환경 빅데이터 분석 및 서비스 개발

착수자문회의(2017.3.30)

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	2017.1 ~2017.12	
연구진	강성원 연구위원(책임) 한국진 전문원 김진형 연구원 김도연 위촉연구원 강선아 위촉연구원 정은혜 위촉연구원 이동현 한국산업기술대 교수(위탁)	
자문위원	내부	명수정 연구위원 배현주 부연구위원 이명진 부연구위원
	외부	김종률 과장 (환경부 정책총괄과) 우석진 교수 (명지대학교 경제학과) 강희찬 교수 (인천대학교 경제학과) 이성호 박사 (한국개발연구원)
자문일정	착수자문회의: 2017년 3월 중간자문회의: 2017년 7월 최종자문회의: 2017년 10월	

2. 연구 목적

환경 빅데이터 분석 및 서비스 개발

1. 빅데이터 방법론 적용 환경연구 개발

- 환경 빅데이터 연구 : 주제 선정, 데이터 수집 및 가공, 데이터 분석 과정에서 빅데이터 연구 기법을 활용

  - 주제선정: 텍스트 마이닝, 자연언어 분석 기법을 미디어에 적용하여 연구수요 파악

  - 수집 및 가공: Scraping, Crawling 등 온라인 자료 수집 기법을 활용하여 연구자료 확보

  - 분석: 시-공간 해상도가 높은 패턴 분석 및 예측이 가능한 빅데이터 방법론 적용
    - [전망]: 오염원, 오염도 결정요인, 건강 정보→ 시간, 공간 해상도 높은 환경위험 예측
    - [정책 설계]: 데이터 기반 오염물질 MRV 시스템 구축 환경 규제 효과성 제고
      - 예: 영상, 소리 등 비정형 데이터와 오염도, 오염물질 배출량 간의 패턴 파악
    - [정책 평가]: 기(旣) 시행 정책수행 지표와 환경위험 지표간 패턴 분석 → 공간 해상도 높은 정책평가

예시: 환경 Bigdata 연구

환경부 조직			빅데이터 연구 주제		
기획조정실	기획재정담당관		지자체예산소요예측		
	창조행정담당관		환경민원NLP	환경미디어NLP	
	규제개혁법무담당관		환경규제성과패턴분석	규제미디어NLP	규제소송NLP
	정보화담당관		환경정보수집알고리듬(MRV)	환경IOT알고리듬(MRV)	
	비상안전담당관		화학사고가능성예측		
	국제협력관	해외협력담당관			
		지구환경담당관			
환경정책실	환경정책관	정책총괄과	환경규제성과패턴분석		
		환경기술경제과	환경기술RandD성과예측		
		환경협력과			
		환경산업과	환경산업체기업성과예측		
	환경보건정책관	생활환경과	환경민원NLP		
		환경보건관리과	환경성질환유병율예측		
		화학물질정책과	화학사고가능성예측		
		화학안전과	화학사고가능성예측		
	기후대기정책관	기후대기정책과	기후예측	이상기후현상예측	
		기후변화협력과	기후예측	이상기후현상예측	
		기후변화대응과	기후예측	이상기후현상예측	
		대기관리과	대기오염물질배출량및오염도예측		
		교통환경과	교통부문대기오염물질배출량예측		
		신기후체제대응팀			

예시: 환경 Bigdata 연구

환경부 조직			빅데이터 연구 주제	
물환경정책국	물환경정책과	물환경정책과	수질오염도예측	
	유역총량과	유역총량과	수질오염도예측	
	수생태보전과	수생태보전과	수질오염도예측	
	수질관리과	수질관리과	수질오염도예측	하수배출량예측
	상하수도정책관	수도정책과	수자원수요예측	
		생활하수과	하수배출량예측	수질오염도예측
		토양지하수과	지하수수질예측	토양오염예측
자연보전국	자연정책과	자연정책과		
	생물다양성과	생물다양성과		
	공원생태과	공원생태과		
	국토환경정책과	국토환경정책과	토양오염예측	
	국토환경평가과	국토환경평가과		
자원순환국	자원순환정책과	자원순환정책과	폐기물배출량예측	
	폐자원관리과	폐자원관리과	폐기물배출량예측	
	자원재활용과	자원재활용과	폐기물배출량예측	
	폐자원에너지과	폐자원에너지과	폐기물배출량예측	

2. 연구기반 구축 및 활용방안 모색

- 환경 빅데이터 연구 인프라 구축: 연구 자료 및 알고리듬을 오픈 소스로 공개하여 향후 환경 빅데이터 연구 인프라 제공
  - 환경 빅데이터 연구자료/알고리듬 오픈소스로 공개하여 커뮤니티 형성
  - 산재된 원내외 환경관련 자료 수집-추출 사례 축적 및 공개
  - 중장기적으로 환경연구에 특화된 빅데이터 플랫폼 구축
- 원내외 빅데이터 서비스 개발: 연구 성과를 활용하여 원내외 연구정보서비스 및 공공서비스를 개발
  - 예) KEI 연구보고서 시각화 서비스, 연구자 네트워트 시각화 서비스

환경 관련 DB 현황

관련 분야별 DB	서 비 스 U R L	관련 분야별 DB	서 비 스 U R L
환경영향평가	http://www.eiass.go.kr/	토양지하수	https://sgis.nier.go.kr/newsgis
기후변화 적응정보	http://vestap.kei.re.kr/ http://ace.kei.re.kr/ http://ccas.kei.re.kr/	화학물질배출량 DB	http://ncis.nier.go.kr/
		대기오염도	http://www.airkorea.or.kr/
		상수도	http://www.waternow.go.kr/
환경공간정보	http://ecvam.kei.re.kr/ http://egis.me.go.kr/	소음정보	http://www.noiseinfo.or.kr/
		기후변화 시나리오	http://sts.kma.go.kr/
환경가치 종합	http://evis.kei.re.kr	한국환경공단 자원순환정보시스템	https://www.recycling-info.or.kr/rrs/main.do
환경오염 방지 지출 통계	http://www.kosis.kr/	konetic 국가 환경산업기술정보시스템	https://www.konetic.or.kr
물환경	http://water.nier.go.kr/	온실가스종합정보센터	https://www.gir.go.kr
환경통계	http://stat.me.go.kr/	산자부 국가에너지통계 종합정보시스템	www.kesis.net

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

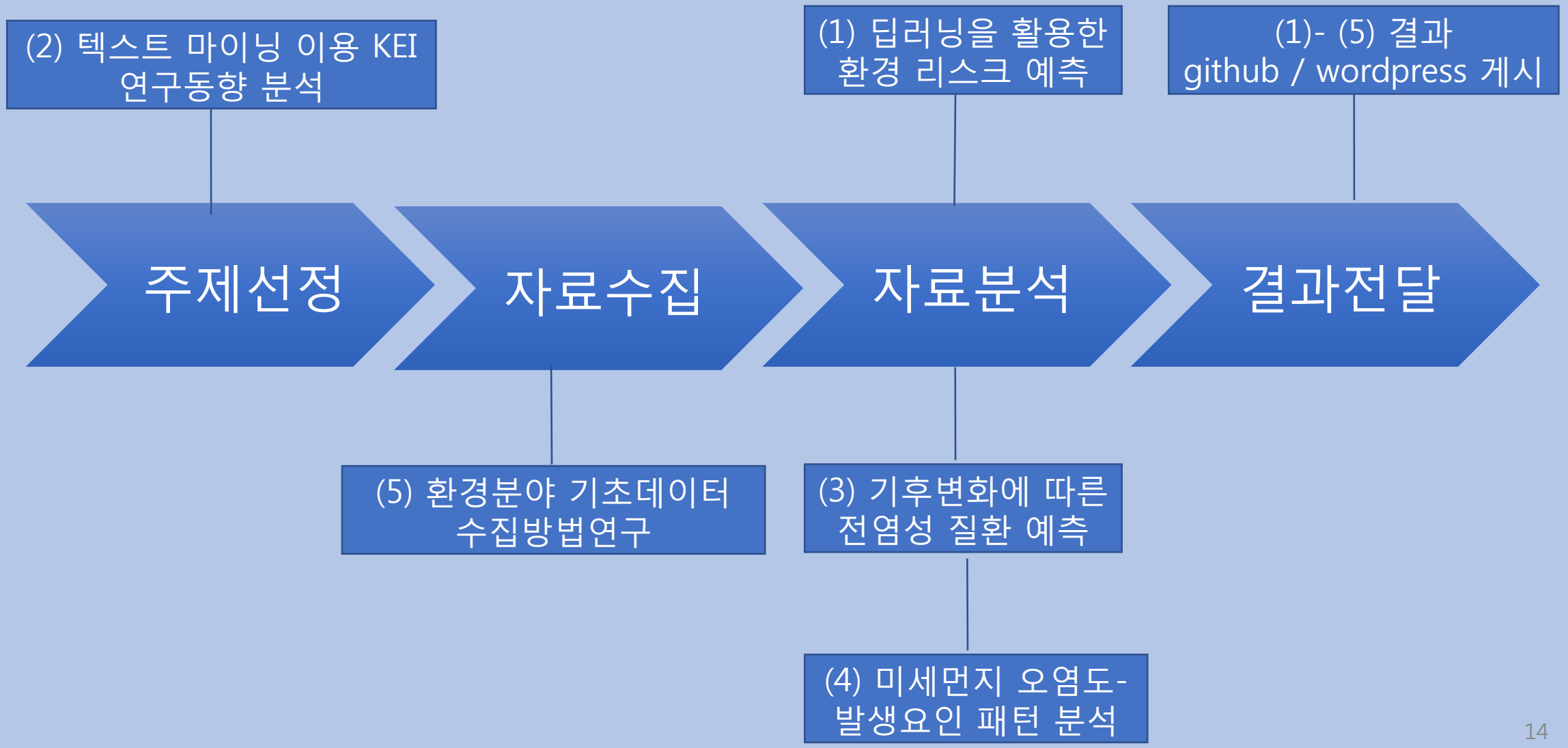
환경 빅데이터 분석 및 서비스 개발 연차계획

	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 수입 –추출하는 사례를 축적하여 오픈 소스로 공개
- 3. 원내외 빅데이터 서비스: 연구 정보 제공 서비스 개발
  - 연구 정보 추출 서비스 제공

