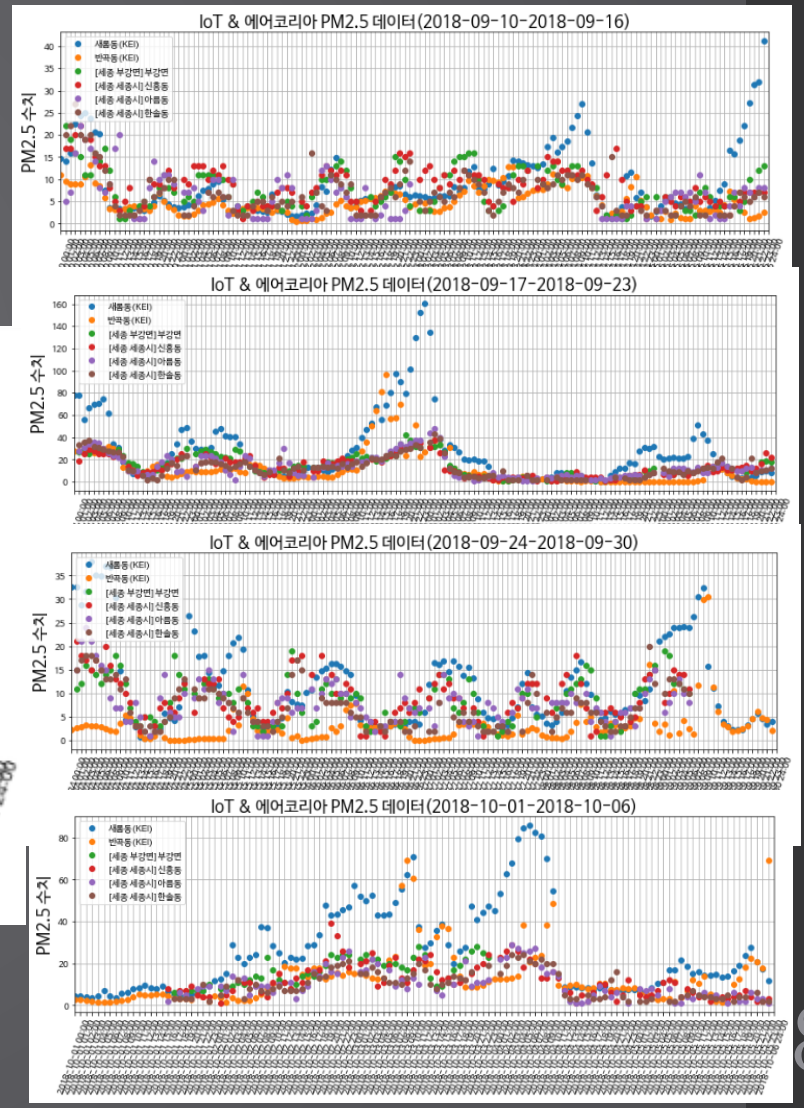
※ 외기 영향 최소화를 위해 창문틀 설치



스마트센서 활용 데이터 수집 현황

- 주 단위 데이터 분석
  - 스마트센서 : 실시간 해상도 -> 시간 단위로 해상도 조정
  - 에어코리아 : 시간 단위 해상도 그대로 사용

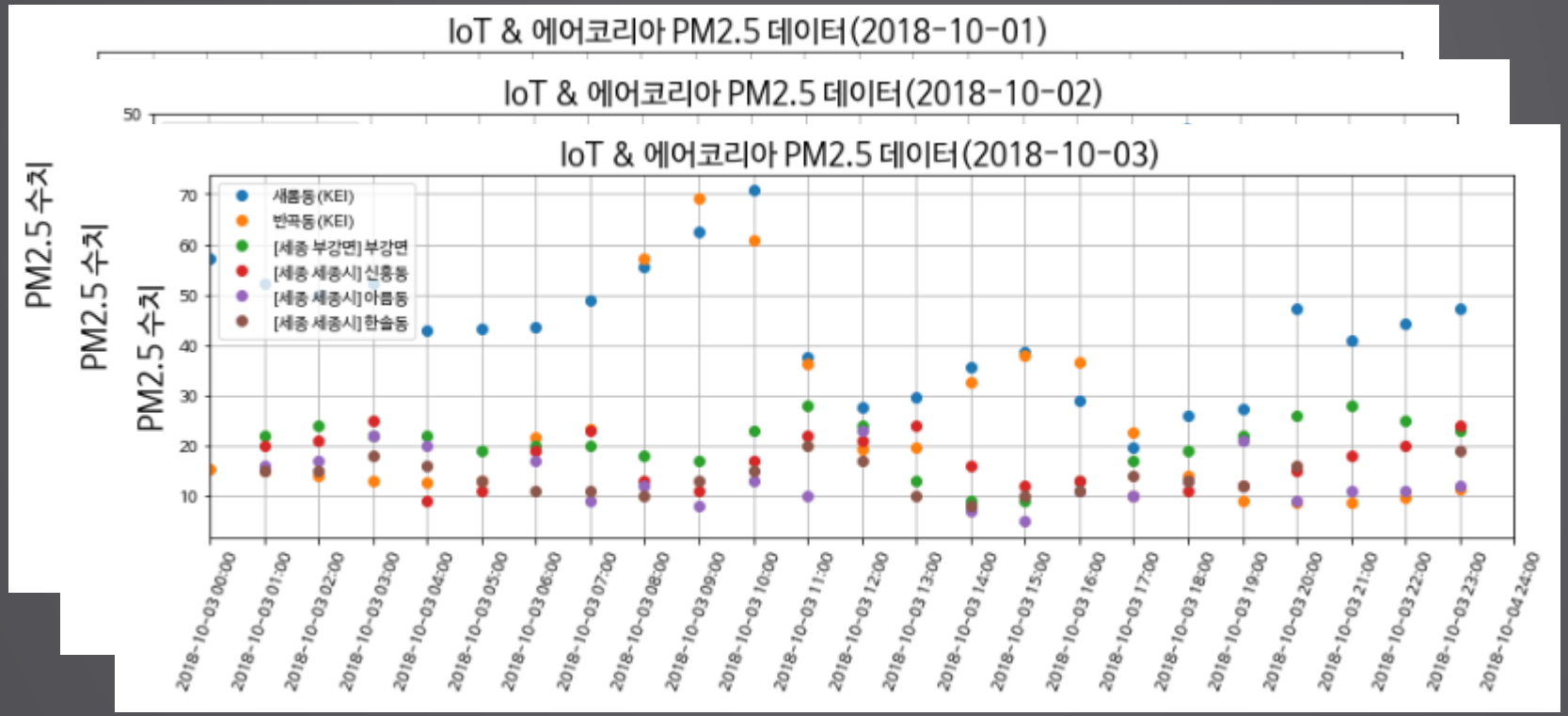




스마트센서 활용 데이터 수집 현황 : 일별 자료

◆ 일 단위 데이터 분석

– 시간 해상도 변경에 따른 보정값 탐색이 필요



외부 영향 최소화 및 고품질 외기 유지를 위한 방안 필요

- 외부 요소(온도, 습도, 바람) 영향 최소화를 위한 하우징은 반드시 필요
- 외기 흡입구 길이 보장(30cm 이상) 및 습기 제거 필요(건조 등)
- 외부 바람 제어 필요(풍속), 따라서 흡입구 방향이 아래서 위로 들어오도록 설치해야 함