2018-03-23

환경 빅데이터 분석 및 서비스 개발

착수자문회의(2018.3.23)

한국 환경정책·평가연구원

강 성 원

1. 연구 일반

2. 연구 목적

3. 연구 내용 및 방법론

4. 사업 관리

5. 기대 효과

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 (한국개발연구원) 오세영 박사 (한국행정연구원)
자문일정	착수자문회의: 2018년 3월 중간자문회의: 2018년 7월 최종자문회의: 2018년 10월	

목적: 빅데이터 연구방법론 환경연구 적용 가능성 모색

- **세부목적 1**: 환경 빅데이터 연구 수행
  - 주제선정 → 데이터 수집 및 가공 → 데이터 분석 → 결과 전달 전 과정 빅데이터 분석 기법 도입
    - [주제 선정] 알려지지 않은 규칙성을 발견하여 연구주제 및 정책과제를 발굴
    - [데이터 수집 및 가공] 연구 주제 분석 관련 대용량 데이터를 수집 및 축적하는 알고리듬 개발
    - [데이터 분석] 단기 예측의 정확도를 제고하고 개인별, 지역별 특성을 반영한 예측치를 도출
    - [결과 전달] 연구결과를 새로운 데이터를 이용하여 주기적으로 update 하여 발신
- **세부목적 2**: 환경 빅데이터 연구 인프라 구축
  - 환경 빅데이터 연구 결과 축적된 자료 및 알고리듬 공유
  - 원내외 환경자료 수집・추출 사례 축적 및 공개
  - 환경연구에 특화된 빅데이터 연구 플랫폼 구축
- **세부목적 3**: 원내외 빅데이터 서비스 개발
  - 환경 빅데이터 연구성과를 활용하여 연구정보 서비스 및 공공 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**

2017년 성과: 예측 및 연구주제 파악 가능성 확인

- 수치 데이터 예측 알고리듬 3개, 텍스트 데이터 연구동향 분석 알고리듬 3개, 환경 데이터 수집 알고리듬 3개 구축
  - 예측 알고리듬: 기존 연구방법론 대비 예측오차 개선
    - LSTM, kNN 공간순환신경망 : 측정소-시간 미세먼지 오염도 예측오차 10% 개선
    - 심층신경망 : 시군구-월 장감염 발생빈도 예측오차 25% 개선
    - 랜덤포레스트/Boosting: 시군구-월 미세먼지 오염도 예측오차 37%/46% 개선
  - 연구동향 분석 알고리듬: 환경뉴스 동향과 KEI 연구보고서 동향 추이 비교 → 새로운 연구주제 도출
    - 새로운 토픽 : 유전자 변형-소음, 보건-데이터 연구
    - 기존 토픽 연구 방향 : 기후변화 총론 연구 → 태풍, 한파. 대설 등 세부 현상 연구
  - 환경 데이터 수집 알고리듬: 공공데이터 포털, AirKorea, 기상자료개방포털 3개 홈페이지 자료수집 자동화

- 연구자료 및 알고리듬 인터넷 공개, 학술대회 발표, 논문 개제를 통해 결과 공유
  - 연구자료 및 알고리듬 : 홈페이지(https://keibigdata.github.io/project.html), Github(https://github.com/keibigdata/)
  - [이동현 교수, 강선아 연구원] SSCI 급 국제학술지 논문 개제: .Lee, D., Kang, S. and Shin, J. (2017), Deep Learning Techniques to Forecast Environmental Consumption Level, sustainability, 9(10). (SSCI)
  - [김도연 연구원] 대한산업경영학회 International Conference on industrial Convergence Best Paper Award. 'A study on Recognition of Climate Change by using Word2Vec'(Do-Yeon Kim, Sung-Won Kang)

2. 연구 목적

환경 빅데이터 플랫폼: 대용량 자료 활용 연구 플랫폼 설계

- ‘수요자 맞춤 지원행정’ 인프라 역할을 수행할 수 있는 환경 빅데이터 플랫폼 설계
  - 환경 빅데이터 플랫폼: 환경 데이터 활용 연구 및 환경 빅데이터 분석기법 개발 연구를 연구자가 수행할 수 있는 연구 환경
- 자료 수집, 축적: 환경 데이터 안내지도(Open Data Map)를 구축하고 기관 자체 자료를 결합하여 환경 데이터 안내지도를 보완
  - 환경 데이터 안내지도(Open Data Map) 구축 : 데이터의 목록과 Link를 제공
  - 기관 자체 자료를 사용하여 환경 데이터 안내지도를 보완: 자체수집, 기존 DB
- 자료분석: 대용량 자료 분석 및 빅데이터 분석 알고리듬 개발 환경
  - 기존 알고리듬 사용자: 사용자 편이성이 높은 Web기반 환경 제공
  - 알고리듬 개발자: 개발자의 자유도가 높은 CLI 기반 환경 제공
- 시험운영: 과제 참여자들이 설계된 플랫폼을 1년간 시험 운영하여 플랫폼의 실용성을 점검

환경 빅데이터 연구: 비정형 대용량 자료분석

- 빅데이터 연구기법이 비교우위를 나타내는 비정형 대용량 자료 분석을 추진하고 연구영역을 확대
  - 분석 자료를 전산화된 자료에서 비정형 자료로 확대: 화상자료, SNS자료
  - 대용량 자료의 장점을 활용할 수 있도록 예측 대상의 해상도를 제고: 시군구, 측정소→ 개인
  - 매체 중심 연구에서 수용체 중심 연구로 분야를 확장: 개인 건강, 유동인구

- 비정형 대용량 자료 분석 : 화상(Image) 분석, 건강보험 자료, SNS 자료 분석
  - 화상 분석: 미세먼지 오염도를 이미지로 전환하여 컨벌루션 신경망 모형(CNN)을 적용
  - 건강보험 자료 분석: 건강보험 코호트 자료를 이용하여 개인별 환경성질환 분석
  - SNS 자료 분석 : 환경이슈와 관련된 SNS 자료를 이용하여 환경이슈에 대한 감성을 분석

- 연구영역 확대: 수질오염 예측 및 환경위험에 대한 수용체 반응 분석
  - 2017년 대기오염 중심 매체 기반 연구→ 수질오염 연구 및 수용체 반응 연구로 확장
  - 수질오염 : 한강수계 측정소별 주간 수질오염 오염도 예측 알고리듬 개발
  - 수용체 반응 : 미세먼지 오염도가 서울시 유동인구에 미친 영향 분석 알고리듬 개발

환경 빅데이터 서비스: 연구동향 서비스 원내 공개 추진

- 2017년 연구성과 중 '텍스트마이닝을 이용한 KEI 연구동향 분석'에서 개발한 연구동향 분석 알고리듬 원내 공개
- LDA 기반 토픽 클러스터링 : 연구보고서를 유관성이 높은 토픽으로 분류하고 연간 토픽 구성을 파악하여 개괄적 연구 동향을 파악
  - KEI 보고서 및 NAVER News 제목
- 네트워크 분석: KEI 보고서의 연관어를 파악하여 연관 빈도가 높은 단어들의 네트워크를 도출
  - 네트워크 구성의 시간 별 추이를 파악하여 연구 동향을 파악
  - KEI 보고서 및 NAVER News 제목

3. 연구 내용 및 방법론(1)

환경 빅데이터 연구 인프라 구축

대용량 자료 수집, 저장, 분석이 가능한 연구환경 설계

- 환경 빅데이터 플랫폼: 대용량 자료 수집, 저장, 분석 수행 연구환경을 제공
  - 개인 PC에서 처리할 수 없는 자료 분석이 필요한 연구를 One Stop으로 수행
    - 개인 PC에서 처리할 수 없는 자료 분석: 고성능 처리, 대용량 데이터, 연속 작업, 주기적 작업

- 연구자 사용 방식을 '자료 이용', '알고리듬 이용', '알고리듬 개발' 3개 방식으로 구분하고 각 방식에 적합한 기능 구비
  - 자료 이용: 대용량 자료 수집의 Gate
    - Open Data Map : 공개된 자료를 찾아 갈 수 있는 Link를 제공
    - 자체 자료 : 사용빈도가 높은 공개된 자료, KEI 연구성과, KEI 자체수집 자료 직접 사용
  - 알고리듬 이용: 기존 연구결과를 재생하거나 이미 개발한 알고리듬을 활용
    - 기존 연구 성과를 공개: 자료 및 알고리듬을 필요에 따라 수집하여 활용할 수 있도록 제공
    - 대용량 자료 분석 기능 부여: 개발된 알고리듬에 새로운 데이터를 적용할 수 있는 환경 제공
  - 알고리듬 개발 : 대용량 자료 처리에 필요한 새로운 알고리듬 개발
    - 알고리듬 개발 및 대용량 자료 분석 기능 부여 : 새로운 알고리듬을 개발하여 성능을 점검할 수 있는 환경 제공

- 자체 서버에 설치하고 과제참여자에게 개방하여 시범운영(2018)
  - 자료 이용 : Open Data Map 과 자체 DB
  - 알고리듬 이용 및 분석: 우분투(Ubuntu), 아나콘다(Anaconda) 등 소프트웨어 업데이트
  - 인프라: 서버 메모리 증설
    - 現 192GB에서 32GB * 18개로 증설 가능 : 예산 범위 내 추가 증설 예정