2. 연구 내용 및 방법론

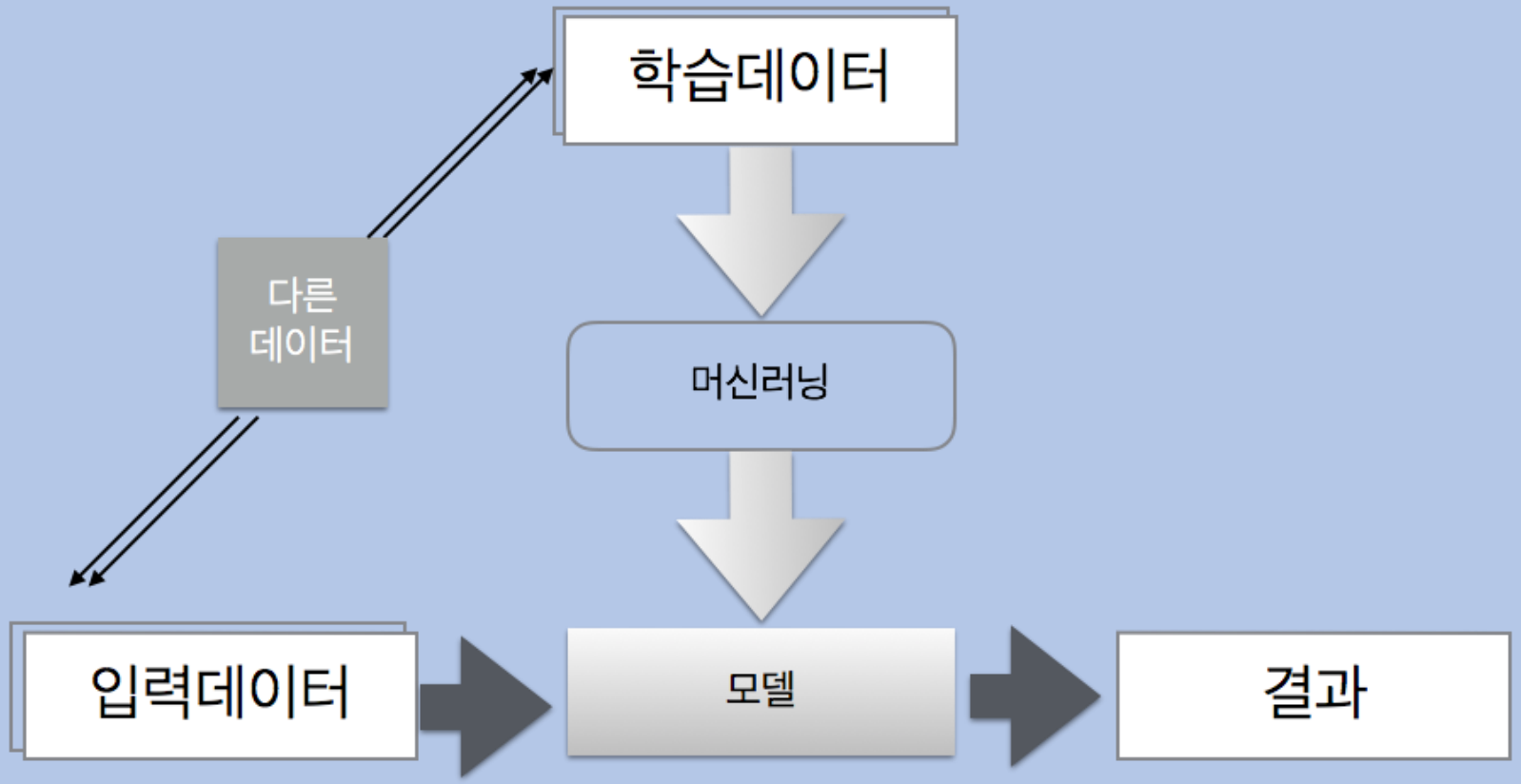

(2) 텍스트 마이닝 이용 KEI 연구동향 분석
(1) 딥러닝을 활용한 환경 리스크 예측
(1)- (5) 결과 github / wordpress 게시
주제선정
자료수집
자료분석
결과전달
(5) 환경분야 기초데이터 수집방법연구
(3) 기후변화에 따른 전염성 질환 예측
(4) 미세먼지 오염도-발생요인 패턴 분석

(1) 딥러닝을 활용한 환경 리스크 예측

- 대기, 수질오염 오염도 자료를 딥러닝 알고리듬으로 분석하여 시간-공간 해상도가 높은 오염도 예측치를 도출
  - 분석이 용이한 1개 매체를 집중적으로 연구
  - 자료축적 → 패턴 파악 → 오염도 예측 process 진행
    - [자료축적] 오염도 및 오염도 영향 요인 자료를 기초지자체 수준에서 축적
    - [패턴파악] 딥러닝 알고리듬을 이용하여 요인과 오염도 간의 패턴을 파악
    - [예측] 오염도 결정요인 전망치를 알고리듬에 투입하여 오염도를 예측
  - 딥러닝 이외의 전통적 추정방식(예: 공간시계열 분석) 과 성과 비교

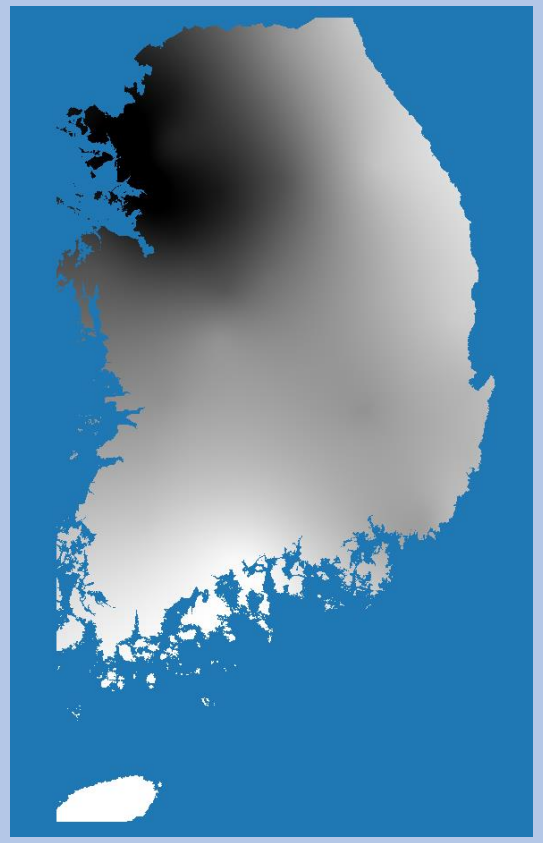
- 대용량 자료 분석도구를 사용 경험 축적
  - 대용량 자료의 병렬처리가 가능한 패키지(Tensorflow) 및 클라우드 서비스 (AWS: Amazon Web Service) 사용 경험 축적

