2018년도 연구계획

: 소셜 빅데이터를 통한 환경이슈에 대한 국민인식 분석

2018.02.28.

김도연

연구주제 : 소셜 빅데이터를 통한 환경이슈에 대한 국민인식 분석

☐ 연구의 필요성 및 목적

○ 환경이슈에 대한 정책이 국민의 의견을 수렴하고 있는 지 여부에 대한 회의 존재

- 국민의견 수렴은 설문지, 전문가 의견, 콜센터 등의 전통적인 방법으로 진행되고 있음
- 기존의 전통적인 방법 **단점**:
  1) 조사 비용 발생
  2) 표본의 수 적음
  3) 정해진 문항만 측정 가능
  4) 사회적 바람직성 등 편향 발생

<법무부>

<교육부>

연구주제 : 소셜 빅데이터를 통한 환경이슈에 대한 국민인식 분석

□ 연구의 필요성 및 목적

○ 실시간으로 생산되는 소셜 빅데이터의 **감성분석(Sentimental Analysis)**을 통해 환경이슈에 대한 국민인식 파악 가능

- 감성분석은 텍스트에 담긴 감성(sentiment), 정서(affect), 주관(subjectivity), 감정(emotion)을 추출하여 점수화하는 기법(Chen and Zimbra, 2010)
- 소셜 미디어의 확산으로 감성분석은 2000년대 이후로 활발히 연구되었음(Liu 2012; Appel et al. 2015)

- "긍정", "부정" 같이 상반되는 극성(polarity)의 분류부터
- "기쁨", "슬픔", "지루함", "흥분", "놀람" 등 다수 범주에 대한 분류까지 텍스트에 내포된 의견이나 감성 등의 의미를 범주화하는 작업

- 본 연구에서는 소셜 미디어에 나타난 환경이슈에 대한 국민들의 반응을 감성분석의 대상으로 활용함
- 환경이슈 관련 국민인식이 다량으로 표출되는 온라인 채널의 텍스트를 기반으로 감성분석을 수행하여 **실시간으로 국민의 니즈 발굴**

연구주제 : 소셜 빅데이터를 통한 환경이슈에 대한 국민인식 분석

□ 연구의 필요성 및 목적

○ 소셜 미디어에 표출되는 환경이슈에 대한 국민인식을 분석하여, 향후 환경정책 수립을 위한 기초자료 제공에 기여하고자 함

- 성공적인 정책운영을 위한 선제적 대응에 본 연구 결과를 활용할 수 있음
- 기존의 전통적인 방식을 보완하는 방법으로 '감성분석을 이용한 국민인식 분석 방법'의 예를 제공

연구주제 : 소셜 빅데이터를 통한 환경이슈에 대한 국민인식 분석

☐ 주요 연구 내용

1) 감성 키워드 분석

· 국민이 가장 부정적으로 체감하는 환경이슈 발굴

· 환경이슈에 대한 세부 감성 키워드 분석

2) 감성 시계열 분석

· 환경이슈 발생 전후 감성 변화 분석

· 기후변화관련 정형 데이터(온도, 습도, 강우량, 강설량, 미세먼지)와 추이 비교분석

3) 매체별, 지역별 비교 분석

· 매체별: SNS(트위터, 페이스북, 인스타그램), 댓글(뉴스 댓글, 블로그 및 카페 댓글)

· 지역: 수집 가능한 위치 데이터를 바탕으로 분석

4) 시각화

연구주제 : 소셜 빅데이터를 통한 환경이슈에 대한 국민인식 분석

□ 연구 추진 방법

① 환경이슈 선정 및 사전 구축
② 데이터 수집
③ 데이터 저장
④ 데이터 전처리
⑤ 환경주제 감성사전 구축
⑥ 감성분석기 개발
⑦ 분석결과 도출

연구주제 : 소셜 빅데이터를 통한 환경이슈에 대한 국민인식 분석

□ 연구 추진 방법

① 환경이슈 선정 및 사전 구축

- 텍스트 데이터 수집(웹 크롤링)을 진행하기에 앞서 불필요한 데이터의 수집을 제거하기 위해 환경이슈 주제에 적합한 키워드 사전구축이 필요함
  · 키워드 사전은 향후 도출되는 분석 결과에 직접적으로 영향을 미치며, 결과의 신뢰성을 확보하기 위해 가장 중요한 작업이라 할 수 있음