자료 이용 : Open Data Map을 구축하고 자체 DB로 보완

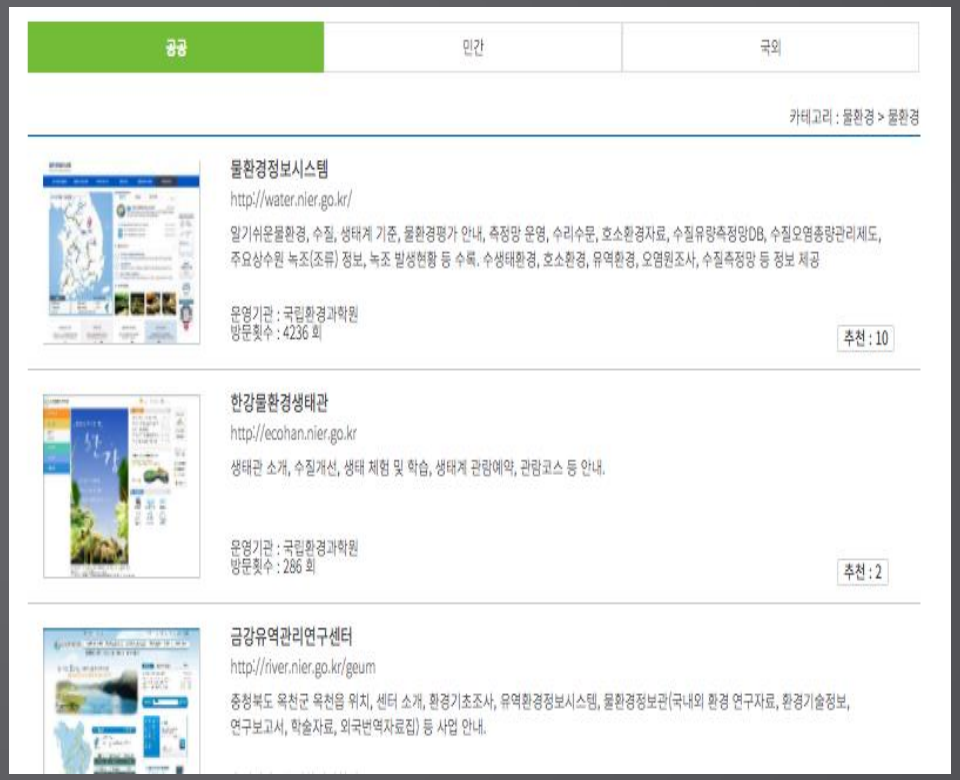
- Open Data Map 구축: KEI 발간 보고서 인용 온라인 문헌 자료를 기반으로 확장
  - 온라인 문헌 자료 추출 → 제목 등 관련 정보 부가 → 분류 → keyword 등 탐색정보 부여
  - 전자도서관 자료(2018.03.01)에서 URL 및 메타 정보(<title> 등) 추출
  - 다양한 분류 기준을 적용하여 분류하고, 원본 보고서의 keyword를 부여

- 자체 DB: 외부 데이터 수집 DB 와 IoT를 활용한 자체 수집 DB로 Data Map 보완
  - 수요가 높은 데이터 수집 과정을 자동화하여 연구자의 자료 수집 부담 경감
    - Open Data Map 사용 log 분석 및 데이터 수요 설문조사를 이용하여 수요 파악
      - KEI 내부 및 외부전문가 DB 수록 전문가 대상: 연 70건 이상 조사결과 확보
  - IoT Data DB 구축: 센서 기반 데이터 수집 방법을 보완
    - 2018년 하반기부터 세종시 2개 지역 이상 미세먼지 오염도 수집
      - 1개소 당 센서(PMS 7003) 3기 * 2개소
  - 원내 기 구축 DB 및 유관 과제와 협조 추진
    - 환경가치 종합정보 시스템 (안소은), 국토환경지리정보 활용성 제고방안(명수정)

- 자료 이용 방식 : 대용량 데이터 파일 다운로드 (Web, DB) 방식 우선 제공
  - 데이터 추출 자동화 UI : 2019년 이후 검토 / 구축

Open Data Map의 예: 환경정보네트워크(환경부)

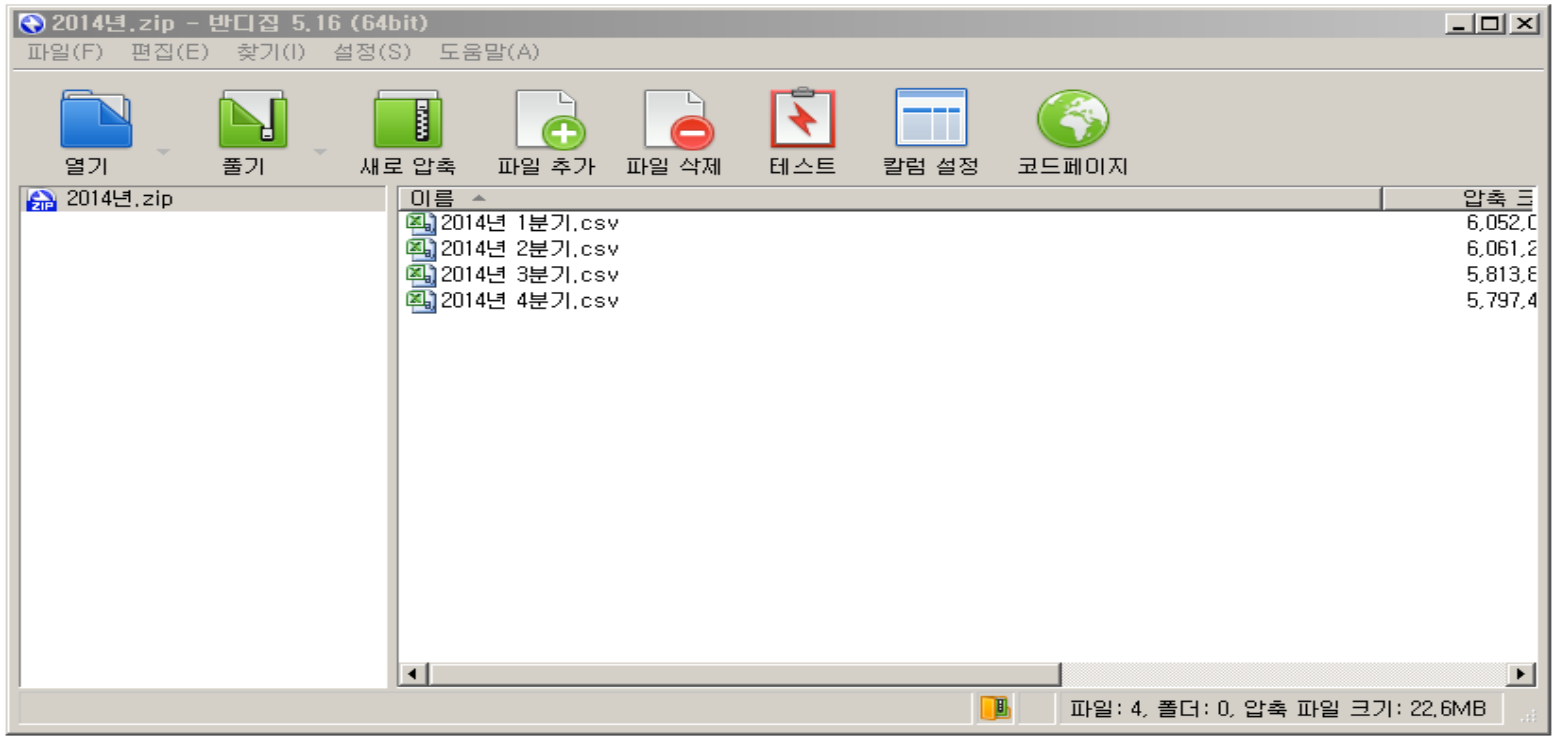
http://www.me.go.kr/home/etips/etipsMain.do



대용량 데이터 파일 다운로드

Index of /FileDB_에어코리아/최종확정자료/

../		
2014년.zip	18-Mar-2018 17:29	23M
2015년.zip	18-Mar-2018 17:29	24M
2016년.zip	18-Mar-2018 17:28	24M
2017년.zip	18-Mar-2018 17:27	39M