빅데이터 분석 방법

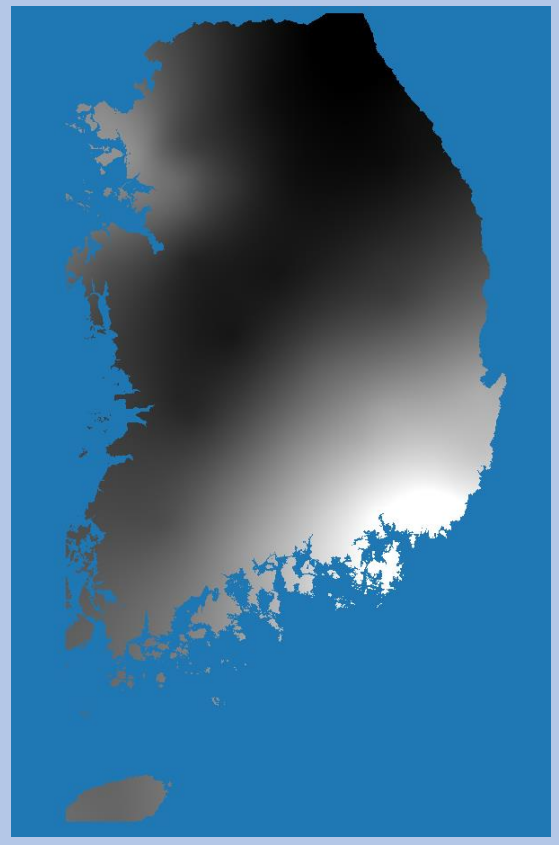


예) 미세먼지 오염도 분석

mean의 지리적 분포

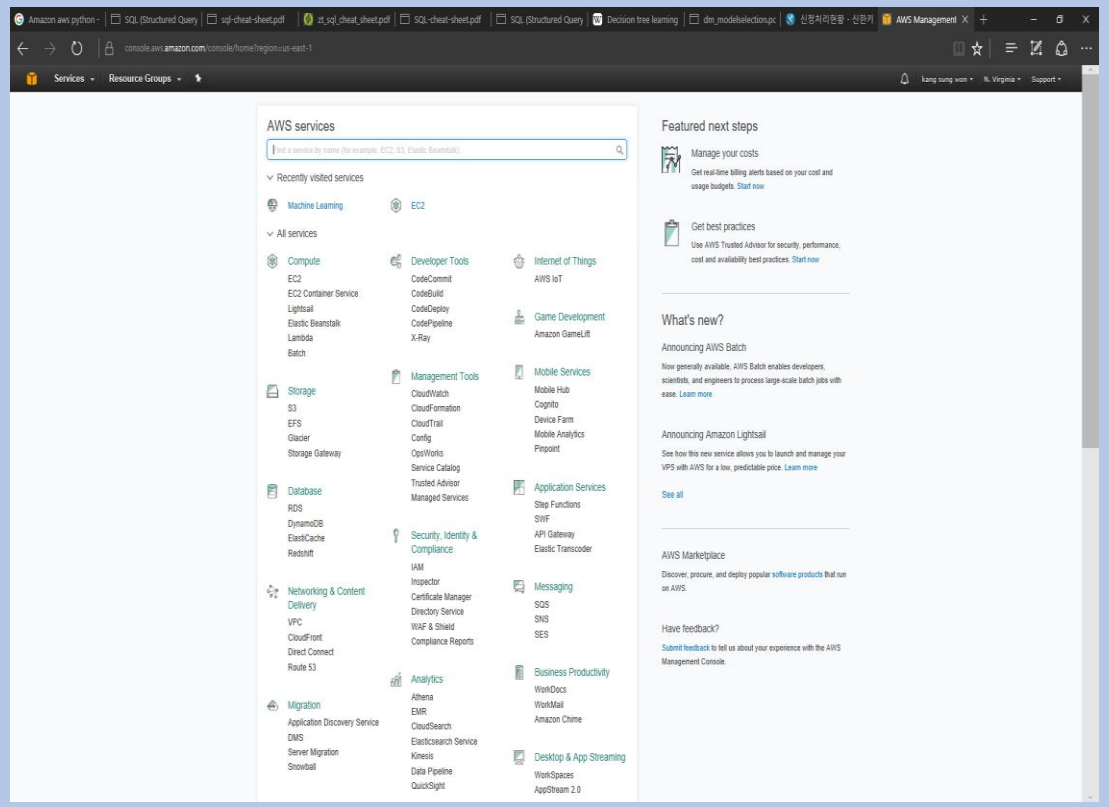


tail index의 지리적 분포

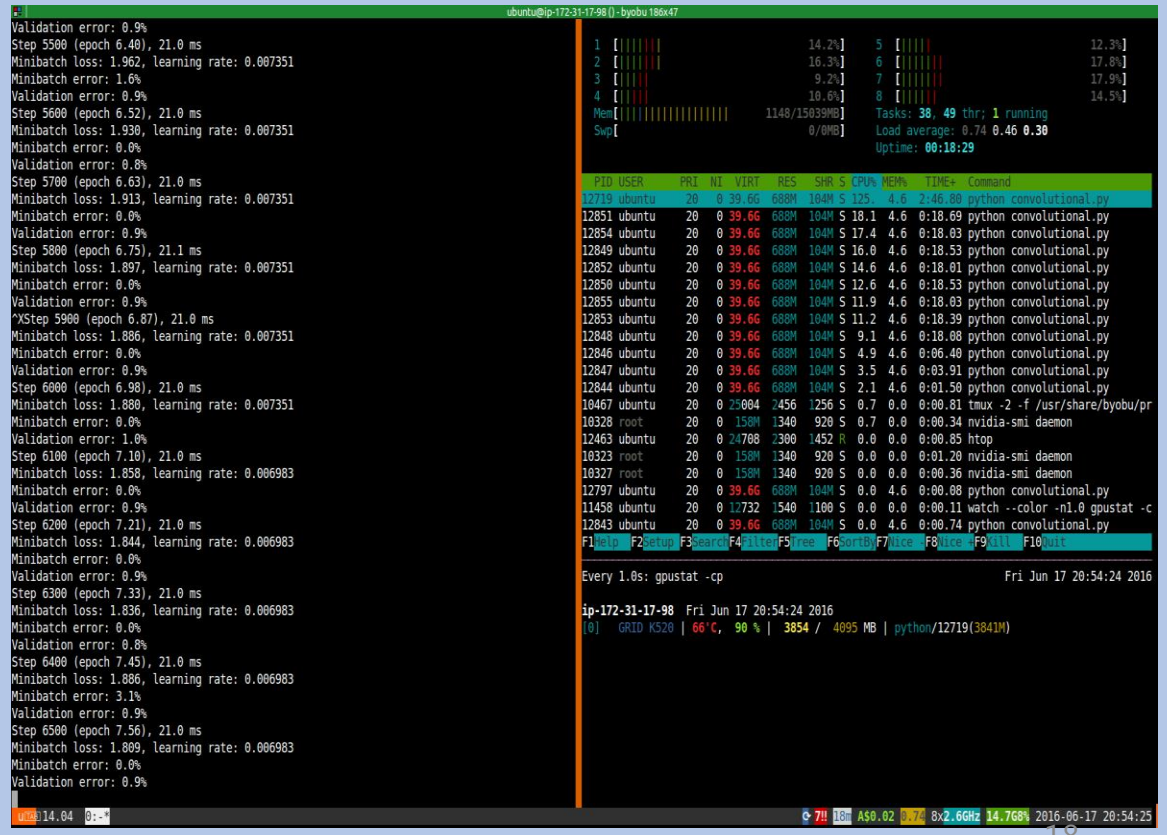


대용량 자료 처리 도구

Amazon Web Service (AWS)



Tensorflow on AWS



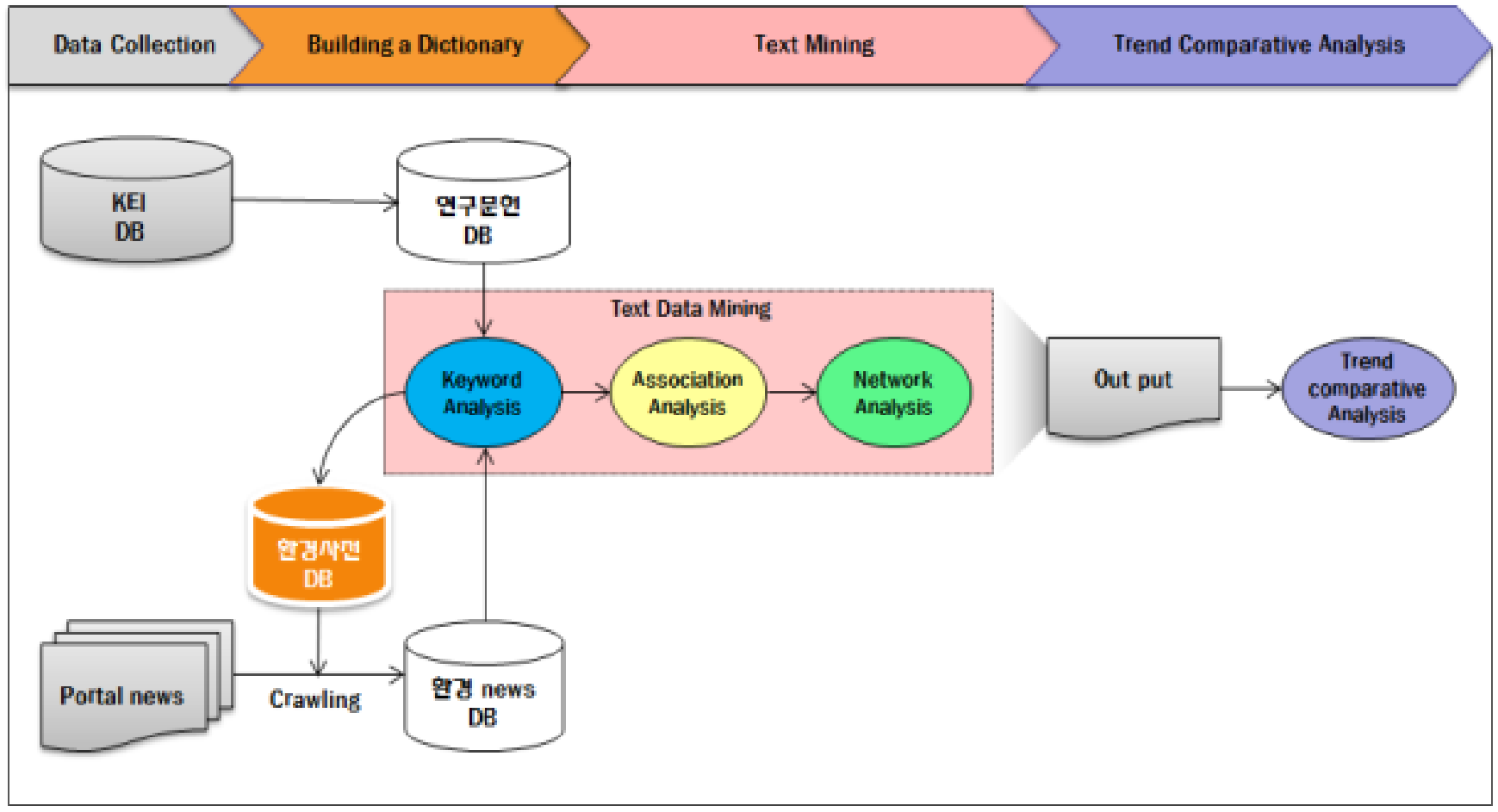
(2) 텍스트마이닝 이용 KEI 연구동향분석

- 1993-2016 KEI 사업계획서, 연구보고서 텍스트 분석
  - 연구보고서(제목, 목차, 요약, 날짜, 연구자) 1,679건 및 사업계획서(제목, 날짜, 연구자): 2,614 건 자료에 텍스트 마이닝 기법 적용
  - 자료 집적 → 환경 키워드 사전 구축 → 텍스트 마이닝 → 추세분석

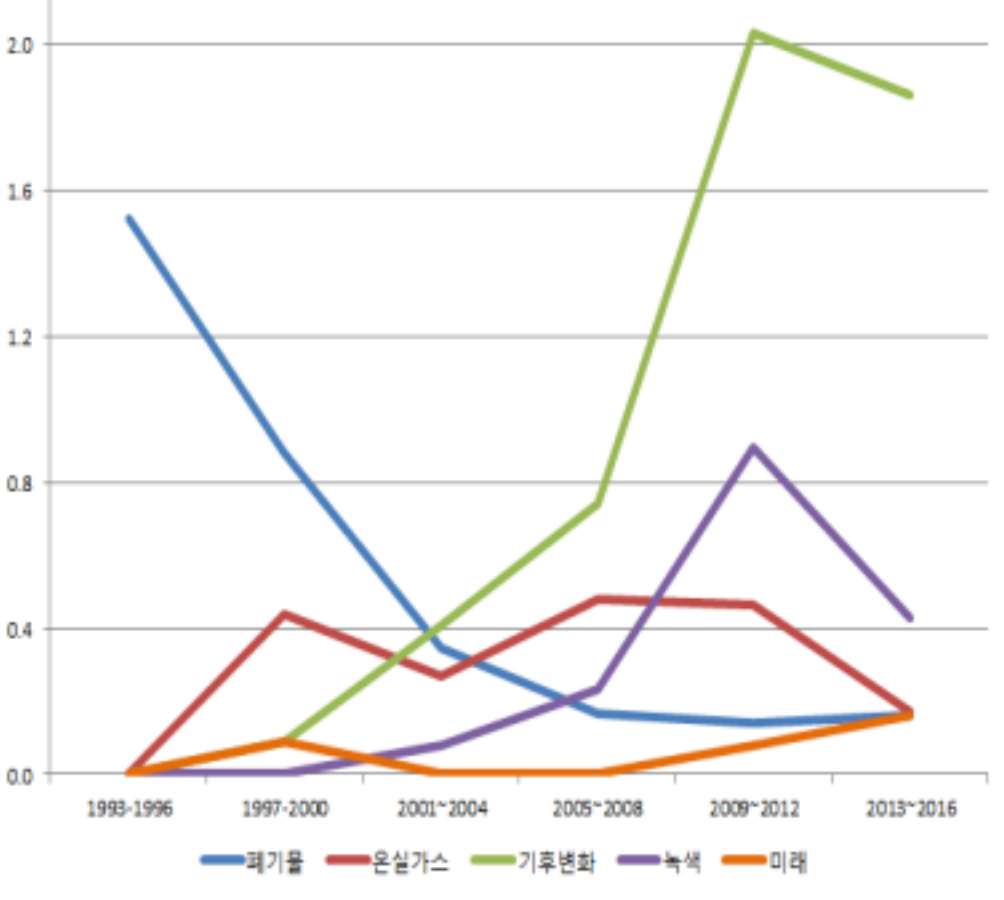
- 연구 보고서의 동향과 민간 연구수요 동향간의 조응 여부 점검
  - 민간 매체 (뉴스, 소셜미디어, 학술논문서지) 텍스트 분석을 병행하여 시계열 추이를 비교
    - 보고서 분석 결과인 연구공급동향과 매체 분석 결과인 연구수요동향간의 관계 파악

