ISSN 2733-9696(온라인)
ISSN 2733-9572(인쇄본)

GTC BRIEF

2020
DECEMBER
Vol.1 No.2

GTC BRIEF는 기후기술과 관련하여 시의성 있는 현안 및 동향정보를 알기 쉽게 정리한 자료임

1. 장기 저탄소 발전 전략(LEDS)과 탄소 중립 정책 동향
_ 최형식

2. 기후변화대응을 위한 산림 분야 국내 정책 및 R&D 동향 분석
_ 이천환 이은창 안세진 한수현

3. Korea's CTCN pro bono activities: experiences and lessons learned
_ 이원아 양리원

4. 태양전지 분야 연구논문 동향 분석
_ 하수진 오상진

5. LDA 기반의 융·복합 녹색·기후기술 도출 방법
_ 신현우 전은진 오지현 정현덕

GTC 녹색기술센터
GREEN TECHNOLOGY CENTER

ISSUE 05

LDA 기반의 융·복합 녹색·기후기술 도출 방법

신현우, 전은진, 오지현, 정현덕 / 기술총괄부 | hwshin@gtck.re.kr, honeysuckle@gtck.re.kr, jhoh@gtck.re.kr, kate5684@gtck.re.kr

하이라이트

- 기후변화 문제해결형 녹색·기후기술을 도출하기 위한 방안으로 LDA(Latent Dirichlet Allocation) 기반의 토픽 모델링 방법론을 적용
- 특히, 개도국 기술수요가 가장 많은 물, 에너지, 식량 분야들이 연계·통합된 물-에너지, 에너지-식량, 식량-물 부문의 토픽 모델링을 실시하고 기후변화 현안 해결을 위한 융·복합 기술 영역을 도출
- 최근 5년간(2016~2020) 논문과 국·영문 특허의 통합 LDA 분석을 통하여 도출된 융·복합 녹색·기후기술 분야는 '태양광을 이용한 물 이용 설비', '하수 슬러지를 활용한 폐자원 에너지화', '스마트워터시티', 'ICT 기반 물 관리', 및 '바이오매스를 이용한 수소생산'등 이었음

키워드

- LDA, 토픽 모델링, 융·복합 기술, 녹색·기후기술, 최적화 모델

융·복합 녹색·기후기술의 필요성

기후변화는 단일 기술로 해결하기 어려운 대표적인 난제이며, 유망 녹색·기후기술 간 융·복합화를 통하여 한계 극복이 가능

- 기후변화는 인류의 지속가능발전을 위하여 반드시 해결해야 할 도전적 과제이며, 경제적·환경적·사회적 문제가 복잡하게 얽혀 있어서 통합적 관점의 접근이 필요
  - 온실가스 감축과 기후변화 적응을 상호보완적으로 해결하기 위하여 국제사회에서 융·복합 녹색·기후기술에 대한 수요*가 지속적으로 증가

    * UNFCCC CTCN TA(기술지원) 및 GCF 사업에 cross-cutting 융·복합 과제 수요 증가
  - 특히, 글로벌 기후변화의 주요 현안인 물-에너지-식량(WEF) 부족 문제를 함께 해결할 수 있는 융·복합 기술개발이 더욱 중요해짐
  - 기후변화 현안과 연계되어 자원 간 상호연결성이 심화되고 효율적 성과활용이 더욱 강조됨에 따라서 WEF Nexus 부문의 융·복합 녹색·기후기술 도출 필요
- 따라서, 우선적으로 개발이 필요한 융·복합 녹색·기후기술을 도출하기 위하여 각기 다른 주제의 문서들 내 밀접하게 연계되어있는 기술 토픽 모델링 추진

융·복합 기술 도출을 위한 LDA 분석

LDA 방법론

- LDA를 번역하면 '잠재 디리클레 할당'이며, 주어진 문서 내 방대한 비정형 데이터로부터 내재하는 주요 토픽 또는 프레임을 도출할 수 있는 확률적 토픽 모델링 기법[1)]
  - (의미망 분석) 텍스트와 텍스트 사이의 관계를 나타내는 텍스트 마이닝은 텍스트 간 연결성 및 네트워크 구조를 분석하는 '의미망 분석(SNA, Semantic Network Analysis)'이며, LDA 분석은 의미망 분석의 확장된 개념
  - (토픽 모델링) 텍스트 데이터에서 사용되는 단어들의 빈도를 통계적으로 분석하여 문서 집합 속에 포함되어있는 잠재적인 주제나 의미 구조 등을 자동으로 추출하는 분석 방법[2)]