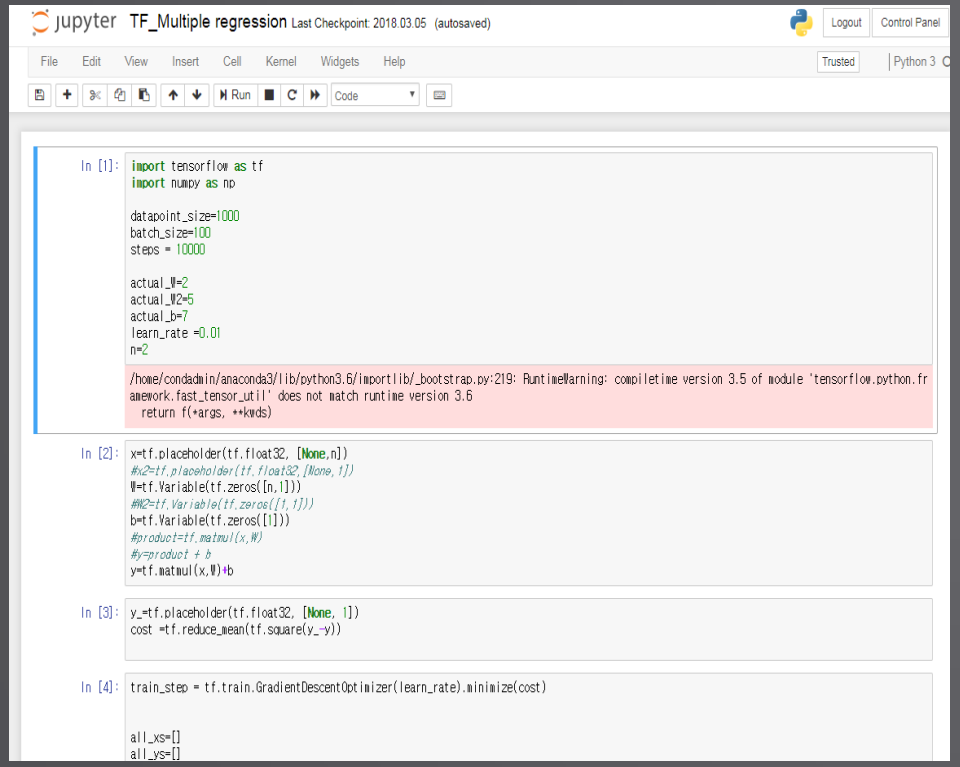


분석 플랫폼: 기존 알고리듬 공유, 활용 및 신규 알고리듬 개발환경

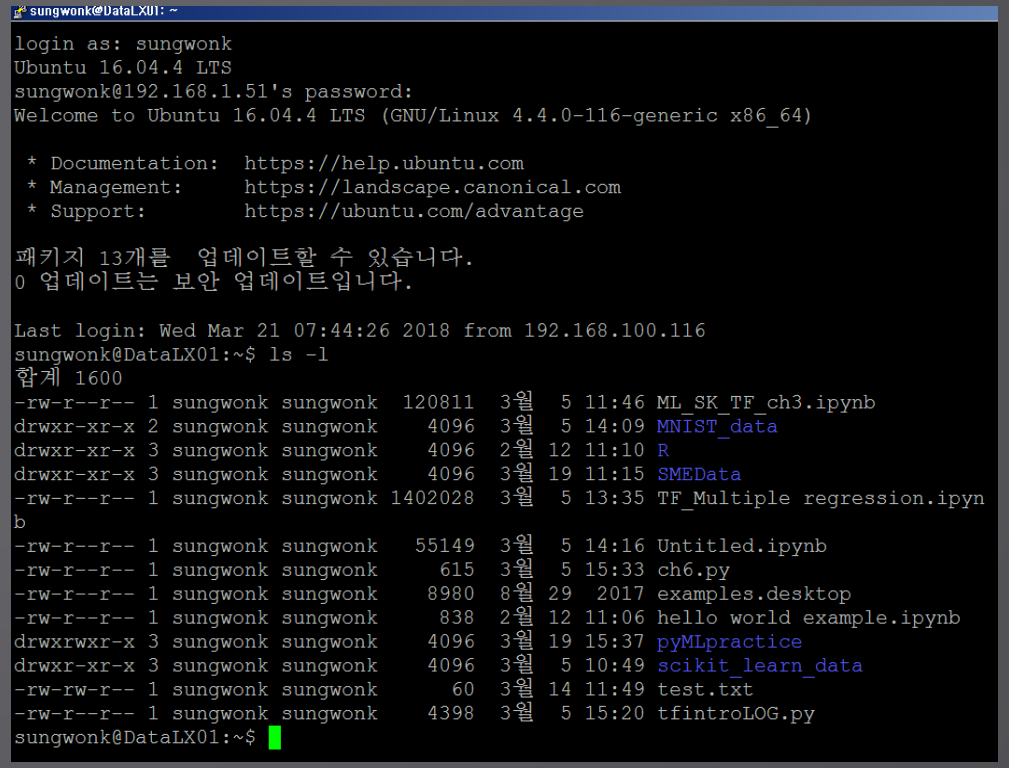
- 분석 플랫폼 : 알고리듬 이용 및 알고리듬 개발에 필요한 기능 제공
  - 알고리듬 이용: 기 개발된 알고리듬을 새로운 자료에 적용하는 연구
  - 알고리듬 개발 : 새로운 알고리듬을 개발하여 활용가능 여부를 점검하는 연구
- 알고리듬 이용: 연구자 선택 알고리듬을 대용량 자료에 적용할 수 있는 기능 부여
  - '연구자 선택 알고리듬' : 공개 알고리듬 및 연구자 기 개발 알고리듬
  - 기존 알고리듬을 새로운 대용량 자료에 적용하는 연구 및 신규 알고리듬의 대용량 자료 분석 기능을 점검하는 연구에 사용
  - 기존 '환경 빅데이터 분석 및 서비스개발' 연구 개발 알고리듬 및 자료 공유
- 알고리듬 개발: 기존 한계 극복을 시도하는 새로운 대용량 자료 분석 알고리듬 개발
  - 알고리듬 개발 및 대용량 자료 대상 실험을 반복할 수 있는 환경 제공
- 사용 목적에 적합한 사용자 환경(User Interface) 제공
  - Spark, Hadoop (대용량 자료 운용) + Python, R (프로그래밍 언어) 사용 가능 환경 제공
  - 알고리듬 이용: 사용이 상대적으로 용이한 Jupyter Notebook, RStudio 웹IDE 제공
  - 알고리듬 개발: 개발자의 재량(discretion)이 폭넓게 허용되는 CLI(Command Line Interface) 제공

분석 플랫폼

웹 IDE



Command Line Interface



3. 연구 내용 및 방법론(2)

환경 빅데이터 분석

1. 컨벌루션 신경망(CNN)을 통한 미세먼지 예측

2. 데이터 기반 한강 수질 예측

3. 딥러닝 이용 국내 노인인구 호흡기 질환 사망 위험 추정

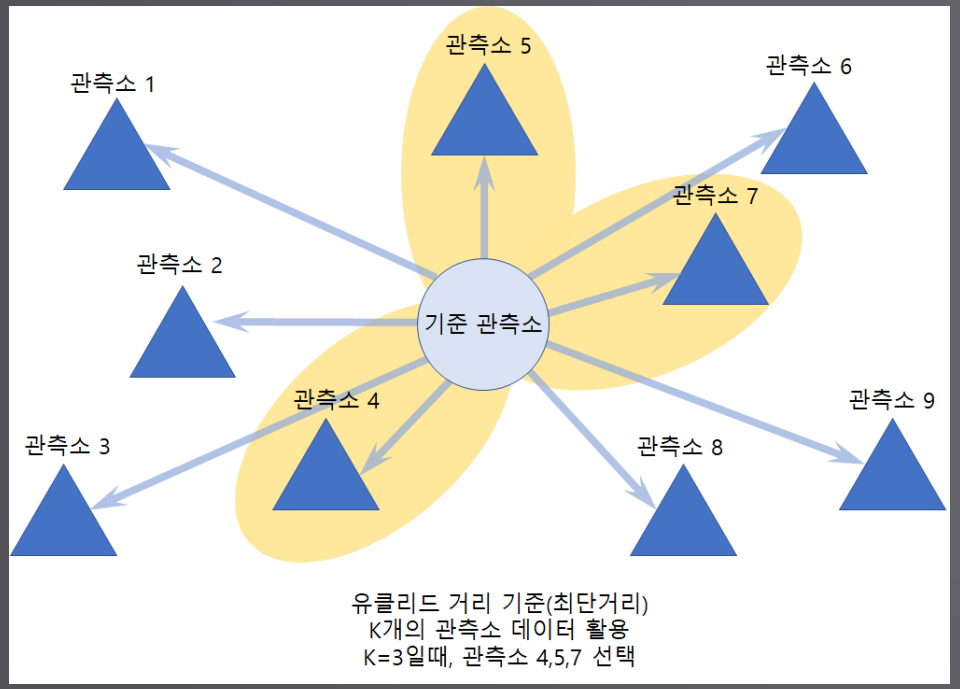
4. 딥러닝 기반 환경이슈 감성분석기 개발

5. 미세먼지 농도 및 예보가 서울 대중교통 이용에 미치는 영향

(1) 컨벌루션 신경망(CNN)을 통한 미세먼지 예측

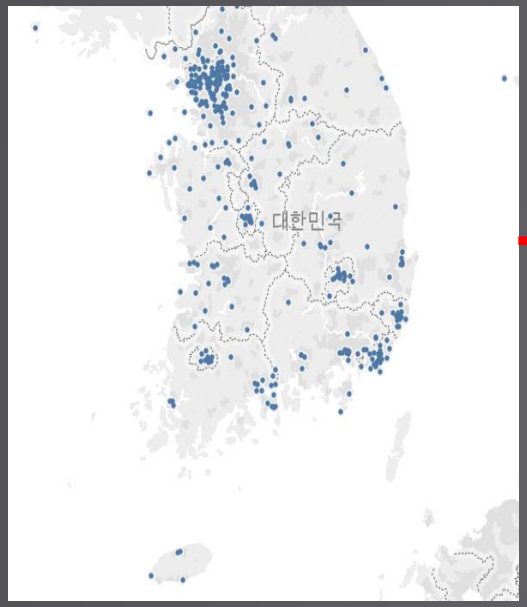
- 목적: 대기오염 오염도 추정 시 주변 지역 정보 반영을 최적화하는 메커니즘을 컨벌루션 신경망을 이용하여 학습
  - 2017년 '딥러닝을 활용한 환경 리스크 예측' 연구에서 주변지역 정보 반영 시 예측 정확도가 제고되는 현상을 확인
    - kNN공간 순환신경망: kNN 주변 지역 선택 알고리듬 + RNN > RNN
  - 주변 지역 정보 반영 방식을 임의적 방식에서 데이터 기반 방식으로 전환
    - kNN 방식의 임의성 극복
- 방법론: 측정소 별 미세먼지를 거리를 반영하여 격자형으로 보간하고 컨벌루션 신경망(CNN: Convolution Neural Network)을 적용하여 오염도 공간패턴 추정
  - 위도, 경도, 대기 및 기상자료, 미세먼지 오염도 → 미래 미세먼지 오염도 추정
  - 개별 자료에 서로 다른 label을 부여하여 모든 label을 예측하는 Many-to-Many 방식 사용
- 기대효과 : 단순 순환신경망보다 예측 오차 축소 기대
  - 주변 지역 정보 활용 범위 및 방식이 예측 오차를 최소화하는 방식으로 결정