연구주제 : 소셜 빅데이터를 통한 환경이슈에 대한 국민인식 분석

□ 연구 추진 방법

① 환경이슈 선정 및 사전 구축

- 환경이슈 사전 구축

· 환경관련 생산문서에서 머신러닝 기법(LDA, Word2Vec)을 이용하여 환경이슈와 근접한 순으로 키워드 추출

* 1) 강성원. (2017). 환경 빅데이터 분석 및 서비스 개발. *일반연구보고서*, 2017

2) 환경부, "환경정책", http://www.me.go.kr/home/web/index.do?menuId=10259, 검색일: 2017.2.26.

· 추출한 키워드를 환경 전문가 집단의 의견을 반영하여 키워드 선정 및 환경이슈 별 키워드 사전 DB 구축

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
...			

연구주제 : 소셜 빅데이터를 통한 환경이슈에 대한 국민인식 분석

□ 연구 추진 방법

② 데이터 수집

- 환경이슈 관련 국민인식이 표출되는 온라인 채널 확인
- 온라인 채널별 이용자 성별 및 연령층 확인 (아래 표 참조 : 2017년 6월 기준, mobiinside)

· SNS(Social Network Service)*와 온라인 댓글**에서 발생하는 구전 데이터 수집

* SNS: Twitter, Facebook, Instagram

** 온라인 댓글: 뉴스 댓글, 카페 댓글

<채널별 사용자 성별 및 연령층 >

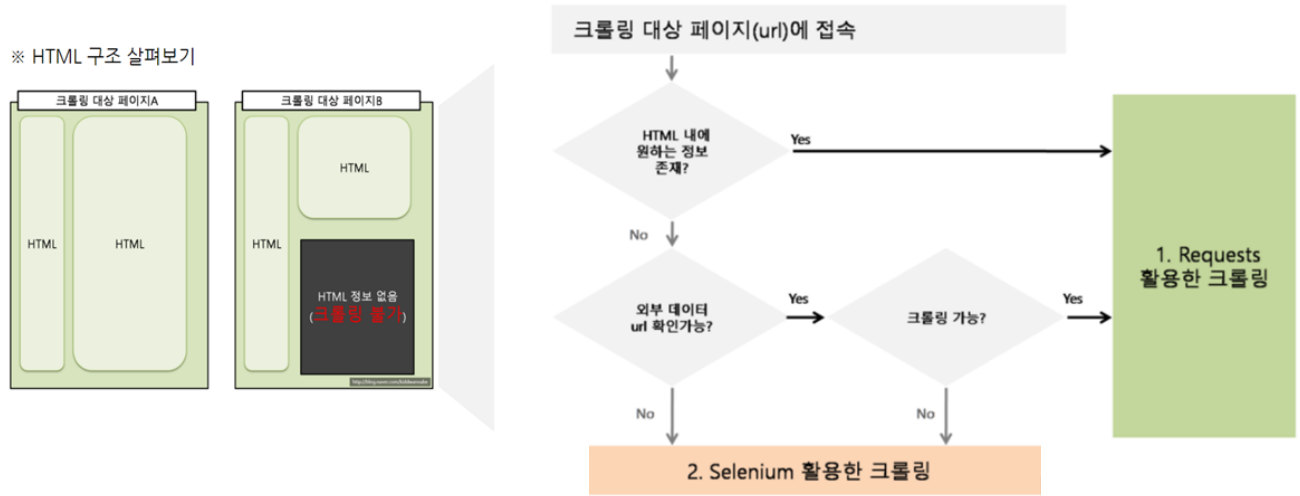
채널	성별 및 연령층
트위터	**20대 여성(21%), 10대 여성(19%)**, 20대 남성(9%), 30대 남성(8%)
페이스북	**20대 남성(18%), 30대 남성(14%)**, 20대 여성(12%), 40대 여성(9%)
인스타그램	**20대 남성(15%), 20대 여성(15%)**, 30대 남성(14%), 30대 여성(13%)
네이버 카페	**30대 여성(20%), 40대 여성(17%)**, 40대 남성(14%), 30대 남성(13%)
다음 카페	**50대 여성(17%), 50대 남성(16%)**, 40대 여성(13%), 40대 남성(13%)