- 텍스트 마이닝 알고리듬을 원내 공개: 연구동향 파악 서비스 제공

KEI 연구동향 분석 작업 흐름도



keyword analysis: 1993-2016 사업기획서



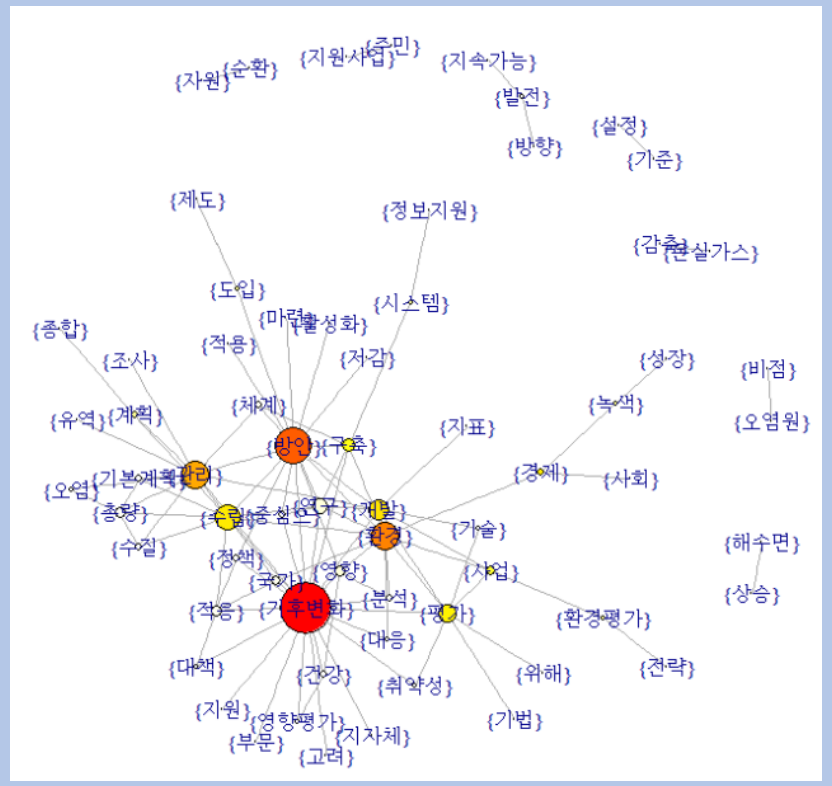
	폐기물	온실가스	기후변화	녹색	미래
1993~96	11	0	0	0	0
1997~00	20	10	2	0	2
2001~04	18	14	21	4	0
2005~08	15	43	66	21	0
2009~12	20	66	288	127	11
2013~16	32	34	366	84	32

Association Analysis : 1993-2016 사업기획서

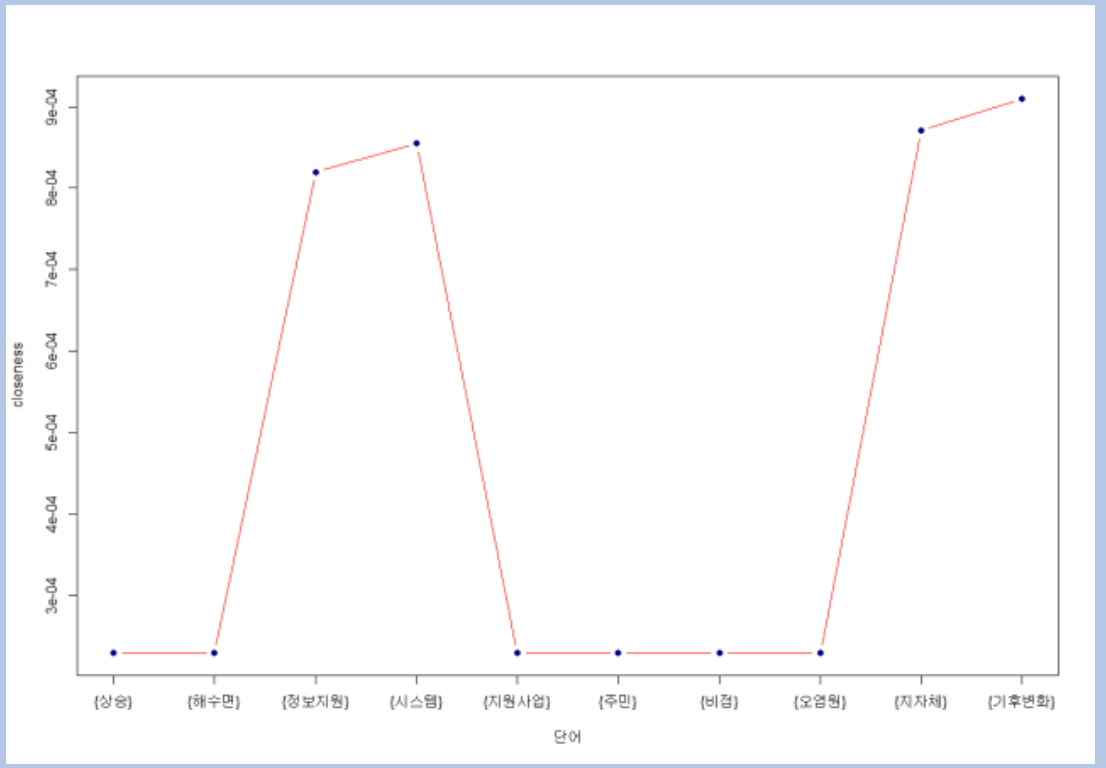
No	lhs		rhs	support	confidence	lift
1	상승	=>	해수면	0.006	1	129.429
2	해수면	=>	상승	0.006	0.714	129.429
3	정보지원	=>	시스템	0.009	1	24.322
4	시스템	=>	정보지원	0.009	0.208	24.322
5	지원사업	=>	주민	0.009	0.941	66.879
6	주민	=>	지원사업	0.009	0.627	66.879
7	비점	=>	오염원	0.006	0.742	68.943
8	오염원	=>	비점	0.006	0.59	68.943
9	지자체	=>	기후변화	0.006	0.606	5.985
10	기후변화	=>	지자체	0.006	0.054	5.985

Network Analysis: 1993-2016 사업기획서

연관어 네트워크 분석



단어 근접 중심성 분석 (상위 10개)



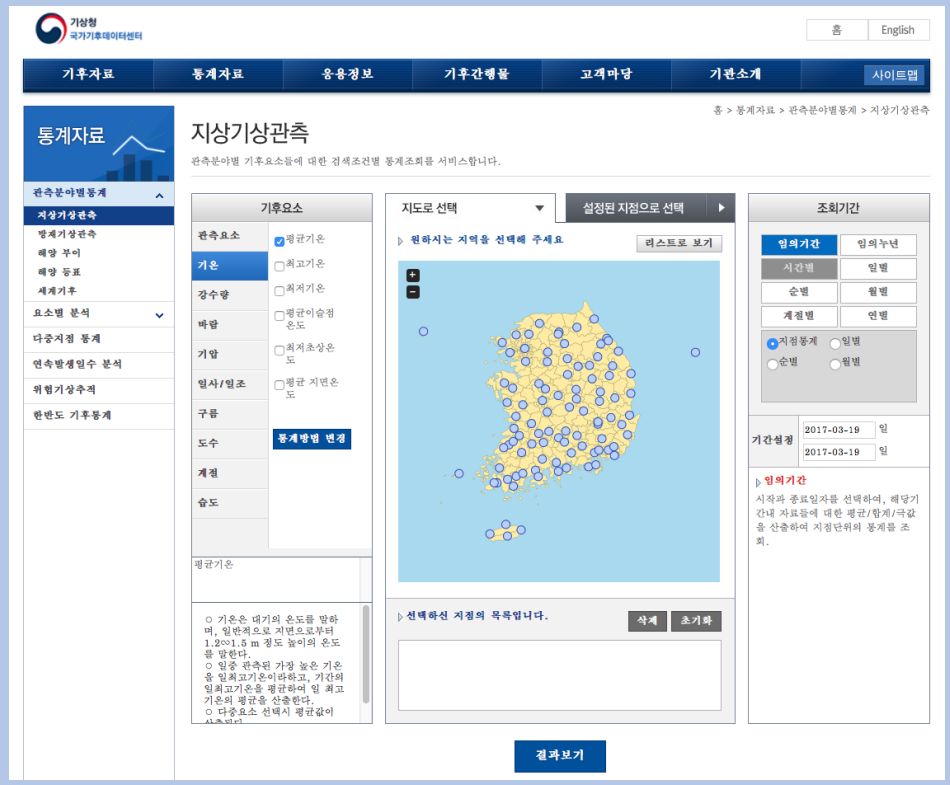
(3) 기후변화에 따른 전염성 질환 예측

- 건강보험 표본 코호르트자료와 기상청 기후자료를 연계하여 기후변화에 민감한 전염병의 발생을 예측
  - 자료: 건강보험 코호르트 자료, 기상청 국가기후데이터센터 자료(2005~2015)
  - 방법론 : RNN 을 적용하여 기초자체 단위 기후변화 민감 전염성 질환 발생 추이를 시계열로 파악
    - RNN(Recursive Neural Network): time dependency 가 있는 자료의 분석에 강점이 있는 Deep Learning 방법론
      - 전기(t-1) Data 학습의 경험을 State 에 축적하여 다음기(t, t+1,,,) Data 학습에 적용
    - 쯔쯔가무시, 말라리아, 세균성이질, 렙토스피라, 장염비브리오 5개 질환 대상