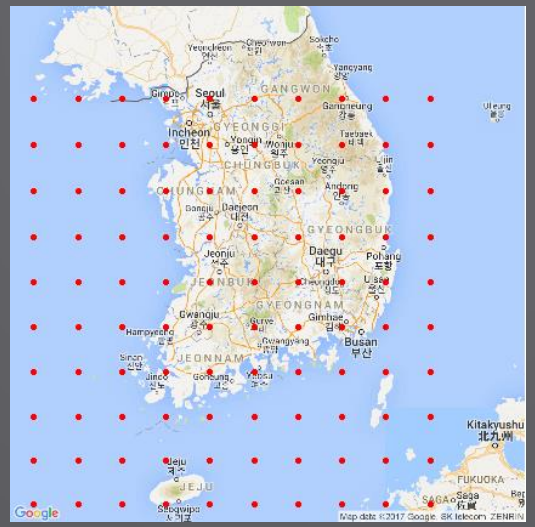
(2017) KNN 공간순환신경망 :k 개 인접지역 정보를 반영

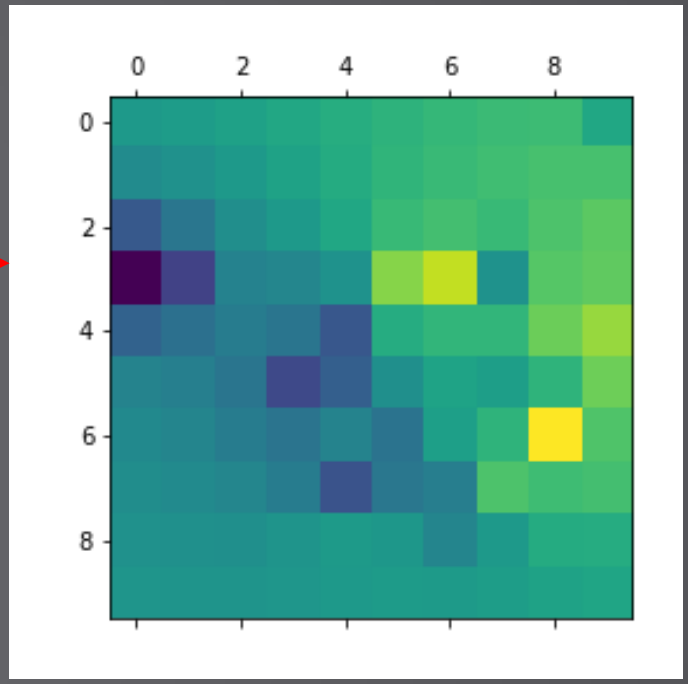


	RMSE
OLS	17.04
ARIMA	8.89
LSTM	8.19
KNN공간순환신경망(k=5)	**7.96**

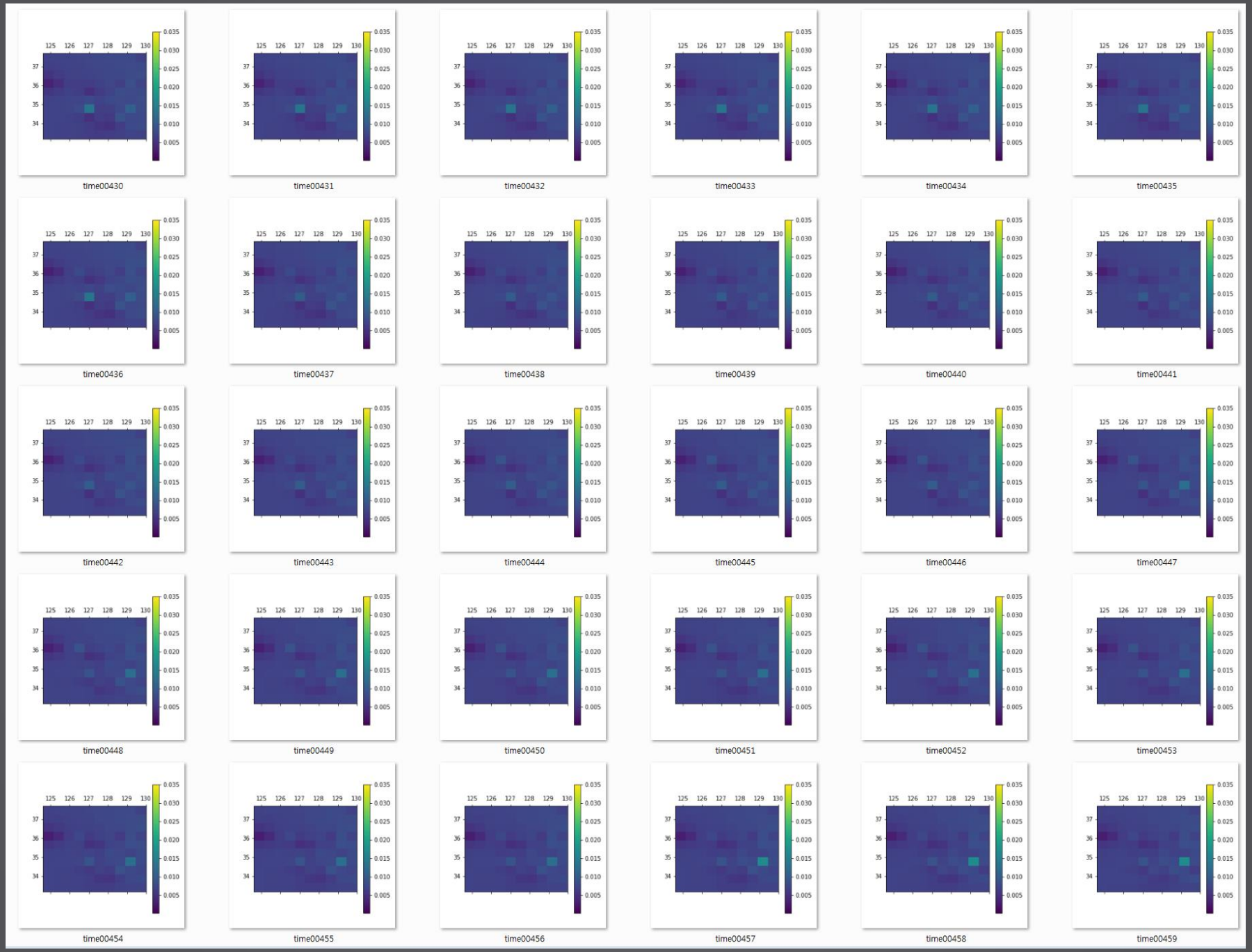
데이터 변환 : 측정소 자료를 격자형 자료로 보간



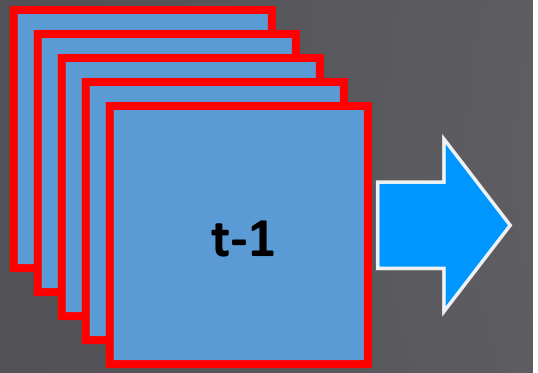




데이터 격자 보간
(IDW: Inverse Distance Weighted)

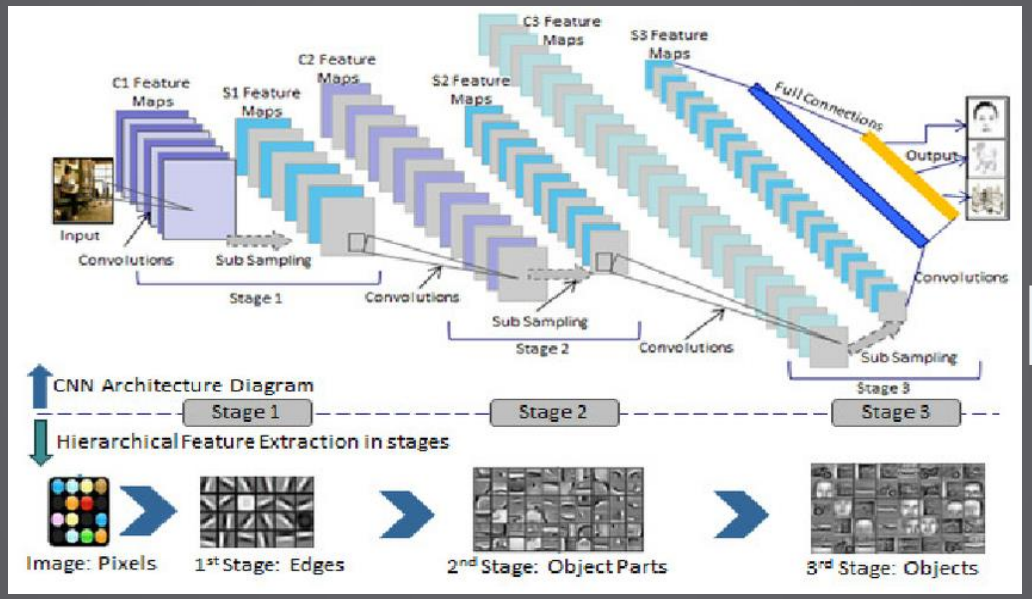

time00430
time00431
time00432
time00433
time00434
time00435
time00436
time00437
time00438
time00439
time00440
time00441
time00442
time00443
time00444
time00445
time00446
time00447
time00448
time00449
time00450
time00451
time00452
time00453
time00454
time00455
time00456
time00457
time00458
time00459

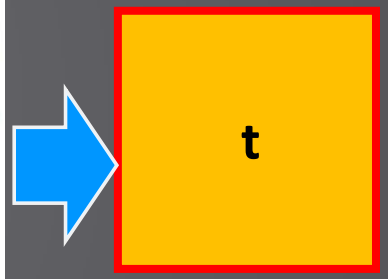
CNN (Convolution Neural Network) 을 활용한 예측



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

CNN





Y:
PM10

4차원 (위도, 경도, 변수, 시간)

(2) 데이터 기반 한강 수질 예측

- 목적: 인공지능 및 공간통계모형을 이용한 데이터 기반 수질 예측 알고리듬 개발
  - 공간해상도가 높은 자료를 요구하는 기존 수질모형과는 다른 시각에서 접근
    - '있는 데이터를 활용해서 얻을 수 있는 최선의 결과를 추구'
  - 특정 하천에 국한되지 않은 예측모형을 개발하여 적용

- 물환경정보시스템의 수질 일반측정망 자료와 기상자료를 활용하여 수질오염도 예측
  - 2013-17 수도권 수질 측정 지역 중 자료가 충분한 측정소를 선택하여 주간 용존산소량 예측
    - 350개 이상의 관측치가 존재하는 12개 측정소 대상으로 예측 진행
    - 과거 오염도 자료 및 기상자료[입력자료] → 용존산소량 예측치 [출력자료]
    - 주변 지역 오염도 변수는 설명 변수로 활용