- 대용량의 문서들로부터 주요 토픽을 자동으로 찾아내기 위한 LDA 알고리즘[3)]은 최초 학습을 시작할 때, 전체 문서에 토픽을 무작위로 1차 배정한 후 토픽의 재할당을 반복 수행하여 문서와 단어에 가장 적절한 토픽을 찾아가는 과정
  - (LDA 확률 산식) 문서에 포함된 모든 단어 수(*V*)와 토픽수(*K*) 지정되었을 때, *d* 번째 문서의 *i* 번째 단어에 해당하는 토픽 $z_{d,i}$가 *j* 번째에 할당될 확률(*p*)는 다음 수식과 같음[4)]

$$p(z_{d,i}=j|z_{-i},w)=\frac{n_{d,k}+\alpha_j}{\sum_{i=1}^{K}(n_{d,i}+\alpha_i)}\times\frac{v_{k,w_{d,n}}+\beta_{w_{d,n}}}{\sum_{j=1}^{V}(v_{k,j}+\beta_j)}=AB$$

$n_{d,k}=k$번째 토픽에 할당된 d번째 문서의 단어빈도
$v_{k,w_{d,n}}=$ 전체 텍스트에서 k번째 토픽에 할당된 단어$w_{d,n}$의 빈도
$w_{d,n}=d$번째 문서에 n번째로 등장한 단어
$\alpha=$ 문서의 토픽분포 생성을 위한 디리클레분포 $parameter$
$\beta=$ 토픽의 단어분포 생성을 위한 디리클레분포 $parameter$
$A=d$번째 문서가 k번째 토픽과 연계되어있는 연관성 정도
$B=d$번째 문서의 n번째 단어$(w_{d,n})$가 k번째 토픽과 연계되어있는 연관성정도

  - (토픽 분포도) 사용자가 초기에 설정한 토픽의 개수에 따라 시각화된 토픽분포도 결과를 얻을 수 있는데, IDM(Intertopic Distance Map)은 토픽간 거리를 통해 주제 관련성을 나타내고, 막대그래프는 토픽별 가장 빈번하게 나타나는 상위 단어를 나타냄
  - (주요 토픽) IDM 토픽 분포도 평면에 위치한 원의 크기는 각각의 토픽들이 전체 문서에서 차지하는 비율, 즉 빈도수에 비례하므로 원의 크기가 클수록 주요 토픽이라고 할 수 있음
  - (토픽 가중치 λ 설정) Sievert & Shirley(2014)[5)]는 단어 간 연관성(*r*)을 나타내는 변수를 도입하고 다음의 식과 같이 λ라는 가중치 parameter와의 관계를 통하여 토픽의 분별력을 설정

※ λ값이 1에 가까울수록 토픽별 가장 빈번하게 사용되는 단어들로 구성되고, 0에 가까울수록 해당 토픽이 다른 토픽들과 구분되면서 상대적으로 빈도수가 낮은 단어들로 구성

$$r(w,k|\lambda) = \lambda \log(\phi_{kw}) + (1-\lambda)\log(\frac{\phi_{kw}}{p_w}),\ 0 \le \lambda \le 1$$

r = 단어간 연관성을 나타내며, 가중치 λ에 따라서 가변
λ = 단어의 가중치로 0과 1사이의 값
ϕ = 단어 w가 토픽 k에 해당할 확률
p_w = 단어 w가 해당 $corpus$(말뭉치)내 나타나는 확률

WEF 분야의 논문과 특허에 대한 LDA 적용 및 분석

물, 에너지, 식량 부문의 주요 토픽 도출을 위한 LDA 적용

- (분석 소프트웨어) Python 3.7[6] 기반의 오픈소스를 사용하여 텍스트마이닝을 수행하고, 토픽모델링 LDA는 Gensim[7], 시각화는 pyLDAvis를 활용