전염성 질환 예측 분석 대상

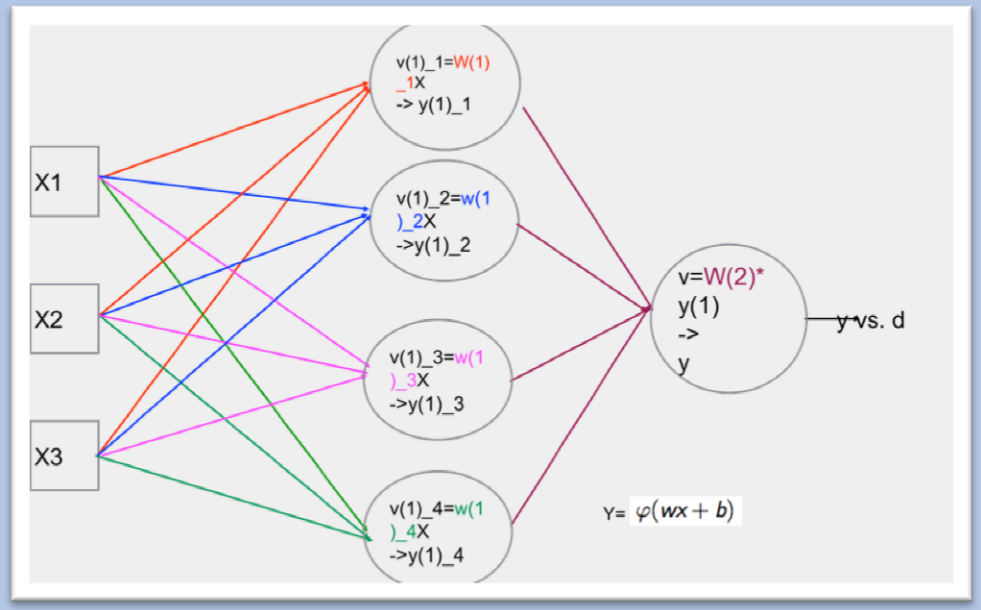
건강보험 코호르트 자료

기상청 국가기후데이터 센터

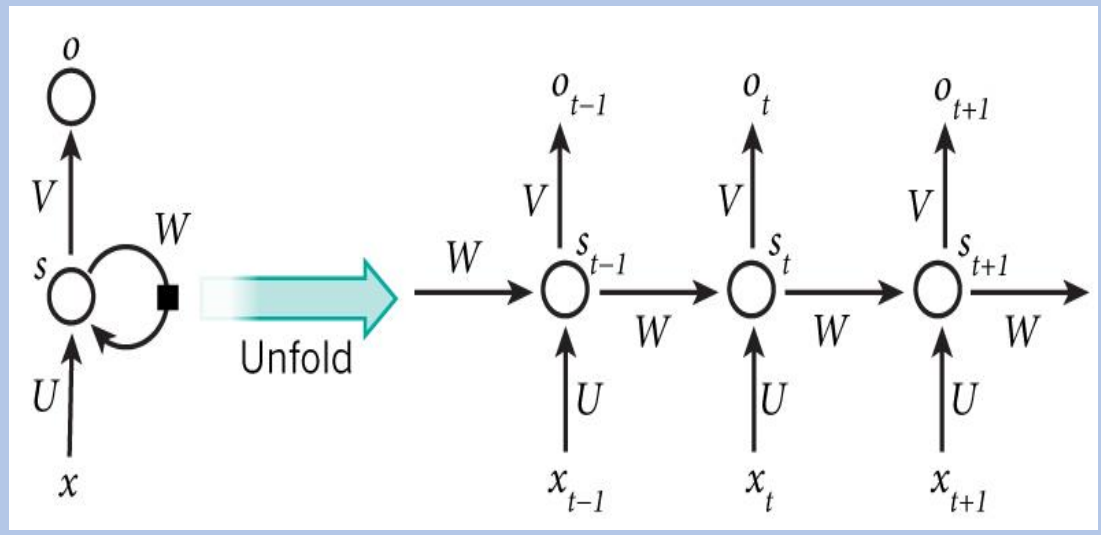


RNN의 특징: 과거학습의 정보 사용

RNN 이 아닌 다층신경망
(feed forward network)



RNN



(4) 미세먼지 오염도-발생요인 패턴분석

- 중요 요인을 파악하는 다양한 기계학습 기법을 활용하여 미세먼지 오염 발생 요인 파악
  - 기계학습 활용 인과관계 파악 방법론의 환경연구 적용 가능성 탐색
  - 변수 선택법(Variable selection), Decision Tree, Random forest 등 다양한 인과 파악 방법론 활용
    - 변수 선택법: Ridge Regression, Lasso Regression 등 최적의 변인을 파악하여 과적합(Overfitting)을 방지하는 방법론 적용
    - Decision Tree: 모형의 결과를 예측할 수 있는 요인들의 임계치를 파악
    - Random Forest: Decision Tree 모형의 예측력을 향상
      - 다수의 Decision Tree 모형의 ensemble 활용 + 각 Tree의 임계치 결정 요인의 과적합억제

미세먼지 발생요인 패턴 분석을 위한 변수 선정 및 데이터 출처

변수 분류	변수	
발생원인 변수	직접 발생 원인	NOx, SOx 등
	간접 발생 원인	VOCs, O3, NH3, 빛에너지 등
기후기상요인 변수	기온, 기압, 강수량, 바람 등	
사회경제적 변수	가계	인구, 인구밀도, 가계 소득, 난방비 등
	기업	대기오염 배출 사업장 및 배출량 정보 등
외부요인 변수	중국 동북부 미세먼지 농도, 황사발생일수 등	

▶ 변수는 문헌분석을 통해 향후 추가 혹은 제거될 수 있음

데이터 출처		변수
기상청 국가기후데이터센터	http://sts.kma.go.kr/jsp/home/contents/main/main.do	기후기상요인 변수
기상자료개방포털	https://data.kma.go.kr/cmmn/main.do	
에어코리아	http://www.airkorea.or.kr/index	발생원인 변수
국가통계포털	http://kosis.kr/	사회경제적 변수
환경공간정보서비스	https://egis.me.go.kr/main.do	
World Air Quality Index Sitemap	http://aqicn.org/map/china/kr/	외부요인 변수