- 방법론 : 기계학습, 시공간자료 분석 모형(통계학) 방법론을 병행
  - 기계학습 : ANN, RNN, GRU or LSTM[심층신경망] , Autoencoder [차원 축소]
    - kNN 을 통해 주변 지역 정보를 반영
  - 시공간자료분석 : 시간 및 공간 정보를 동시에 예측에 활용하는 기법

- 기대효과: 선형회귀모형, VARMA(vector ARMA) 모형보다 예측오차 축소 기대

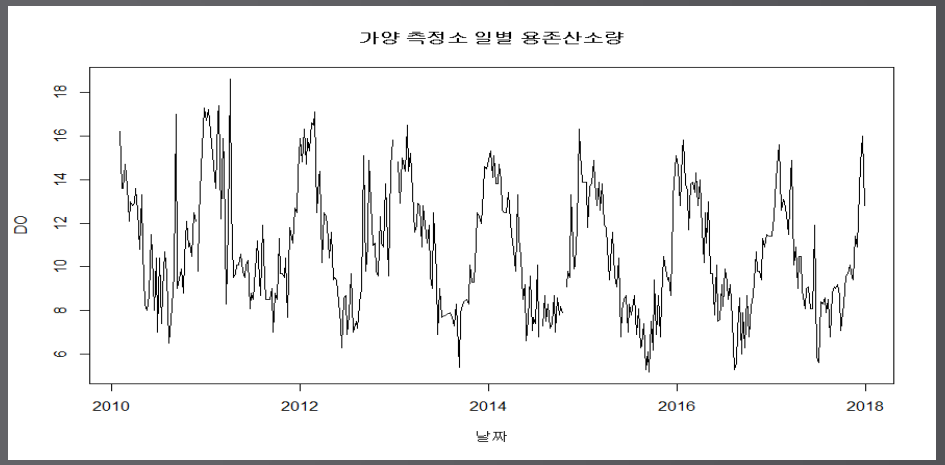
자료 구성

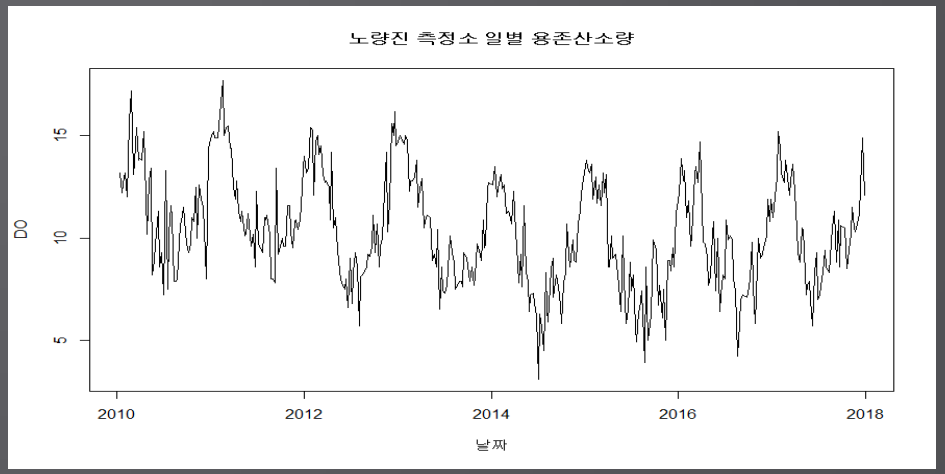
◆ 입력 자료 : 수질 일반측정망 자료 및 기상자료

- 수질 일반측정망: 수소이온농도(pH), 용존산소량(DO), 총인(TP), TOC, 수온, 전기전도도, 암모니아성질소(NH3-N), 질산성질소(NO3-N), 클로로필-a
- 기상 : 강우량, 습도
- 오염물질 배출량은 연간자료이기 때문에 주간 용존산소량 추정 정확도 제고 효과를 기대하기 어려움

◆ 추정 대상 : 용존산소량

용존 산소량(가양, 노량진 측정소)

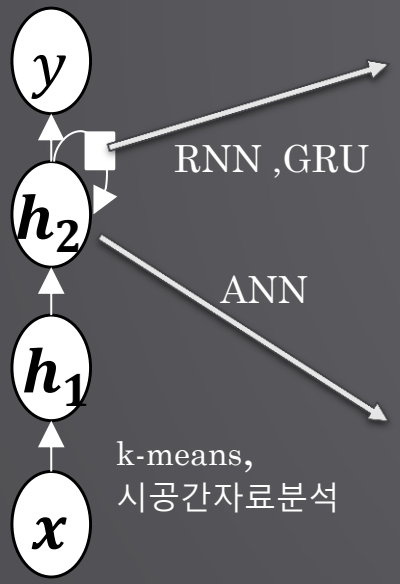


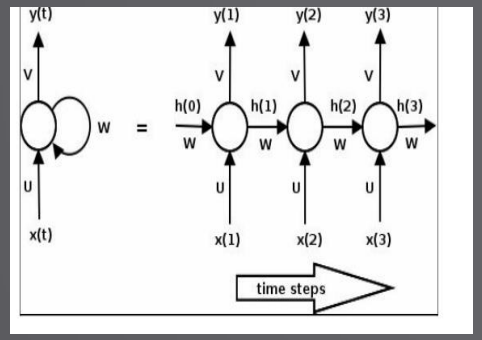


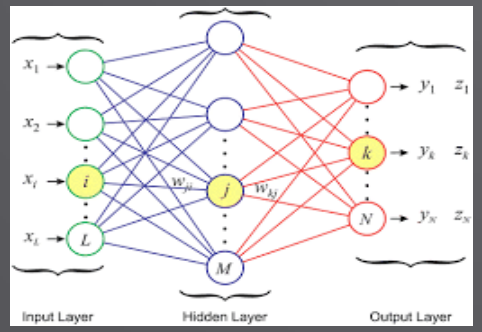
데이터 기반 한강수계 예측

ANN, RNN 기반 알고리듬

◆ 시계열 자료: Lag 반영 ANN/ RNN 기반 알고리듬 적용

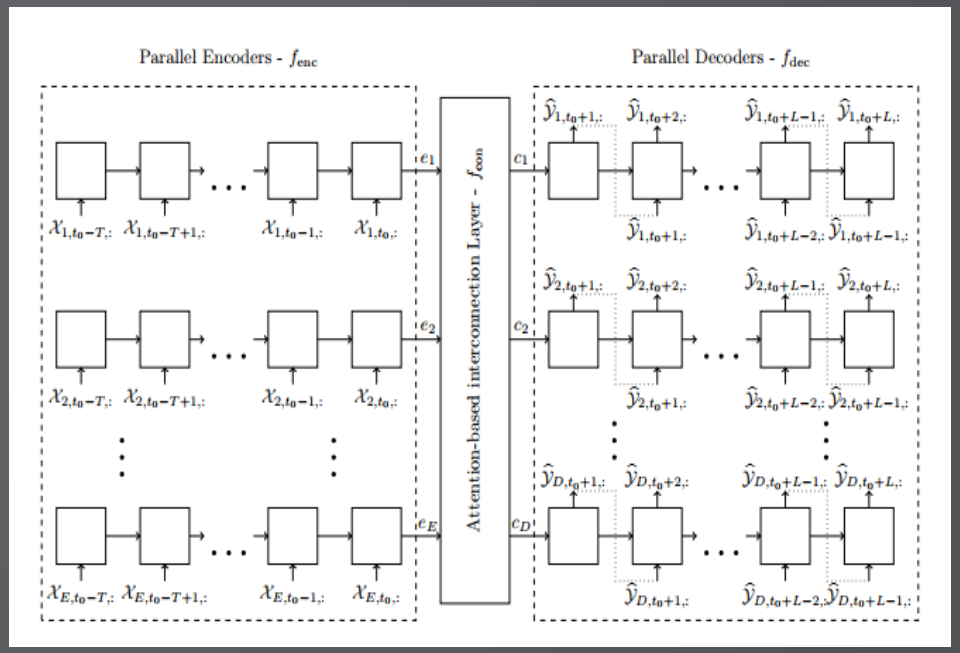






Auto encoder 적용

◆ 하천의 특성을 반영하는 1차원 변수로 2차원 자료를 축소



(3) 딥러닝 이용 국내 노인인구 호흡기 질환 사망 위험 추정

- 목적: 딥러닝을 활용한 예측 의학 성과를 환경성 질환 분석에 적용
  - 예) 머신러닝 알고리듬으로 파악한 심혈관 발병 인자를 이용하여 추정한 발병 위험 추정치가 기존 학회 제공 발병 인자 이용 추정치보다 더 정확함을 확인(Stephen et al. 2017)
  - 실시간으로 갱신되는 데이터를 반영하여 결과를 update 할 수 있는 딥러닝의 장점 활용

- 내용: 만성폐쇄성 폐질환 사망 위험을 딥러닝을 이용하여 추정
  - 연구 대상 :65세 이상 만성폐쇄성폐질환(COPD) 환자
    - 2009년 현재 치료 중 환자 192,496명/ 2010년 전체 사망원인 중 7위에 해당
  - 자료 : 건강보험 맞춤형연구 DB , 2006-2015년 건강보험 코호트 DB version 2.0, 인구, 기후, 대기오염도 및 대기오염물질 배출량 자료를 연계
    - 맞춤형연구 DB: 만성폐쇄성 폐질환 질병에 영향을 끼치는 요인 분석
    - 건강보험 코호트 DB: 인구 특성 (성별, 연령) , 건강 관련 특성(병력, 식전혈당..), 진료기록
    - 기후: 기상청 제공 시군구별 기후 데이터
    - 환경자료: 대기오염물질 오염도, 대기오염물질 배출량

- 방법론: 딥러닝과 일반적인 호흡기 질환 사망위험 예측 모델링의 예측 정확도 비교
  - 머신러닝 방법론: Lag 변수를 변인(feature)으로 포함하는 ANN/시계열 분석이 가능한 RNN 적용 점검
  - 일반적으로 알려진 위험인자: 대한결핵 및 호흡기 학회/WHO 제공

머신러닝 의료분야 적용 성과 예: 심혈관 질환 위험 예측

머신러닝을 적용하여 Baseline 예측의 성과를 1.7~3.2%p 개선 (AUC 기준)

Table 4. Performance of the machine-learning (ML) algorithms predicting 10-year cardiovascular disease (CVD) risk derived from applying training algorithms on the validation cohort of 82,989 patients. Higher c-statistics results in better algorithm discrimination. The baseline (BL) ACC/AHA 10-year risk prediction algorithm is provided for comparative purposes.