- (분석 대상 텍스트) 녹색기술센터 기후기술 분류체계 중 물(소분류27~30), 신재생에너지(소분류 4~13, 41)* 및 식량 부문(소분류 23~26, 34, 38)에 해당하는 기술들과 관련된 영문 논문과 등록 특허의 텍스트마이닝 결과에 대한 LDA 분석

  * 신재생에너지 중심 에너지전환 등의 국제동향을 고려하여 온실가스 감축 기술 중 재생에너지·신에너지 및 신재생에너지 하이브리드 기술로 한정

  - (대상기간) 2016년도 1월 1일 ~ 2020년도 6월 30일

  - (대상 논문) Web of Science의 database로부터 수집된 WEF 관련 논문

[표 1] 텍스트마이닝 및 LDA분석에 사용된 WEF 분야별 논문 개수

에너지	논문 수	식량	논문 수	물	논문 수
4. 수력	300	23. 유전자원 유전개량	16,189	27. 수계 수생태계	33,786
5. 태양광	17,166				
6. 태양열	3,061	24. 작물 재배 생산	16,431		
7. 지열	17,938			28. 수자원 확보 및 공급	12,490
8. 풍력	18,091	25. 가축 질병 관리	20,654		
9. 해양에너지	25,322				
10. 바이오에너지	11,607	26. 가공 저장 유통	10,985	29. 수처리	3,648
11. 폐기물	23,222				
12. 수소제조	3,343	34. 수산자원	3,894		
13. 연료전지	8,446			30. 수재해 관리	17,556
41. 신재생에너지 하이브리드	16,398	38. 산림 생산 증진	9,975		
총계	**144,894**	**총계**	**78,128**	**총계**	**67,480**

※DB 출처: GTC 기술총괄부 기후기술 수준조사

  - (대상 특허) WINTELIPS database로부터 수집된 WEF 관련 등록특허

[표 2] 텍스트마이닝 및 LDA분석에 사용된 WEF 분야별 특허 개수

에너지	특허 수		식량	특허 수		물	특허 수	
	국문	영문		국문	영문		국문	영문
4. 수력	239	1,553	23. 유전자원 유전 개량	321	660	27. 수계 생태계	111	1,068
5. 태양광	849	3,672						
6. 태양열	345	677	24. 작물 재배 생산	916	2,292			
7. 지열	82	1,235				28. 수자원 확보 및 공급	3,636	6,313
8. 풍력	598	3,432	25. 가축 질병 관리	346	1,097			
9. 해양 에너지	408	2,744						
10. 바이오 에너지	138	839	26. 가공 저장 유통	550	4,196	29. 수처리	596	3,321
11. 폐기물	1,037	1,369						
12. 수소제조	498	1,442	34. 수산 자원	618	3,435			
13. 연료전지	458	1,497				30. 수재해 관리	362	1,090
41. 신재생 에너지 하이브리드	1,132	2,884	38. 산림 생산 증진	555	906			
총계	**5,784**	**21,344**	**총계**	**3,306**	**12,586**	**총계**	**4,705**	**11,792**