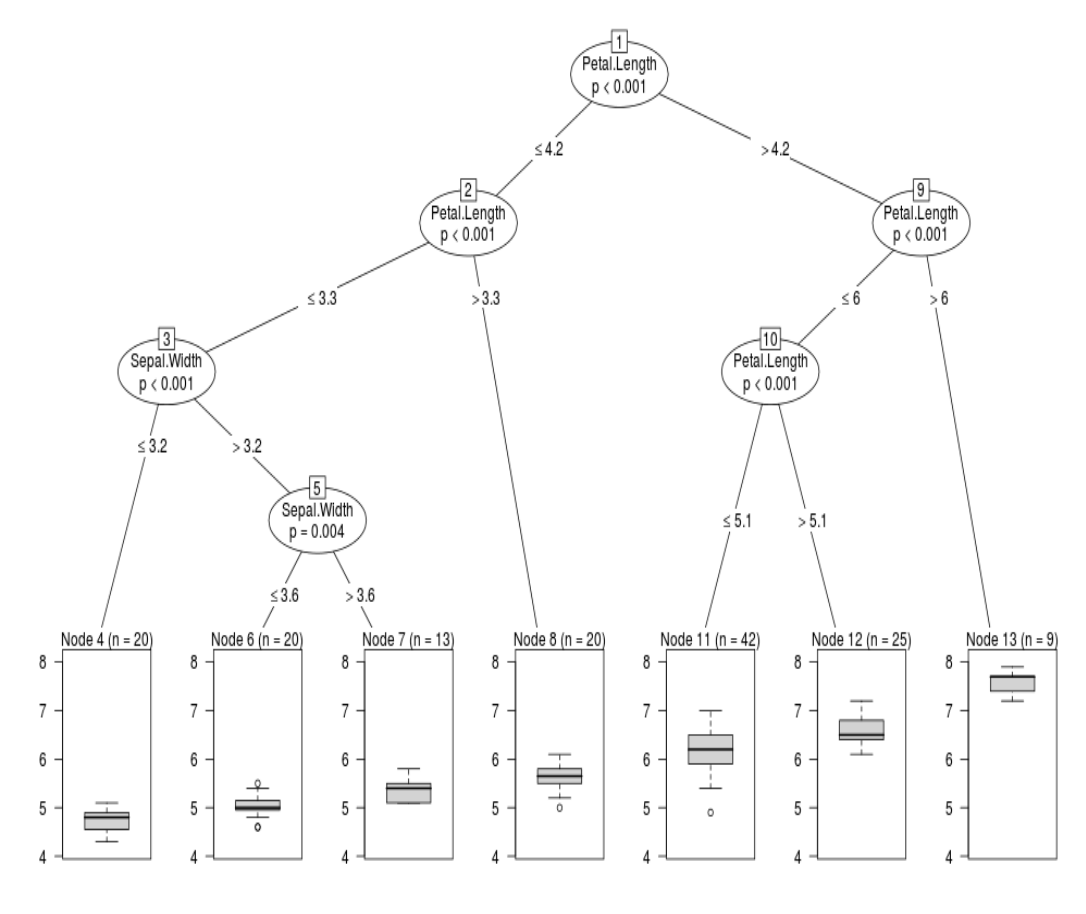
변수선택법 : 변수의 수에 penalty 부여

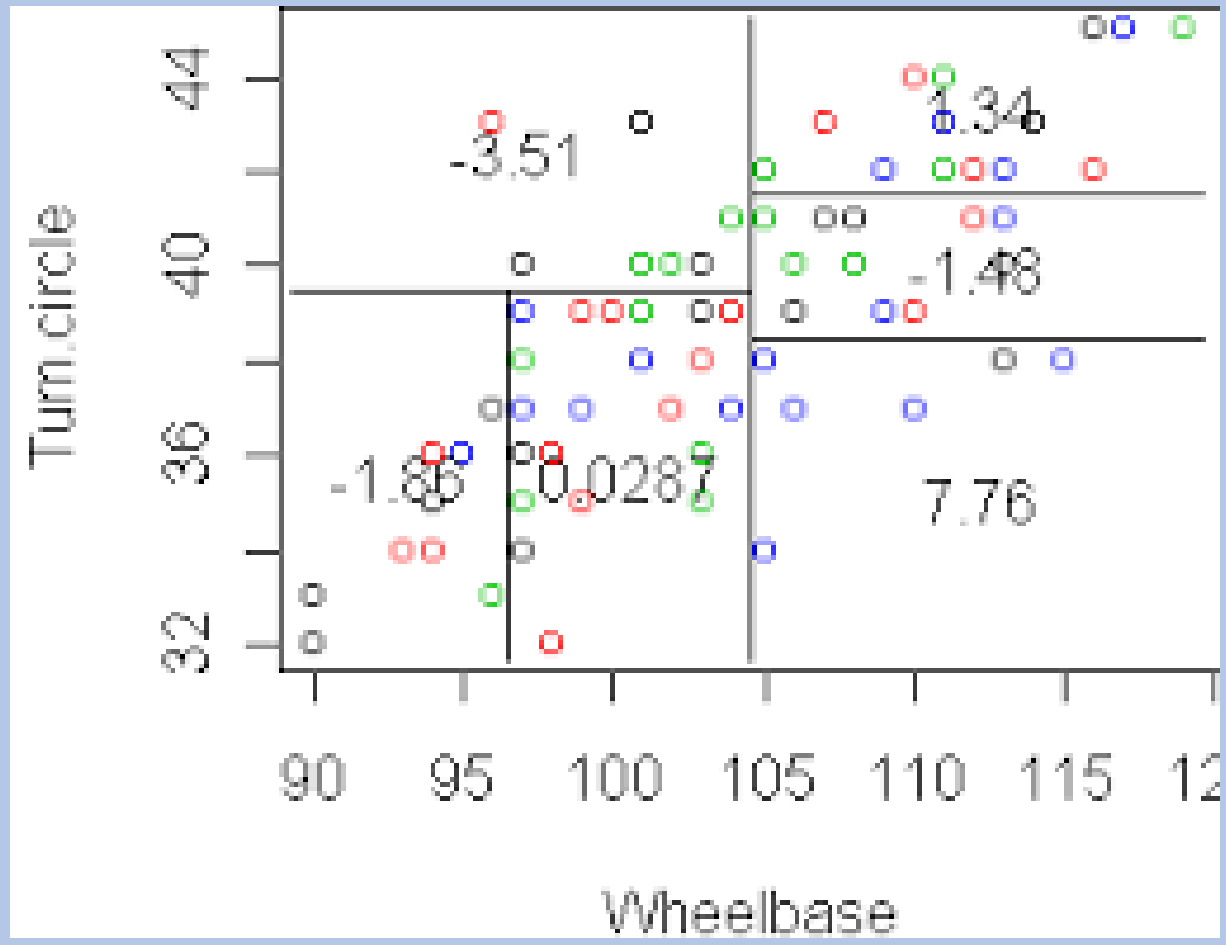
- Ridge Regreesion, Lasso Regression: 추정 모수의 Norm 에 페널티를 부과하는 항목을 비용함수에 포함
  - Rigde Regression 은 L2 norm, Lasso Regression은 L1 norm을 사용

Ridge: $$\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

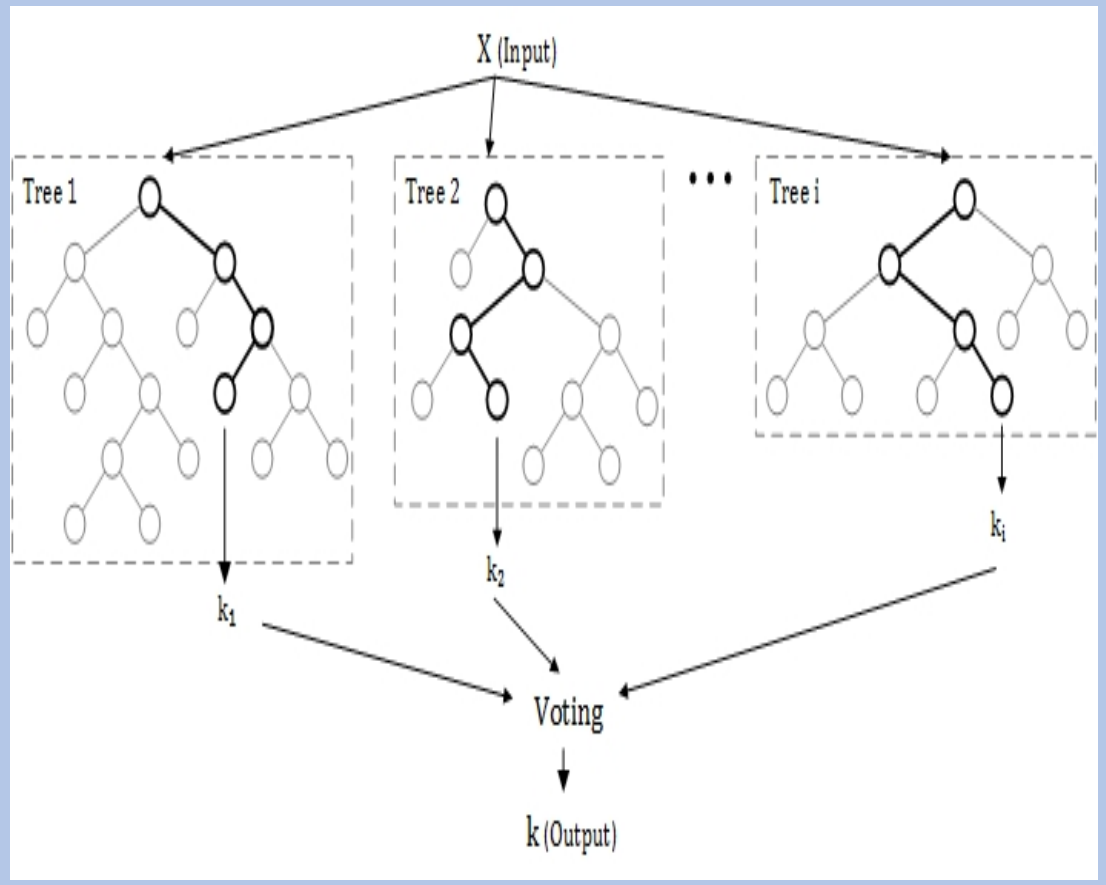
Lasso: $$\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Decision Tree: 결과 구분 변인 임계치 파악





Random Forest : Decision Tree ensemble



- Decision Tree의 예측력 제고
  - Decision Tree는 과적합 (Overfitting)에 따른 예측력 저하 문제에 취약
  - 복수의 Tree를 이용
  - 각각의 Tree 내에서 변수 선택법 등을 활용하여 최적 조합을 추출
  - 모든 Tree의 추정 결과를 종합

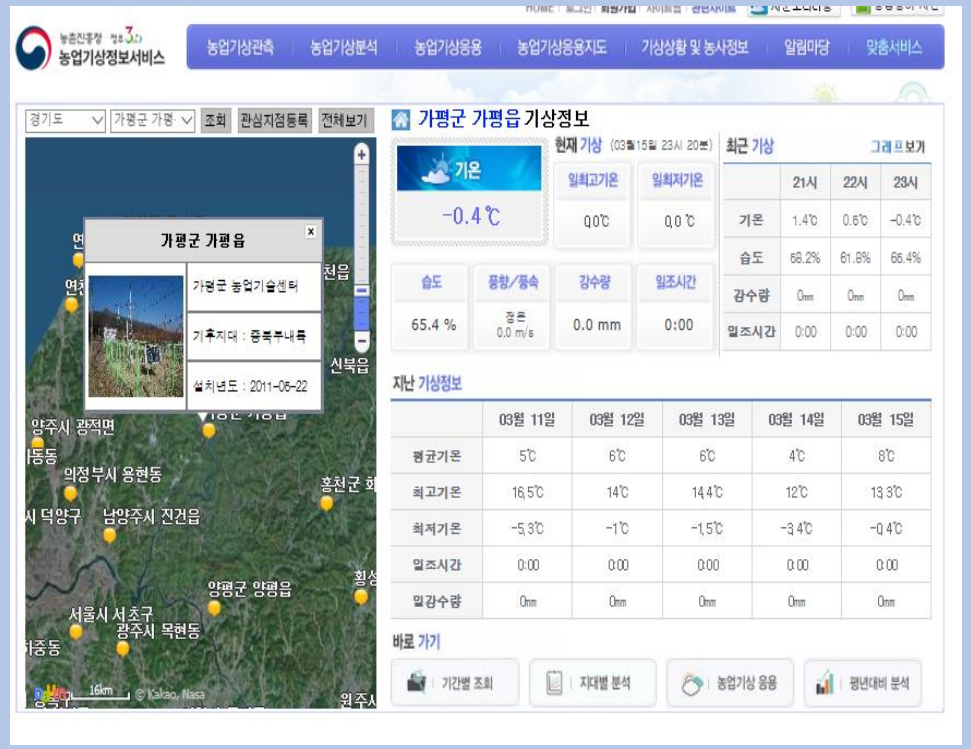
(4) 환경분야 빅데이터 수집방법연구

- KEI 및 유관기관 보유 자료 수집-가공 방안 연구
  - 가공법 및 제공형태(파일, DB, API)의 다양성으로 인한 연구목적 데이터 수집의 어려움을 극복하는 방안을 고민
    - 자료 수집 전용 S/W ELK (ElasticSearch + LogStasch + Kibana) 활용
    - Rvest, Rcurl (R), BeatifulSoap (phython) 등 웹 게시물을 자료 형태로 읽을 수 있는 package 사용

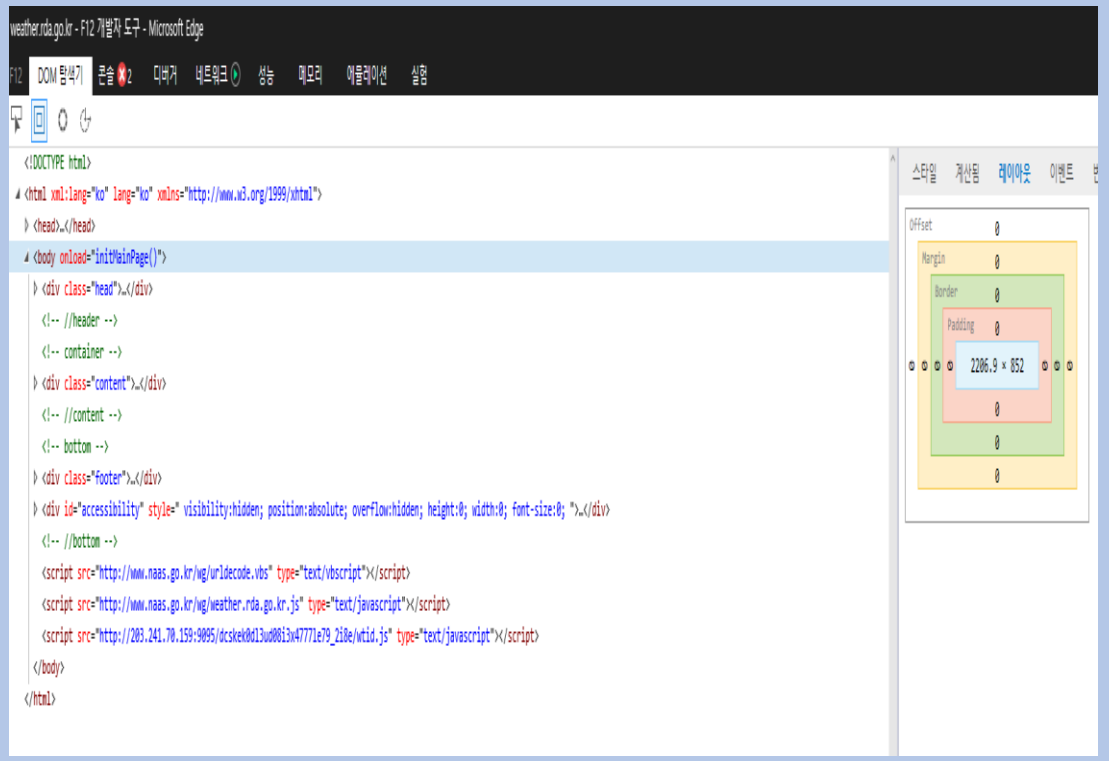
- 기존 자료의 추출-집적 과정 개발 및 결과물을 Open source DBMS로 축적
  - 정적 데이터 사례 1건: 파일 형태
    - 예) 대기오염물 배출(한국환경산업기술원) : .XLS
  - 동적 데이터 사례 1건: 실시간, 주기적 update 자료
    - 예) 농업기상정보시스템 (국립농업과학원): OPEN API
  - 공간정보 기반 웹서비스 데이터 1건 : Web DB 형태
    - 예) 민간서비스(직방, 다방) JSON, HTML, XML