Algorithms	AUC c-statistic	Standard Error*	95% Confidence Interval		Absolute Change from Baseline
			LCL	UCL	
BL: ACC/AHA	0.728	0.002	0.723	0.735	—
ML: Random Forest	0.745	0.003	0.739	0.750	+1.7%
ML: Logistic Regression	0.760	0.003	0.755	0.766	+3.2%
ML: Gradient Boosting Machines	0.761	0.002	0.755	0.766	+3.3%
ML: Neural Networks	0.764	0.002	0.759	0.769	+3.6%

일반적으로 알려진 만성폐쇄성 폐질환 위험인자

	위험인자
대한결핵 및 호흡기학회	- 흡연 - 숙주인자: 유전자, 노령, 성별, 폐성장, 기도과민반응 - 외부인자: 외부 유해물질(흡연, 직업성 분진과 화학물질, 실내외 대기오염), 사회 경제적 수준, 만성기관지염, 호흡기 감염
WHO	- 흡연 - 실내공기오염, 실외대기오염 - 직업성 분진 및 화학물질 - 어린 시절 잦은 호흡기 감염

분석 자료 구축

개인신상정보

Medical history

개인코드	연도	성별	연령	소득	장애유무	입원기록	거주지	기후, 대기 오염 및 배출량 데이터		사망
1111	2006									1
1111	2007									0
…										0
1111	2015									1

(4) 딥러닝 기반 환경이슈 감성분석기 개발

- 목적: 소셜 미디어에 표출되는 환경이슈에 대한 국민인식을 분석하여 환경정책 수립 기초자료로 제공
  - 기존 의견수렴 방식을 보완하는 데이터 기반 의견수렴 방식 개발
  - 소셜 미디어를 통해 표출되는 환경이슈 관련 국민인식을 실시간으로 파악

- 연구내용: SNS 및 주요 포털 댓글 환경관련 데이터 수집 및 감성분석 알고리듬 개발
  - 수집 : 온라인 환경관련 데이터를 파악하고 수집하는 과정을 자동화
    - 수집대상 : Twitter, Facebook, Instagram, NAVER Café 댓글, Daum Café 댓글
  - 감성분석 : 텍스트 데이터의 감성을 6개 감성으로 분류하는 감성분석 알고리듬을 개발하여 적용
    - 6개 감성 Category : 긍정, 중립, 두려움, 슬픔, 분노,객관 (Robert Plutchik의 Wheel of Emotions 활용)

- 방법론: 환경사전 구축, 감성 분류 학습 데이터 구축, 감성분석 알고리듬 개발
  - 수집: 환경사전을 구축하여 환경관련 데이터를 파악하고 수집하는 과정을 자동화
    - 환경사전 : 2017년 '텍스트마이닝 이용 KEI 연구동향분석' 연구성과 활용
  - 감성분류 학습데이터 구축 : 전체 분석 자료의 2~5% (5만 건)의 감성을 직접 분류
  - 감성분석 알고리듬 : 단어 간 패턴(CNN) 및 문장 내 전후관계(RNN)를 모두 반영하는 알고리듬 구축

환경사전 구축: 환경 관련 텍스트 데이터 파악 기준

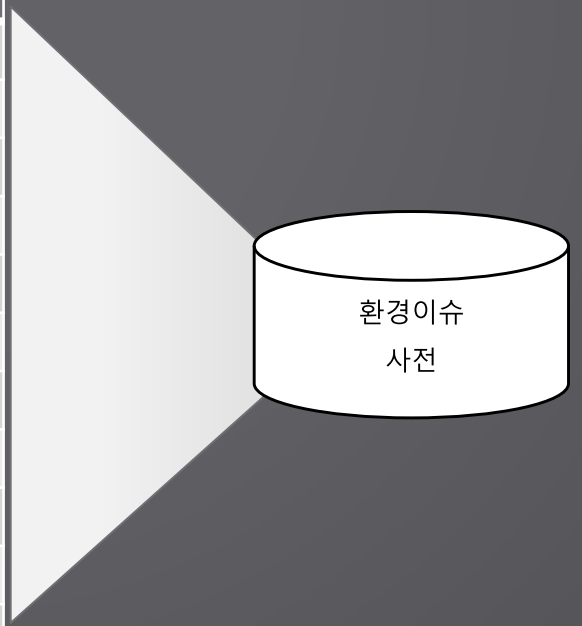
◆ 불필요한 데이터 수집을 제거하기 위해 환경이슈 주제에 적합한 키워드 사전구축

– 환경관련 생산문서에 워드 임베딩 방법(LDA, Word2Vec)을 적용: 키워드 후보군 추출

• 2017년 '텍스트마이닝 이용 KEI 연구동향분석' 연구 결과 활용 (KEI 보고서, NAVER 환경뉴스)

– 전문가 집단의 의견을 반영하여 키워드 후보군에서 키워드를 선정하고 이슈별로 구분

기후변화	에너지자원	폐기물	환경보건
미세먼지	온실가스	산업폐기물	환경성질환
온난화	신재생에너지	생활폐기물	환경성질병
이상기온	친환경에너지	폐수	유전자변형
폭염	청정에너지	하수	유전자조작
한파	전력	소각장	화학물질
가뭄	천연가스	폐기물처리장	아토피
홍수	풍력	하수처리장	석면피해
태풍	수력	쓰레기	가습기살균제
폭설	화력	악취	곰팡이
폭우	원자력	폐기물부담금	독감
.....			



감성분석 학습자료 구축 : 6개 감성으로 분류된 5만건

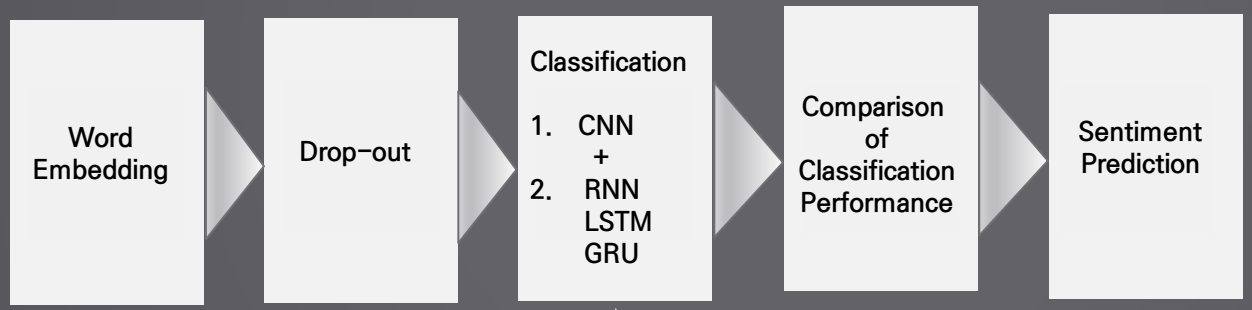
- 기존 범용 감성데이터는 환경 분야에서 쓰는 어휘 및 감정표현을 반영하기 어려워서 감성분류 작업이 필요
- 약 5만개 문장을 6개 감성으로 분류한 학습자료(Training Data Set)를 구축
  - 최소 3주 소요 예정

		감성태그
긍정		POS
부정	두려움	TER
	슬픔	SAD
	분노	RAG
중립		NEU
객관		OBJ

TER	이제 사계절 내도록 미세먼지와 함께 해야하다니..
SAD	봄의 불청객 ㅋㅋㅋ미세먼지인듯...
RAG	지긋지긋한 미세먼지
OBJ	대륙에서 홍수 피해로 구조중인 돼지.jpg
RAG	미세 먼지 참 0같네 진짜 —.—
TER	눈코입이 너무 따가운데 미세먼지때문이라고 믿고싶다..
POS	난 우리동네 공기가 이렇게 좋은거 첨봐.....미세먼지없음 진리구나 이제.....
NEU	댓글창 지진났네 쿵쾅쿵쾅
NEU	지각 할까봐 지진이 깨워 줬나 봄더 쿨럭
RAG	쓰나미 지진 와서 일본땅이바닷속으로 가라앉았으면 좋겠다
RAG	단군이 자리 잘못 잡아 나라 세웠네. 홍수와 가뭄을 겪는 땅에
NEU	가뭄난곳이 어디죠? 제 눈물로 단비를 뿌려주겟어여ㅠㅠㅠㅠ
RAG	비라도 많이와서 가뭄해갈에 도움됐으면 미세먼지는 쫌 꺼지고
TER	비좀 많이와라..가뭄 심각하다...
POS	미세먼지 없는 주말^^
TER	미세먼지에.. 가뭄에 점점 살기어려워지고 있네요..
TER	지구 온난화가 점점 문제를 일으키는구나 ㅠ
NEG	22도 오른것중에 12도정도는 짱개탓이지..거대암세포의 증식이 시작되고나서부터 지구온난화심해짐..
OBJ	철도공단, 25.8kV 친환경 개폐장치 전격 도입 추진한다!
OBJ	#가뭄 프리뷰 영상
POS	울 오빠들 폭우속 공연 최고였음!!!