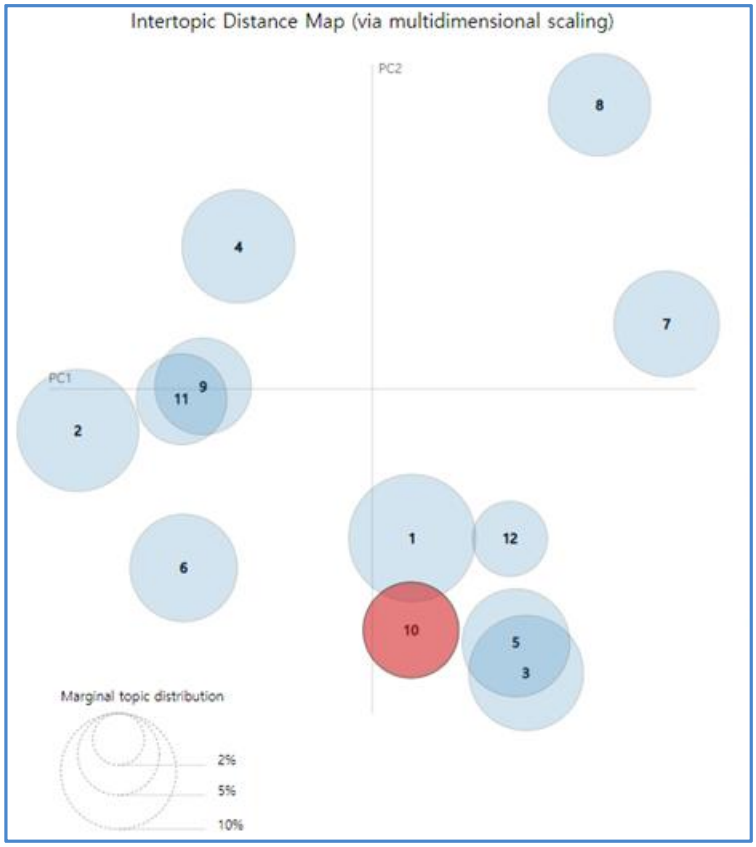
※DB 출처: GTC 기술총괄부 기후기술 수준조사

- (주요 토픽과 가중치) 토픽간 관련성과 비중을 나타내는 IDM과 연관된 상위 텍스트 분석시 λ값은 Sivert & Shirley(2014)의 실험 최적치인 0.6을 중심으로 활용

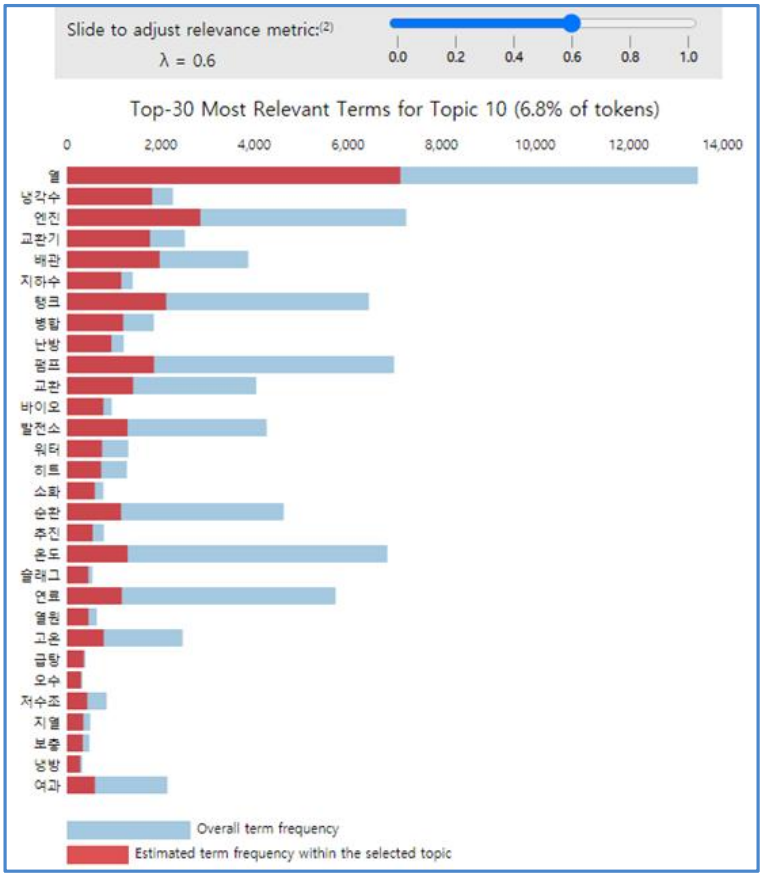
  - 본 토픽 연구의 결과에 λ값을 0~1 사이에서 변화시켰을 때에도 λ값이 0.6 근처에서 최적치*를 나타냄

    * GTC 내부 연구진이 최적치(λ=0.6)를 기준으로 1차 해석안을 도출한 후, 전문가 심층 분석 단계에서 λ값 변경에 따른 주요 키워드 변화를 관찰하여 토픽 해석을 확정