OPEN API 자료 수집

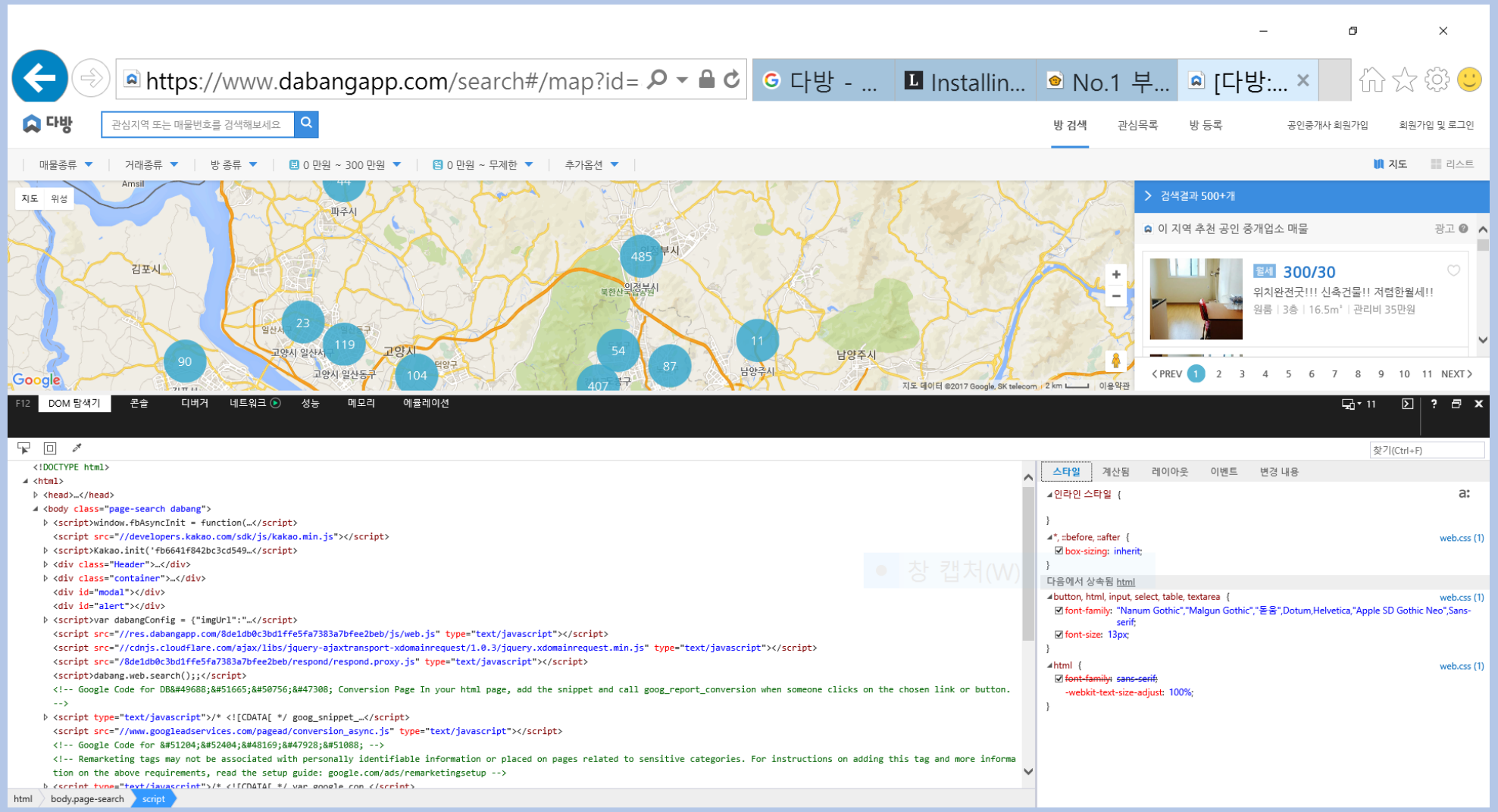
농업기상정보 서비스



농업기상정보 서비스 구조



공간정보 기반 웹서비스 : 직방



4. 사업관리

기간, 인력, 예산

- 기간: 2017년 3월 30일 – 2017년 12월
- 인력: 박사급 연구원 2명(1명 원외), 전문원 1명, 연구 보조인력 3명 투입
  - 박사급 연구원 2명 채용 예정
- 예산: 3억 6백만 원 책정
  - 위탁연구비 4천 만원 책정: '딥러닝을 활용한 환경리스크 예측'
    - 위탁과제 책임자: 한국 산업기술대학교 이동현 교수

연구진 구성

연구진	역할
강성원 연구위원(책임)	- 과제 총괄 - 빅데이터 연구 방법론 활용방안
한국진 전문원	- 환경분야 빅데이터 수집방법연구
이동현 한국산업기술대 교수(위탁)	- 딥러닝을 활용한 환경리스크 예측
강선아 위촉연구원	- 기후변화에 따른 전염성 질병 예측
김도연 위촉연구원	- 텍스트 마이닝 이용 KEI 연구동향 분석
김진형 연구원	- 미세먼지 오염도- 발생요인 패턴 분석 - 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. 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										

연구관리

- 월 2회 Team Seminar 중 1회 세부과제 연구상황 공유
  - 일정 및 결과물을 자문위원진과 실시간으로 공유할 수 있는 온라인 공간 마련
- 월 2회 Machine Learning Study를 실시하여 연구능력을 함양
- Working paper 상태의 중간 산출물을 온라인에 게시하여 피드백 기회를 확대

Team Seminar Plan

연간 계획

2017 년 3 월 3 일 금요일
오전 8:55

1. 목적: Bigdata team 2017 년 연구 진도 관리 /Idea 공유/Brain Storming
2. 일정
   a. 2 월
      i. 2/9 김오석 박사: GIS 소개
      ii. 2/23 배현주 박사: 건강보험 Cohort 자료 소개
   2. 3 월
      a. 3/9 이성호 박사: 인공지능 소개
      b. 3.27-3.31 : Machine Learning Platform 비교 분석
   3. 4 월
      a. 2 주(4.10-4.14): Proposal Seminar 1-김진형, 강선아, 김도연
      b. 4 주(4.24-4.28): Proposal Seminar 2- 강성원, 한국진
   4. 5 월
      a. 2 주(5.8-5.12): Brain Storming 1: 전원
         i. Bigdata 로 KEI 에서 할 수 있는 일
         ii. Bigdata 로 내가 하고 싶은 일
      b. 4 주(5.22-5.26): Progress Report- 전원
   5. 6 월
      a. 2 주(6.12-6.16): 발제 1 - 강성원(CNN)
      b. 4 주(6.26-6.30): Progress Report- 전원
   6. 7 월
      a. 2 주(7.10-7.14): 발제 2 - 김진형
      b. 4 주(7.24-7.28): Progress Report- 전원
   7. 8 월
      a. 2 주(8.7-8.11): 발제 3 - 한국진
      b. 4 주(8.21-8.25): Progress Report- 전원
   8. 9 월
      a. 2 주(9.11-15): Brain Storming 2
         i. Bigdata 로 KEI 에서 할 수 있는 일
         ii. Bigdata 로 내가 하고 싶은 일
      b. 4 주(9.25-9.29): 발제 4-김도연
   9. 10 월 : 휴강(최종보고 준비)
   j. 11 월
      i. 2 주(11.6-11.10): 발제 5-강선아
      ii. 4 주(11.20-11.24): 발제 6-강성원(RNN)
   k. 12 월
      i. 2 주(12.11-12.15): 외부강사 (마지막)