감성분석 알고리듬: 단어 패턴 및 문장 내 전후관계 고려

- Word2Vec – CNN – RNN 이 결합된 감성분석 알고리듬 구축
  - Word2Vec : 텍스트의 특성을 반영하는 수치화된 입력자료 도출(word embedding)
  - CNN : 문장 내 전후관계에 관계 없는 단어 간 패턴(예: 해수면 오염, 2℃ 시나리오) 반영
  - RNN : 문장 내 전후관계에 따른 감성에 미치는 영향의 차이를 반영



		감성태그
긍정		POS
부정	두려움	TER
	슬픔	GRI
	분노	RAG
중립		NEU
객관		OBJ

(5) 미세먼지 농도 및 예보가 서울 대중교통 이용에 미치는 영향

- 목적 : 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 농도 및 예보가 지하철 이용에 미치는 영향을 분석하고 이를 반영하여 지하철 이용을 예측
  - 자료: 서울시 지하철 승하차 정보(서울 열린 데이터 광장, 공공데이터 포털) , 기상기후 데이터(기상자료개방포털), 미세먼지 데이터(에어코리아)
  - 미세먼지 농도가 높거나 예・경보가 발령되었을 경우 지하철 이용의 변화를 통계 분석 및 머신러닝 기법을 적용하여 분석
  - 머신러닝 기법, 시계열 분석 방법을 적용하여 지하철 이용 및 혼잡도를 예측
    - 시간 특성(첨두, 비첨두 시간), 요일, 지역인구 데이터를 활용하여 미세먼지 이외 요인을 제어
- 방법론: 다양한 통계 방법론을 사용하여 분석 알고리듬을 개발
  - 방법론 후보군: 회귀분석, SVM(Supporting Vector Mechanism), Boosted Tree
  - 실시간으로 변화하는 자료의 특성을 반영하여 추정 결과를 상시적으로 갱신하는 발신 방식을 고민

유동인구 정보: 서울시 지하철 승하차 정보

- 서울열린데이터광장, 공공데이터포털
- 2010~2017.04 1~4호선 역별 시간별 승하차 인원
- 2015~2018.01 1~8호선 역별 일별 승하차 인원
- 컬럼 정보: 날짜, 호선, 역명, 구분(승하차), 시간(1시간)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
1	날짜	호선	역명	구분	00~01	01~02	02~03	03~04	04~05	05~06	06~07	07~08	08~09	09~10	10~11	11~12	12~13	13~14	14~15	15~16	16~17	17~18	18~19	19~20	20~21
2	2014-01-01	1호선	서울역(15(	승차	91				115	428	512	601	1099	1617	1941	2757	3423	3550	2957	4179	3633	3212	2667	3279	272(
3	2014-01-01	1호선	서울역(15(	하차	172				3	463	1231	1073	1280	1516	1738	2074	2426	2743	2803	2718	3013	3033	2849	2274	213
4	2014-01-01	1호선	시청(151)	승차	8				10	134	134	99	161	253	251	369	503	592	853	1087	1266	1266	1398	1081	118
5	2014-01-01	1호선	시청(151)	하차	38					39	231	226	463	576	585	876	900	1170	1158	1143	699	628	451	304	26
6	2014-01-01	1호선	종각(152)	승차	28				37	853	448	261	286	314	471	568	1006	1423	1601	1963	2620	2687	2646	2091	218
7	2014-01-01	1호선	종각(152)	하차	35					66	276	392	768	1057	1202	1547	1887	2080	2202	1930	1827	1632	1346	832	50
8	2014-01-01	1호선	종로3가(1!	승차	15				4	406	255	154	210	315	437	715	1153	1403	1575	1925	2178	2284	1857	1549	158
9	2014-01-01	1호선	종로3가(1!	하차	84					33	102	171	332	685	952	1469	2035	2351	2371	1936	1613	1258	919	595	40
10	2014-01-01	1호선	종로5가(1!	승차	1					64	92	70	121	179	294	416	610	743	870	1032	944	1015	891	692	54
11	2014-01-01	1호선	종로5가(1!	하차	33					24	89	125	249	290	464	716	987	1074	1112	1004	791	791	526	332	20
12	2014-01-01	1호선	동대문(15!	승차	3				15	167	150	180	279	410	490	556	651	859	923	950	1005	869	674	534	40
13	2014-01-01	1호선	동대문(15!	하차	89					22	123	105	153	326	472	858	1040	1150	1140	1025	972	824	602	510	42
14	2014-01-01	1호선	신설동(15(	승차	2				1	104	108	157	276	294	356	415	424	548	580	540	567	558	457	295	27
15	2014-01-01	1호선	신설동(15(	하차	37					27	127	103	157	213	300	370	505	463	447	402	446	465	387	350	38
16	2014-01-01	1호선	제기동(15	승차	1				5	86	95	118	218	351	410	584	772	933	1067	1239	1341	979	542	329	23

환경, 기후정보: 에어코리아 및 기상자료공개포털

- 서울시 미세먼지 데이터 (에어코리아)
  - 2010~2017 측정소별 미세먼지 농도: 지역, 측정소명, 측정일시, 주소, SO2, NO2, CO, O3, PM10
- 서울시 기상기후 데이터 (기상자료공개포털)
  - 2010~2017 관악산, 서울 측정소 : 지점, 시간, 기온, 강수량, 풍속, 풍향, 습도, 기압, 일조, 일사, 적설 등

서울시 미세먼지 데이터

	A	B	C	D	E	F	G	H	I	J	K
1	지역	측정소명	측정일시	SO2	CO	O3	NO2	PM10	PM25	주소	
2	서울	중구	2016070101	0.004	0.3	0.024	0.025	39	30	서울 중구 덕수궁길 15	
3	서울	중구	2016070102	0.004	0.2	0.027	0.019	44	38	서울 중구 덕수궁길 15	
4	서울	중구	2016070103	0.004	0.1	0.03	0.017	28	26	서울 중구 덕수궁길 15	
5	서울	중구	2016070104	0.003	0.2	0.031	0.013	18	12	서울 중구 덕수궁길 15	
6	서울	중구	2016070105	0.003	0.2	0.028	0.014	32	28	서울 중구 덕수궁길 15	
7	서울	중구	2016070106	0.003	0.1	0.02	0.024	22	19	서울 중구 덕수궁길 15	
8	서울	중구	2016070107	0.003	0.1	0.012	0.035	14	13	서울 중구 덕수궁길 15	
9	서울	중구	2016070108	0.003	0.2	0.009	0.039	20	13	서울 중구 덕수궁길 15	
10	서울	중구	2016070109	0.003	0.3	0.006	0.044	24	12	서울 중구 덕수궁길 15	
11	서울	중구	2016070110	0.003	0.3	0.006	0.044	22	13	서울 중구 덕수궁길 15	

서울시 기상기후 데이터

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	P	Q	R	S
1	지점	일시	기온(℃)	강수량(mm)	풍속(m/s)	풍향(16방위)	습도(%)	증기압(hPa)	이슬점온도(℃)	현지기압(hPa)	해면기압(hPa)	일조(hr)	일사(MJ/m2)	적설(cm)	전운량(10분위)	중하층운량(10분위)	운형(운형약어)	최저운고(100m)
2	108	2017-01-01 1:00	0		1.4	20	78	4.8	-3.3	1018.9	1029.9					4		7
3	108	2017-01-01 2:00	-0.3		1.9	50	81	4.9	-3.1	1018.5	1029.4					1		8
4	108	2017-01-01 3:00	-0.7		2	50	84	4.9	-3	1018.8	1029.8				0	0		
5	108	2017-01-01 4:00	-1.1		1.6	20	85	4.8	-3.3	1018.6	1029.6				0	0		
6	108	2017-01-01 5:00	-1.4		1.4	50	86	4.8	-3.4	1018.3	1029.3				0	0		
7	108	2017-01-01 6:00	-1.5		1.6	20	87	4.8	-3.3	1018.1	1029.1				3	3	Sc	7
8	108	2017-01-01 7:00	-1.5		1.4	20	87	4.8	-3.3	1018.6	1029.6				8	8	Sc	7
9	108	2017-01-01 8:00	-1.3		1.4	20	87	4.9	-3.1	1019	1030	0	0.01		8	8	Sc	10
10	108	2017-01-01 9:00	-0.4		1.6	20	83	4.9	-2.9	1019.4	1030.4	0	0.16		9	9	Sc	10
11	108	2017-01-01 10:00	0.8		2.1	50	77	5	-2.7	1020.1	1031	0.1	0.28		9	9	Sc	10
12	108	2017-01-01 11:00	2.5		1.9	50	71	5.2	-2.2	1019.9	1030.7	0.6	0.71		9	9	Sc	10
13	108	2017-01-01 12:00	4		1.3	50	69	5.6	-1.1	1018.7	1029.5	0.2	0.86		9	9	Sc	8
14	108	2017-01-01 13:00	5.1		1.4	20	65	5.7	-0.9	1017.9	1028.6	0.1	0.73		8	7	ScCi	8
15	108	2017-01-01 14:00	6.7		0.7	340	61	6	-0.3	1017.1	1027.7	0.7	0.9		8	7	ScCi	7
16	108	2017-01-01 15:00	6.9		0.9	230	65	6.4	0.7	1016.8	1027.4	0	0.41		9	9	Sc	8

4. 사업관리

기간, 인력, 예산

- 기간: 2018년 1월 – 2018년 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

연구진 구성

연구진	역할
강성원 선임연구위원(책임)	• 과제 총괄
진대용 부연구위원(부책임)	• 플랫폼 구축 및 연구동향 서비스 개발 총괄
명수정 연구위원	• Open Data Map 구축 참여
홍한움 부연구위원	• 데이터 기반 한강 수질 예측
이동현 한국산업기술대 교수(위탁)	• 컨벌루션 신경망(CNN)을 통한 미세먼지 예측
한국진 선임전문원	• 플랫폼 구축 및 연구동향 서비스 지원
강선아 위촉연구원	• 딥러닝 이용 국내 노인인구 호흡기 질환 사망 위험 추정
김도연 위촉연구원	• 딥러닝 기반 환경이슈 감성 분석기 개발
김진형 연구원	• 미세먼지 농도 및 예보가 서울 대중교통 이용에 미치는 영향 • 분석 결과 온라인 출판