연구주제 : 소셜 빅데이터를 통한 환경이슈에 대한 국민인식 분석

□ 연구 추진 방법

② 데이터 수집

- 환경이슈 키워드 사전기반으로 온라인 텍스트 데이터 수집
  · 빅데이터팀 서버에 웹 크롤링 자동화 구현
- 크롤링 방법
  1) **Requests**를 이용한 접근 방법: Python-Beautiful soup
  2) **Selenium**(Browser controller)을 이용한 접근 방법

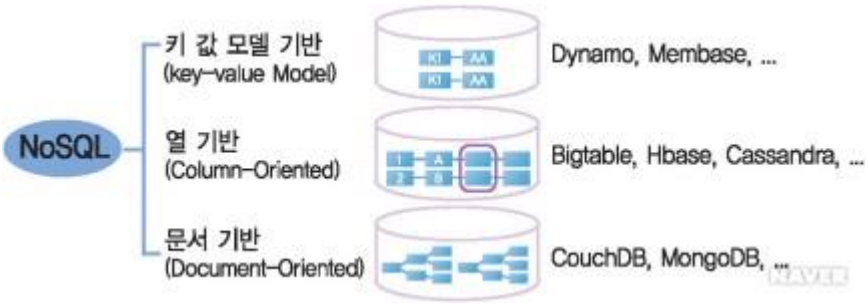


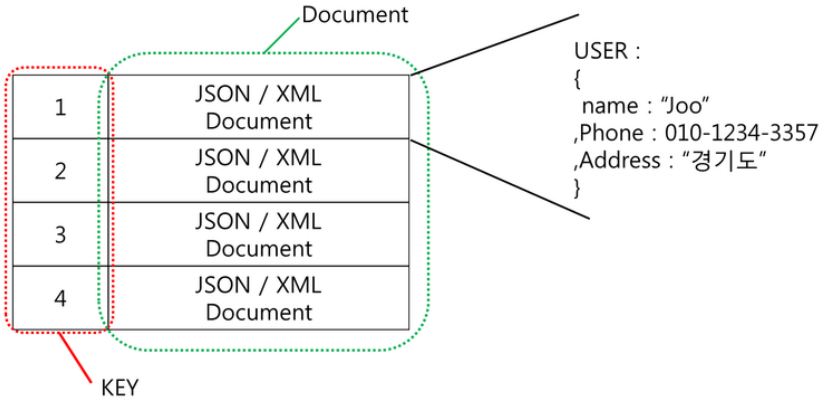
연구주제 : 소셜 빅데이터를 통한 환경이슈에 대한 국민인식 분석

□ 연구 추진 방법

③ 데이터 저장

- 수집한 온라인 텍스트 데이터를 실시간으로 **MongoDB**에 저장
  - · MongoDB: 비정형 데이터를 NoSQL(Not Only SQL) 형태로 저장
  - · Oracle DB: 숫자기반의 데이터를 MySQL 테이블에 저장 및 연산
- 서버에 MongoDB 설치
- 크롤링한 결과를 **JSON** 형태로 변환하여 MongoDB에 저장





Why NoSQL?
-> MySQL은 scheme을 미리 정해주고, 데이터를 넣어야 하는데 크롤링한 데이터가
만약 url, 제목, 본문을 가져오다가 url, 제목, 본문, 날짜 이렇게 가져오면 DB 전체 구조를 migrate 해줘야 하기 때문...