Team Study Plan

Study plan revision

2017 년 3 월 2 일 목요일

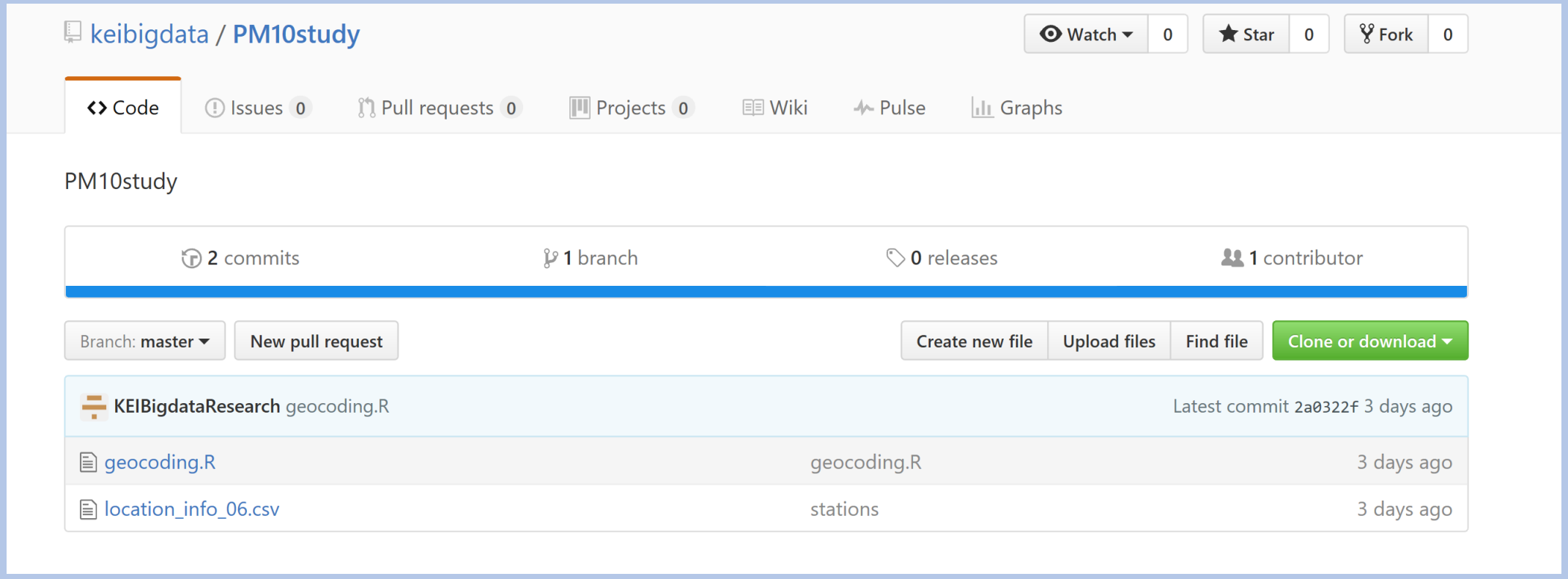
오후 3:34

1. 목적
   a. Machine Learning Technique 배우기: Supervised/Unsupervised/Deep Learning
   b. Machine Learning platform technique 배우기 : SQL, Tensorflow, AWS, BI software
   c. Why? 환경 Bigdata 연구 결과를 만들기 위해서.

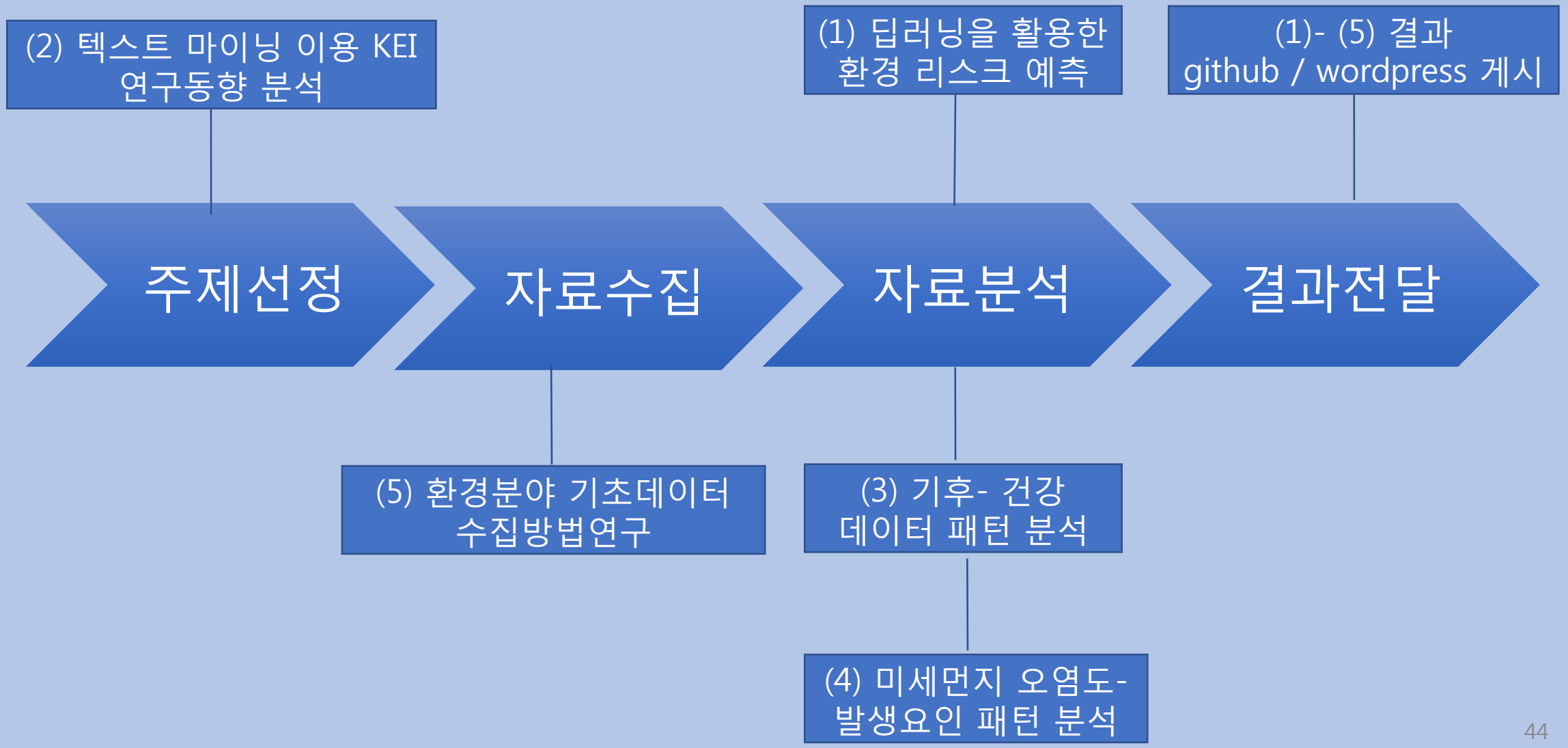
2. 내용
   a. SQL Study (한국진) : Big 'Data'의 'Data'를 저장하는 양식에 대한 선행 학습
   b. Deep Learning 기초(강성원) : 김성필. "딥러닝 첫걸음 머신러닝에서 컨벌루션 신경망까지".
      i. "http://www.kyobobook.co.kr/product/detailViewKor.laf?ejkGb=KOR&mallGb=KOR&barcode=9788968487323&orderClick=LAA&Kc="
      ii. Matlab code 를 R 로 reverse engineering 하는 작업을 함께 진행
   c. Machine Learning 강의 소개 (팀원 전원)
      i. Coursera : Machine Learning Class by Andrew Ng. (강성원)
         https://www.coursera.org/learn/machine-learning/
         1. Matlab code 를 R 로 reverse engineering 하는 작업을 함께 진행
      ii. Fast Campus Machine Learning Course
         1. R Text Mining (김도연)
         2. R Machine Learning (강선아)
         3. Python Programming (한국진)

3. 일정
   a. 3 월
      i. 3 주(3.13-3.17) : SQL
      ii. 4 주(3.20-3.27) : 딥러닝 첫걸음
         1. 1 장: 머신러닝
         2. 2 장 : 신경망
         3. 3 장 : 다층 신경망의 학습
   b. 4 월
      i. 1 주 (4.3-4.7) : 딥러닝 첫걸음
         1. 4 장: 신경망과 분류
      ii. 3 주 (4.17-4.23): 딥러닝 첫걸음
         1. 6 장: 컨벌루션 신경망(CNN)
   c. 5 월: Coursera Machine Learning
      i. 1 주(5.1-5.5): Coursera Machine Learning - Regression
         1. Week1: Introduction
         2. Week 2: Regression with multiple variables
         3. Week 3: Logistic regression
      ii. 3 주(5.15-5.21) Coursera Machine Learning-Neural Network
         1. Week4: Representation
         2. Week5: Learning
   d. 6 월: R text mining (세부내역은 담당자가)
      i. 1 주(6.1-6.9)
      ii. 3 주(6.19-6.24)
   e. 7 월 : R text mining (세부내역은 담당자)
      i. 1 주(7.3-7.7)
      ii. 3 주(7.17-7.21) R text mining (세부내역은 담당자)
   f. 8 월 : R machine learning
      i. 1 주(7.31-8.5) R machine learning(세부내역은 담당자)
      ii. 3 주(8.14-8.18) R machine learning(세부내역은 담당자)
      iii. 5 주(8.28-9.1) R machine learning (세부내역은 담당자)
   g. 9 월 : R machine Learning
      i. 1 주(9.4-9.8) R machine learning (세부내역은 담당자)
      ii. 3 주(9.18-9.23) R machine learning (세부내역은 담당자)
   h. 10 월: Coursera Machine Learning
      i. 1 주(10.2-10.5): 휴식-추석
      ii. 3 주(10.16-10.20) : Coursera Machine Learning -SVM
         1. Week6: Advice for Applying Machine Learning/Machine Learning System Design
         2. Week7: Supporting Vector Machine (SVM)
   i. 11 월: Coursera Machine Learning/ Python Programming
      i. 1 주 (10.31-11.3) Coursera Machine Learning -Clustering/PCA/Anomaly detection
         1. Week8: Unsupervised Learning/Dimension Reduction
         2. Week9: Anomaly Detection (세부내역은 담당자가)
      ii. 3 주(11.13-11.17)
   j. 12 월: Python Programming (세부내역은 담당자가)
      i. 1 주(12.4-12.8)
      ii. 2 주(12.18-12.22)

온라인 게시 (예)


keibigdata / PM10study
Watch 0
Star 0
Fork 0
Code
Issues 0
Pull requests 0
Projects 0
Wiki
Pulse
Graphs
PM10study
2 commits
1 branch
0 releases
1 contributor
Branch: master
New pull request
Create new file
Upload files
Find file
Clone or download
KEIBigdataResearch geocoding.R
Latest commit 2a0322f 3 days ago
geocoding.R
geocoding.R
3 days ago
location_info_06.csv
stations
3 days ago

5. 기대효과


(2) 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
(3) 기후- 건강 데이터 패턴 분석
(4) 미세먼지 오염도- 발생요인 패턴 분석

빅데이터 분석 적용 사례 및 역량 축적

- 다양한 빅데이터 연구 방법론 환경연구 적용 가능성 점검
  - 빅데이터 연구 방법론의 장점인 예측, 패턴 파악 등이 환경 연구· 환경 정책 개발에서 활용될 수 있는지 점검
- 환경 빅데이터 연구 역량 축적
  - 빅데이터 연구 전 단계에 걸쳐 1 건 이상의 연구 실적 획득
  - 3개 수치자료 분석 알고리듬, 1개 텍스트자료 분석 알고리듬 구축
    - 딥러닝, Random Forest, Text mining 알고리듬 각 1개 이상 확보
  - 환경 3개 이상 기초데이터 및 환경 사전 베타 버전 구축
  - 연구 과정 및 결과를 공유하여 민간 연구인력과 교류의 기반을 마련

감사합니다