[그림 1] WEF 특허 관련 토픽 10개에 대한 IDM 그래프 예



[그림 2] 가중치 λ=0.6일 때, 그림 1의 토픽10과 연관된 주요 토픽 빈도수 그래프 예



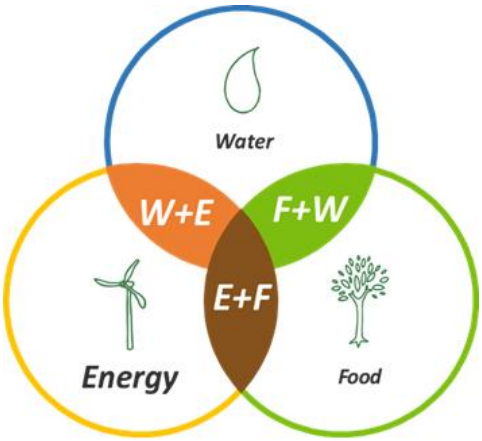
WEF Nexus 부문의 융·복합 녹색·기후기술 및 모델 도출

- (2개 부문 통합·연계 분석) W-E, E-F, F-W 연계 부문별 LDA 분석을 실시하여 융·복합 녹색·기후기술 영역 조사

  - 추출된 토픽 자체에서 융·복합 기술에 해당하는지 여부를 검토*하고, 서로 다른 분야에 속하는 토픽의 IDM이 가깝게 나올 경우 융·복합 여지가 높은 것으로 추정

    * ICT 등 공통기반기술을 포함하여 융·복합 여부를 주로 검토

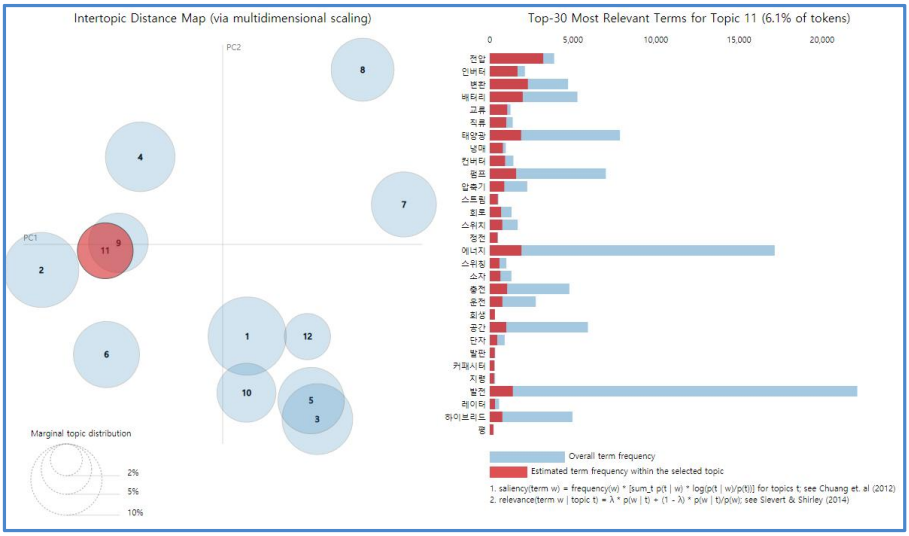
[그림 3] 물-에너지-식량 연계(WEF Nexus) 부문의 융·복합 기술 영역



- (물-에너지 부문 연계분석 사례) 국문 특허내 물과 에너지 분야의 텍스트를 통합하여 토픽모델링 분석을 수행하고 두분야 간 융·복합 기술 영역이 존재하는지 조사

  - (융·복합 기술 도출) 주요 키워드 중 전압, 인버터, 배터리, 태양광, 컨버터, 펌프, 압축기 스트림 등의 빈도 수가 높아서 전문가 심층 분석시, '태양광을 이용한 물 이용 설비 기술'로 도출