연구주제 : 소셜 빅데이터를 통한 환경이슈에 대한 국민인식 분석

☐ 연구 추진 방법

④ 데이터 전처리

- 광고, 스팸 데이터 삭제
- 중복데이터 삭제
- 특수문자, 특정단어 등 불용어 삭제

연구주제 : 소셜 빅데이터를 통한 환경이슈에 대한 국민인식 분석

□ 연구 추진 방법

⑤ 환경주제 감성사전 구축

- 감성분석을 위해 문장의 긍정, 부정 기준이 되는 감성사전이 필요함
- 기존의 범용감성사전은 환경분야에서 쓰이는 어휘 또는 감정표현을 반영하지 못하므로 정확성이 떨어질 수 있음
- 로버트 푸루칙의 감정 수레바퀴 (Robert Plutchik's Wheel of Emotions) 활용하여 감성사전 구축
  · 64가지 감정 중 환경주제에 적합한 감정 선택
- 6개의 감성 Class가 태깅된 환경주제 감성사전 설계
  · Positive, Negative(Terror, Grief, Rage), Neutral, Objective

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

TER	이제 사계절 내도록 미세먼지와 함께 해야하다니..
GRI	봄의 불청객 ㅋㅋㅋ미세먼지인듯...
GRI	지긋지긋한 미세먼지
RAG	미세 먼지 참 0같네 진짜 —.—
TER	눈코입이 너무 따가운데 미세먼지때문이라고 믿고싶다..
NEU	댓글창 지진났네 쿵쾅쿵쾅
NEU	지각 할까봐 지진이 깨워 줬나 봄더 쿨럭
GRI	쓰나미 지진 와서 일본땅이바닷속으로 가라앉았으면 좋겠다
GRI	단군이 자리 잘못 잡아 나라 세웠네. 홍수와 가뭄을 겪는 땅에
NEU	가뭄난곳이 어디죠? 제 눈물로 단비를 뿌려주겟어여ㅠㅠㅠㅠ
TER	비라도 많이와서 가뭄해갈에 도움됐으면 미세먼지는 쫌 꺼지고
TER	비좀 많이와라..가뭄 심각하다...
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	지구온난화의 문제는 더이상 남의 일이 아닙니다. 조금이라도 환경에 신경써야합니다.

연구주제 : 소셜 빅데이터를 통한 환경이슈에 대한 국민인식 분석

□ 연구 추진 방법

⑥ 감성분석기 개발

- Input Design
  - · 감성사전 데이터를 기반으로 X, Y, W 값 구축
  - · X: Tokens, Y: Target, W: Weight

TER	비좀 많이와라..가뭄 심각하다...

X	비	좀		많	이	와	라	.	.	가	뭄		심	각	하	다	.	.	.	ϕ	ϕ	ϕ	ϕ	ϕ	ϕ	ϕ	ϕ	ϕ
Y	TER																										
W	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0

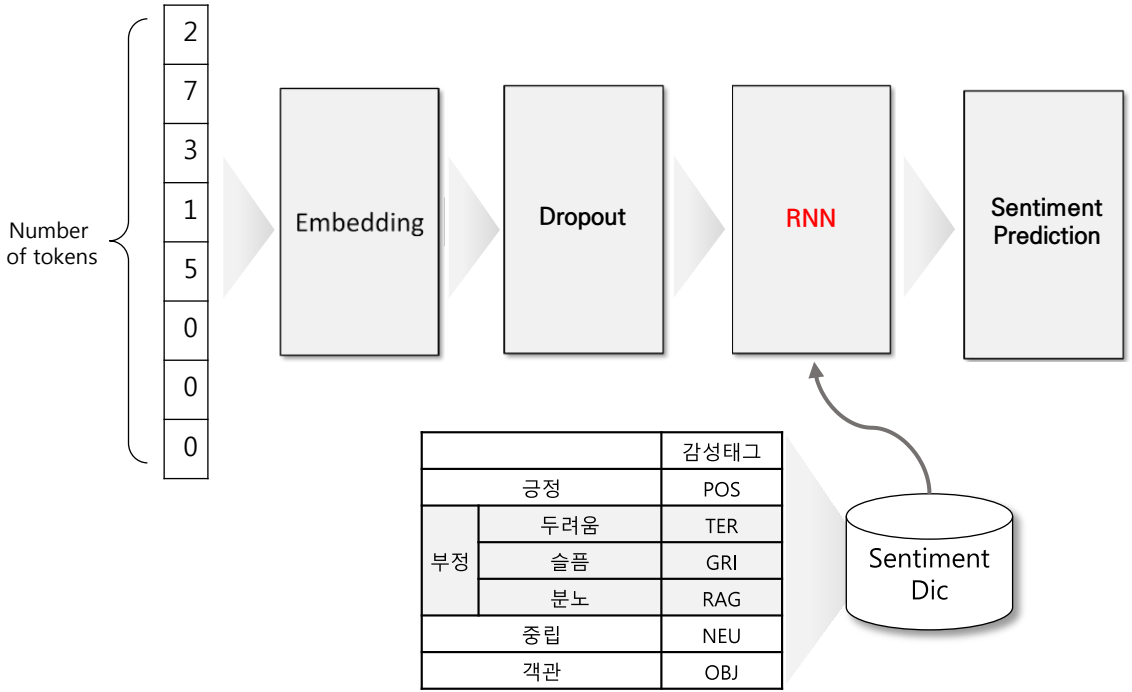
Φ : padding symbol
W : to mark padding positions