보고서 목차 및 작업계획

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										
	2) 빅데이터 분석 플랫폼										
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) 딥러닝 기반 환경이슈 감성 분석기 개발										
	5) 미세먼지 농도 및 예보가 서울 대중교통 이용에 미치는 영향										
4. 환경 빅데이터 서비스	연구동향 파악 서비스										
5. 결론	연구결과 요약 및 시사점										

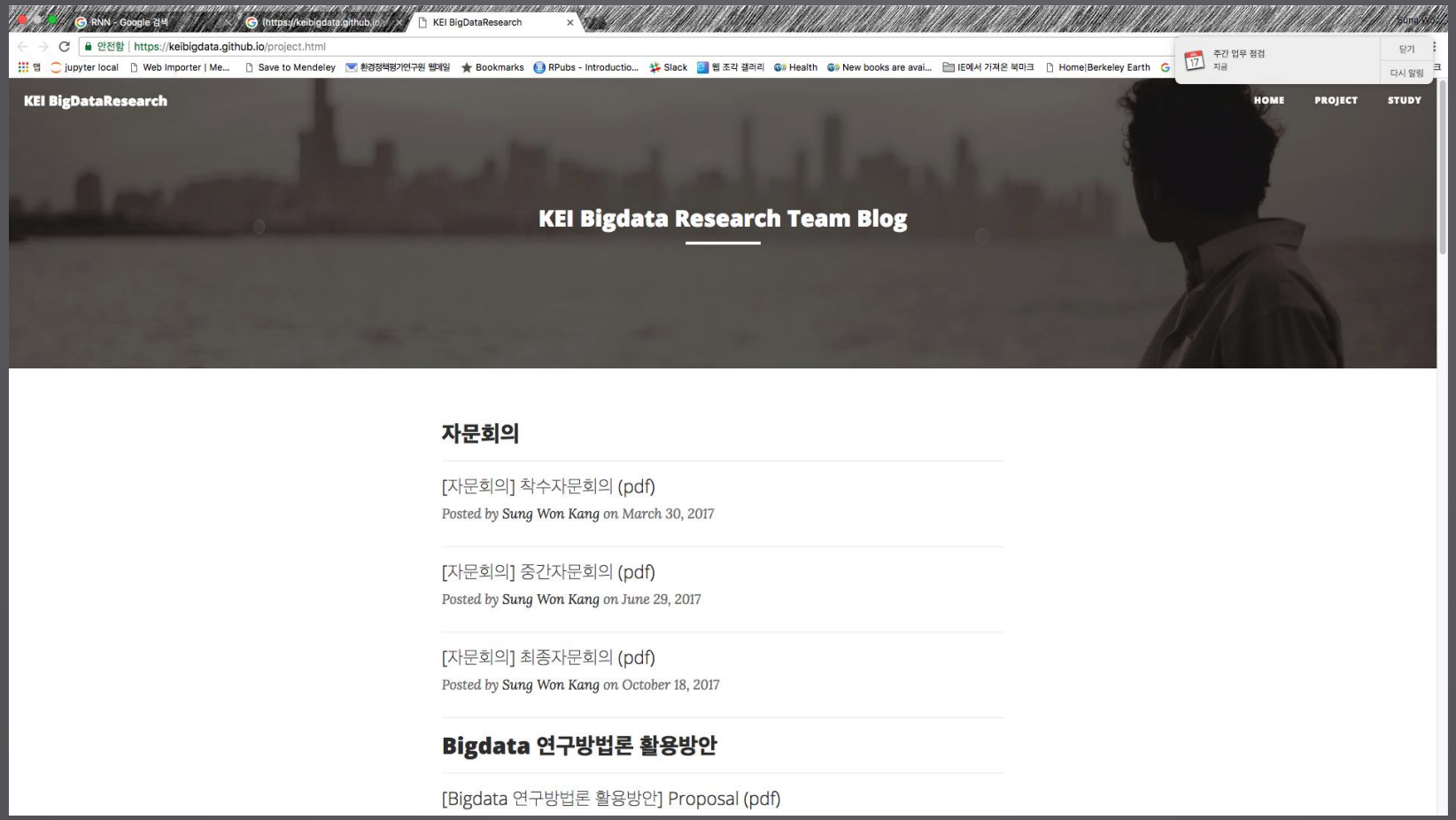
연구관리

◆ 주 1회 정기 meeting : 세부과제 연구상황 공유
- 매주 수요일 오후 3시 : 환경 빅데이터 연구 인프라 구축
- 매주 목요일 오전 10시 : 환경 빅데이터 연구

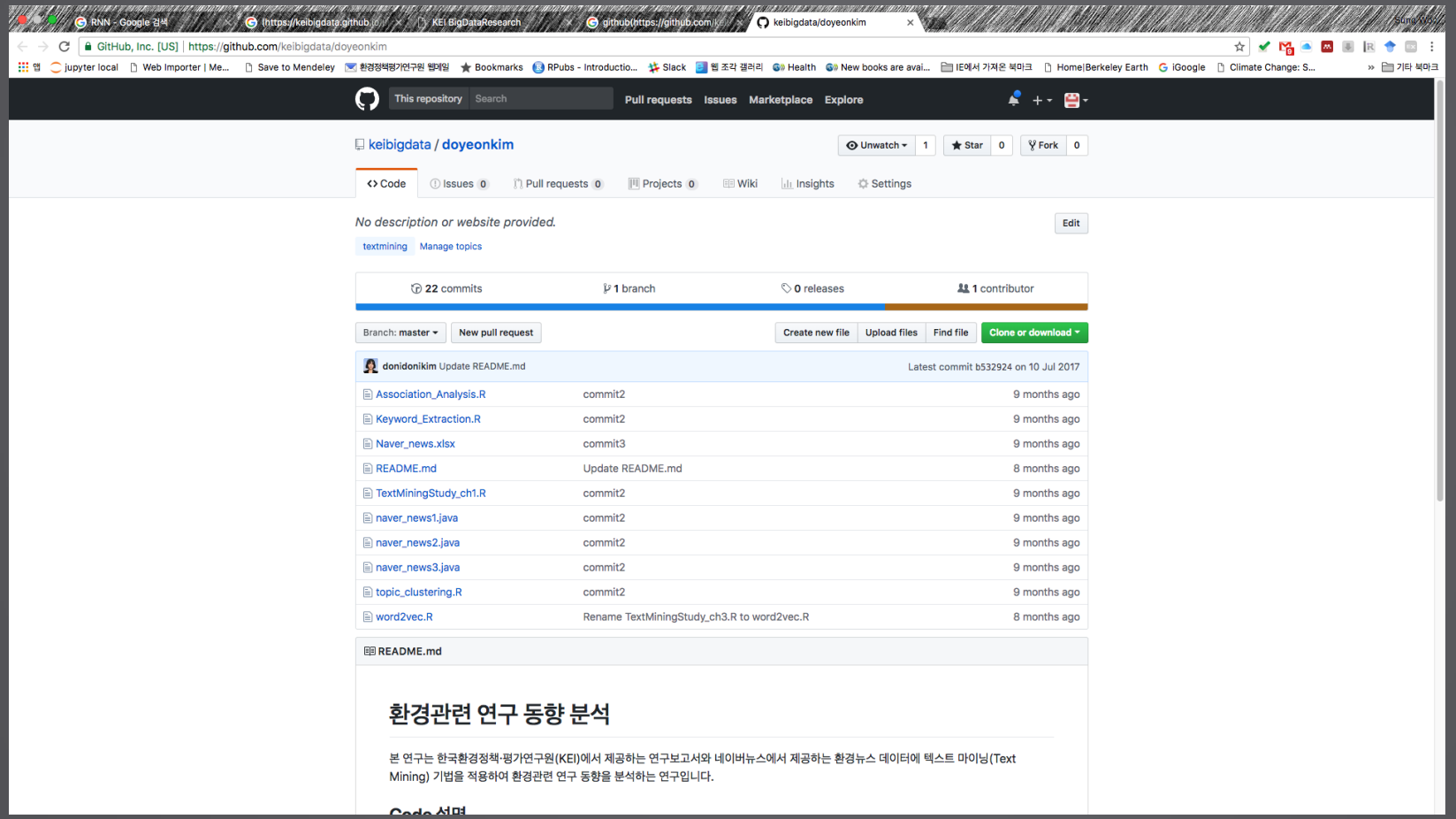
◆ 월 1회 Progress Seminar 실시 : 연구진 전원 참여 및 외부 전문가 자문

◆ Working paper 상태의 중간 산출물을 온라인에 게시하여 피드백 기회를 확대
- 홈페이지(https://keibigdata.github.io/project.html),
- GitHub(https://github.com/keibigdata/)

연구 결과물 게시 : 홈페이지



연구 결과물 게시 : 알고리듬과 자료



5. 기대효과

빅데이터 분석 적용 사례 및 역량 축적

◆ 환경 빅데이터 연구 인프라 구축 : 연구자 친화적 환경 데이터 접근 Gate 를 구축하고 빅데이터 분석 인프라를 설계
- Open Data Map : 연구자 친화적 환경 데이터 접근 Gate
- 빅데이터 분석 서버 시험운영: 향후 원내 빅데이터 연구 공간 운영의 테스트베드 구축

◆ 환경 빅데이터 연구: 비정형 빅데이터 연구 환경연구 적용 가능성 점검
- 실시간 수집 텍스트 데이터·화상 데이터 분석 기능 환경연구 적용 가능성 진단
- 환경 빅데이터 연구 역량 축적
  - 4개 수치자료 분석 알고리듬, 1개 텍스트자료 분석 알고리듬 구축
  - 1개 이상 실시간 데이터 반영 결과 update 발신 경로 구축 (유동인구)
  - 환경 사전 및 환경 텍스트 감성 분류 데이터 구축 (감성분석)

◆ 환경 빅데이터 서비스 : 연구 결과 기반 서비스 제공 시작
- 2017년 연구 결과를 이용하여 연구동향 파악 서비스를 제공

감사합니다