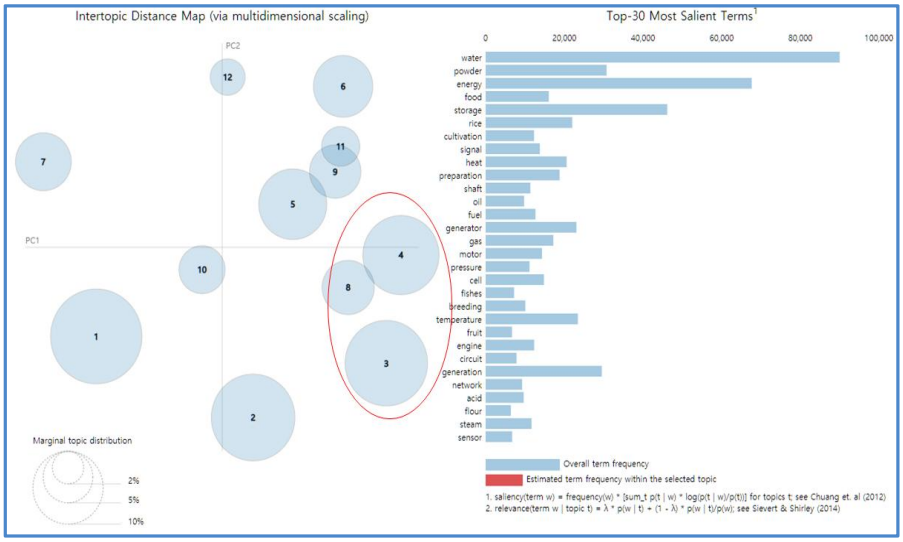
[그림 4] 물-에너지 분야의 국문 특허 텍스트마이닝 LDA 분석 결과(토픽 11)



• (에너지-식량 부문 연계분석 사례) 영문 특허 내 에너지와 식량 분야의 텍스트를 통합하여 토픽모델링 분석을 수행하고 두분야 간 융·복합 기술 영역이 존재하는지 조사

- (융·복합 기술 도출) λ=0.6일 때 토픽 1~5의 주요 영문 키워드는 [표 3]과 같으며, 토픽 3 및 4 영역에서 도출된 에너지와 식량이 연계된 융·복합 기술은 '바이오매스를 활용한 수소생산 기술'분야로 유추됨

[그림 5] 에너지-식량 분야의 영문 특허 텍스트마이닝 LDA 분석 결과



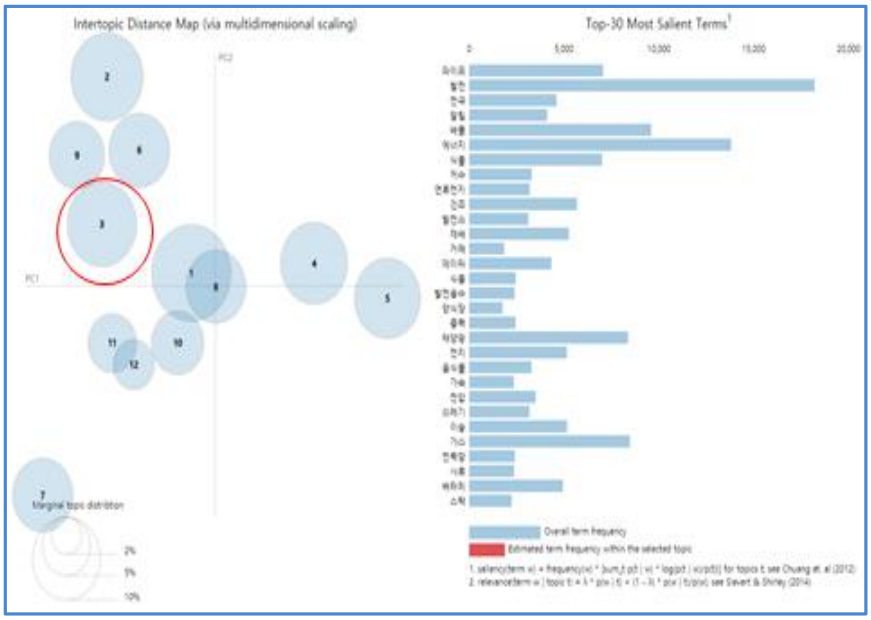
[표 3] 에너지-식량 부문 영문특허 내 각 토픽의 비중과 주요 키워드(λ=0.6)