연구주제 : 소셜 빅데이터를 통한 환경이슈에 대한 국민인식 분석

□ 연구 추진 방법

⑥ 감성분석기 개발

- Tensorflow를 활용한 '한국어 감성분석기' 구현



연구주제 : 소셜 빅데이터를 통한 환경이슈에 대한 국민인식 분석

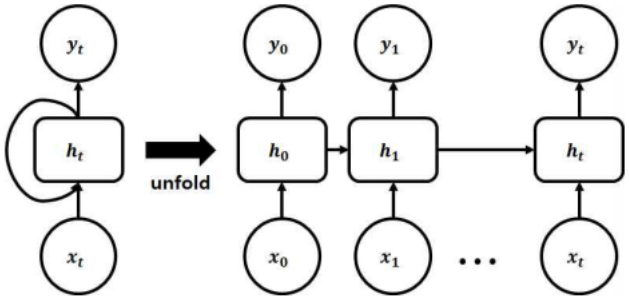
□ 연구 추진 방법

⑥ 감성분석기 개발

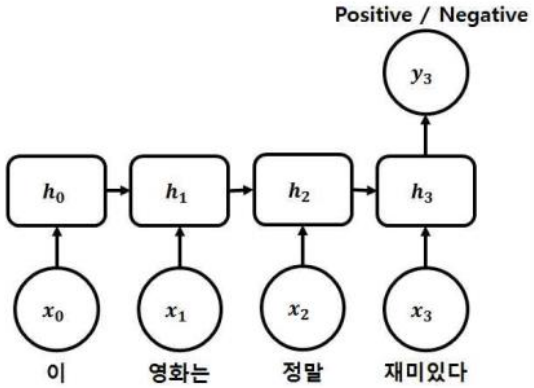
- 순환신경망(RNN, Recurrent Neural Networks)

· 시계열 데이터 학습에 주로 사용됨

· 장기 의존성(long-trerm dependencies)문제: 문장이 길어질 경우 성능이 저하됨.