토픽	비중 (%)	λ=0.6 주요 키워드	토픽해석
1	16.9	rice, preparation, cell, compound, salt, seed, protein, production, corn, membrane, growth, sodium, stack, vitamin, calcium, metal, bacteria, cake, yeast, fermentation, additive, slice	간척지 작물재배
2	14.2	water, breeding, tank, feeding, treatment, container, channel, fertilizer, sterilization, medicine, seeds, moisture, pump, peel, irrigation, shrimp, drainage, manure, injection	불명확
3	13.8	energy, storage, fuel, gas, heat, steam, generation, hydrogen, combustion, turbine, waste, exchanger, compressor, health, micro-grid, pump, generator, greenhouse, exchange, electricity, oxygen, soil medium, biomass	바이오매스를 이용한 수소생산
4	11.8	network, plant, energy, generation, converter, grid, distribution, management, inverter, battery, station, photovoltaic, storage, optimization, monitoring, electricity, response, capacity, population, interface, prediction	에너지/ 전력변환
5	9.4	block, wall, rod, cylinder, cavity, panel, shell, barrel, centimeters, spring, seat, ball, connecting, blocks, survival, frames, sleeve, connection, screw, assembly	장치관련

• (식량-물 부문 연계분석 사례) 국문 특허 내 식량과 물 분야의 텍스트를 통합하여 토픽모델링 분석을 수행하고 두분야 간 융·복합 기술 영역이 존재하는지 조사

- (융·복합 기술 도출) λ=0.6일 때 토픽 1~5의 주요 영문 키워드는 [표 4]와 같으며, 토픽 2의 '발전소 폐온수의 영농 활용'과 토픽 3 영역의 'ICT 기반 물관리'기술이 융·복합 기술 영역으로 추정

[그림 6] 식량-물 분야의 국문 특허 텍스트마이닝 LDA 분석 결과



[표 4] 식량-물 부문 국문 특허 내 각 토픽의 비중과 주요 키워드(λ=0.6)

토픽	비중(%)	λ=0.6 주요 키워드	토픽해석
1	16.2	펌프, 탱크, 열, 수위, 밸브, 정수, 물, 우량, 수원, 교환기, 수조, 센서, 순환, 전극, 냉수, 배관, 교환, 온수, 지하수, 냉매, 정수기, 목표, 유량, 온도, 응축, 필터, 출수, 전기, 히트	정수설비
2	11.9	폐수, 데크, 칼슘, 과수원, 틸리스, 가공, 해양, 황칠, 혼합, 분리, 원자력, 수력, 저온, 발전소, 분배, 온수, 이중, 유체, 수산물, 가열, 원자로, 증기, 조절, 출현, 진공, 구조, 발전기, 건설	발전소 폐온수의 영농 활용
3	11.3	데이터, 선정, 관리, 영상, 컴퓨터, 서버, 수신, 예측, 계획, 전송, 통신, 하천, 이미지, 신호, 자연환경, 기상, 모듈, 수질, 강우량, 촬영, 원격, 감시, 감지, 무선, 데이터베이스, 현장, 센서, 계측	ICT 기반 물관리
4	10.8	절곡, 수액, 입, 공원, 개부, 온도, 여과, 타, 평행, 끼움, 물고기, 물, 금속재, 양식장, 결합, 삽, 수질, 걸림돌, 프린팅, 보호, 수처리, 측정, 부력, 어류, 버블, 공간, 조절, 수조, 종묘, 이탈	불명확
5	10.4	레트, 급부, 공기, 포켓, 오존, 배출, 멍, 댐퍼, 결합, 적재, 흡입, 임펠러, 양방향, 분사, 모터, 피스톤, 후방, 압축, 상하, 배관, 산소, 직수, 밸브, 덕트, 공간, 염소, 대칭, 하천수, 노즐, 흡입구	불명확

결론

융·복합 기술 도출을 위한 LDA 활용 확대

- 대량의 비정형화된 기후기술 데이터의 텍스트마이닝 LDA 분석을 활용하여 기후변화 현안 해결을 위한 융·복합 기술 도출
  - 통계적 모델 방법론을 활용하여 빈도수가 높은 텍스트군 집합에서 잠재적인 유망 기술 분야 및 주제의 발굴에 활용
  - 특히, 기후변화 현안 키워드 도출 및 워드클라우드 분석에서부터 기후변화 현안 해결을 위한 단일 및 융·복합 기술 도출에 이르기까지 체계적인 문제해결형 기후기술 도출 방법론 마련

- 다양하고 복잡한 기후변화 대응을 위하여 기술수요가 가장 많은 물, 에너지, 식량 분야별 유망 기술들을 1차 선정하고 유망 기술들이 연계·통합된 물-에너지, 에너지-식량, 식량-물 부문의 토픽 모델링을 실시하여 융·복합 기술 영역을 도출하고 시각화
  - 파리협정 이후의 논문과 특허를 대상으로 물-에너지-식량 부문의 텍스트마이닝 LDA 분석 결과, '태양광을 이용한 물 이용 설비', 'ICT 기반 물 관리' 및 '바이오매스를 이용한 수소생산'등 이 융·복합 기술로 도출

- 기후변화문제 해결을 위한 기술트리와 도출된 융복합 기술들을 상호연계하여 궁극적이고 통합적인 녹색기후기술 모델 도출이 가능
  - (예시 1) 태양광 기술을 이용한 물 이용 설비 부문은 태양광 부문과 지하수 확보 분야간 매칭을 통하여 청정에너지 부족과 물 수요/공급의 불균형을 동시 해결
  - (예시 2) ICT 기반의 물관리 부문은 물부족 문제 해결-재난재해 예측-효율적 저류시설을 통합하는 스마트워터시티 모델과 밀접한 연계성을 나타냄
  - (예시 3) 바이오매스를 이용한 수소 생산 부문은 음식물을 포함한 유기성 폐기물을 활용하여 그린수소 생산 적용 필요성을 제시

시사점

- LDA 토픽 모델링은 확률에 기반한 분석결과가 비교적 직관적이고 정교한 편이나, 최적의 토픽 수 선정과 다양한 주제의 내용을 혼용하여 발생하는 토픽의 중복 문제를 보완할 필요가 있음

- 텍스트마이닝 빈도 분석과 텍스트 간 토픽 모델링을 통하여 유망 융·복합 기술 분야의 도출은 가능하였으나 해당 모델을 구성하는 세부기술항목 도출을 위해서는 추가 분석이 필요
  - IDM을 통하여 토픽 간 관련성과 주제를 정확하게 해석하기 위해서는 LDA 변수 최적화에 대한 추가 연구가 필요하며, 특히 최적 가중치(λ) 선정을 위한 체계 연구도 중요

- LDA로부터 도출된 기술 토픽과 모델을 실질적으로 적용하기 위해서는 해당 융·복합 분야만을 대상으로 한 기술 인벤토리 구축이 필요

• 전체 녹색·기후기술 특허 및 논문을 대상으로 한 통합 LDA 분석을 위해서는 방대한 양의 데이터가 예상되므로 이를 효율적으로 처리할 수 있는 방안 검토

- 특히, WEF Nexus 이외의 타 분야 기술 및 산업과의 융·복합화를 고려하기 위해서는 지능형 데이터 플랫폼 구축을 통한 통합 모델링 검토 필요

참고문헌

1) 미디어 빅데이터 연구소(2020), 텍스트마이닝의 시각화, 토픽모델링 분석과 활용 : https://brunch.co.kr/@bflysoft1117/199

2) Ramage et al (2009), "Topic Modeling for the Social Sciences", RAND Journal of Economics NIPS Workshop on Application for Topic Models : R=Text and Beyond

3) https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation#inference

4) https://ratsgo.github.io/from%20frequency%20to%20semantics/2017/06/01/LDA/

5) Sievert and Shirley (2014), "LDAvis: A method for visualizing and interpreting topics", Proceedings of the Workshop on Interactive Language Learning, Visualization and Interfaces, p.63-70

6) Python : https://pypi.python.org/pypi/lda

7) Gensim : https://radimrehurek.com/gensim/

본 내용은 녹색기술센터(GTC)의 주요사업(신현우, 이구용, 전은진, 오지현, 신종석, 정현덕, 「문제해결형 융·복합 녹색·기후기술 도출 및 적용을 위한 전략연구」)으로 수행되었던 내용의 일부를 요약·정리한 것입니다.