<순환신경망의 기본구조>



<감성분석에 활용되는 순환신경망 예시>

연구주제 : 소셜 빅데이터를 통한 환경이슈에 대한 국민인식 분석

□ 연구 추진 방법

⑦ 분석결과 도출

1) 감성 키워드 분석

· 국민이 가장 부정적으로 체감하는 환경이슈 발굴

· 환경이슈에 대한 세부 부정 키워드 분석

2) 감성 시계열 분석

· 환경이슈 발생 전후 감성 변화 분석

· 기후변화관련 정형 데이터(온도, 습도, 강우량, 강설량, 미세먼지)와 추이 비교분석

3) 매체별, 지역별 비교 분석

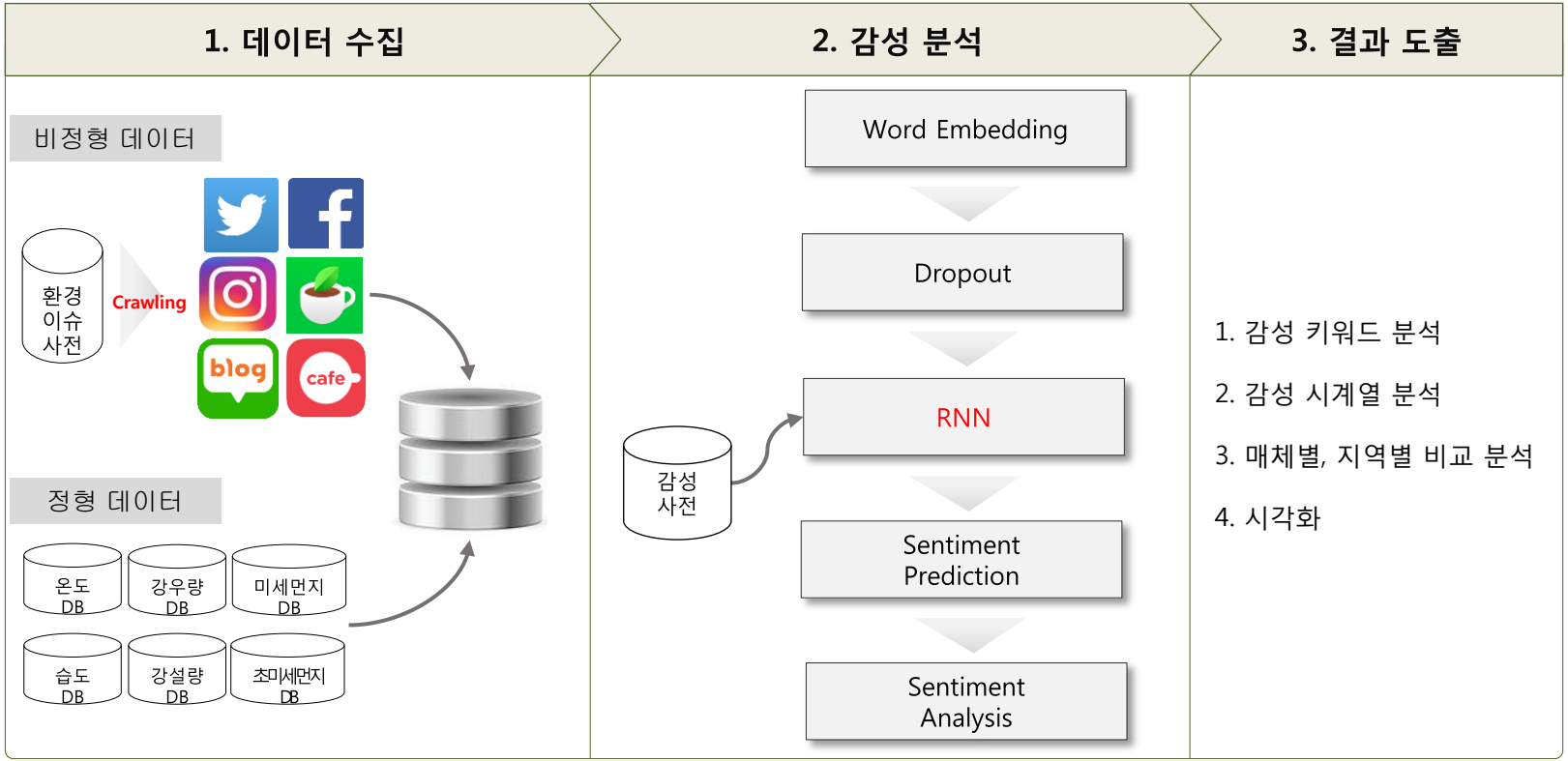
· 매체별: SNS(트위터, 페이스북, 인스타그램), 댓글(뉴스 댓글, 블로그 및 카페 댓글)

· 지역: 수집 가능한 위치 데이터를 바탕으로 분석

4) 시각화

연구주제 : 소셜 빅데이터를 통한 환경이슈에 대한 국민인식 분석

□ 연구 추진 방법



연구주제 : 소셜 빅데이터를 통한 환경이슈에 대한 국민인식 분석

☐ 연구 관리

중간보고 ▼ 6월 / 최종보고 ▼ 10월

구분 \ 월	1	2	3	4	5	6	7	8	9	10	11	12
연구주제 선정	■											
환경이슈사전 구축	■	■										
데이터 수집			■	■	■	■	■					
데이터 전처리						■	■					
감성사전 구축		■	■	■	■	■	■					
감성분석기 개발			■	■	■	■	■					
감성분석							■	■	■	■		
보고서 작성										■	■	■

■ 진행완료 및 보완 예정

■ 진행예정