기관-2020-020

인공지능 기반 국내외 바이오헬스 기술개발 동향 비교분석 연구

(AI-based comparative analysis of domestic and overseas bio-health technology development trends)

유 거 송

KISTEP 한국과학기술기획평가원
Korea Institute of S&T Evaluation and Planning

• 연구진

- 연구책임자	유거송 (한국과학기술기획평가원 부연구위원)
- 참여연구원	김한해 (한국과학기술기획평가원 부연구위원)
	여창민 (한국과학기술기획평가원 연구원)
	여은주 (한국과학기술기획평가원 연구원)
	김은정 (한국과학기술기획평가원 연구위원)
	김주원 (한국과학기술기획평가원 연구위원)
	한지아 (한국과학기술기획평가원 연구원)
	조해주 (한국과학기술기획평가원 연구원)
	손미림 (한국과학기술기획평가원 연구원)

기관 2020-020

인공지능 기반 국내외 바이오헬스 기술개발 동향 비교분석 연구
(연구기간 : 2020.4.1.~2020.12.31)

• 발행인 : 김상선

• 발행처 : 한국과학기술기획평가원
(27740) 충청북도 음성군 맹동면 원중로 1339
Tel) 043-750-2300 Fax) 043-750-2680

• https://www.kistep.re.kr

• 인쇄 : 주식회사 동진문화사

목 차

표 목 차

그림목차

제1장 서 론

제1절 연구의 필요성

제2절 연구의 목표 및 내용

제3절 연구 추진전략 및 방법

제1절 연구의 필요성

□ 바이오헬스 분야의 비약적 발전으로 R&D의 규모가 지속적으로 확대됨에 따라, 정책수립 및 투자의사결정을 위한 동향분석의 중요성 및 복잡도가 증가하고 있음

○ 유전자가위, 의료인공지능 등 판도를 바꾸는 기술의 등장으로 바이오헬스에 대한 관심이 증가하고 있으며, 정부는 BIG3 분야 중 하나로 바이오헬스를 선정하여 향후 예산을 크게 확대할 예정임

※ 생명 · 보건의료 분야 정부R&D 투자 : ('17) 2.6조원 → ('25) 4조원 이상[1)]

○ 대규모 R&D예산에 대한 정책적 의사결정을 위한 객관적 근거의 중요성이 증대되고 있으며, 과거와 현재의 트렌드 파악 및 미래 예측을 위해 방대한 데이터를 효과적으로 분석할 수 있는 방법론이 요구됨

- 바이오헬스 연구규모의 확대에 따라 연구과제, 논문 · 특허 등 문헌자료의 양이 방대해짐에 따라, 기술개발 동향분석을 위한 시간과 비용 역시 증가하고 있음

□ 최근 빠르게 발전하고 있는 인공지능 기술은 이러한 정책적 수요에 대응할 수 있는 방법으로 제시되고 있으며, 동 연구진은 인공지능 기반의 지능형 R&D정보데이터 분석시스템을 개발하고 활용기법을 연구한 바 있음[2)]

○ 인공지능(기계학습) 기반 자연어처리 기술의 발전으로 연구개발 관련 빅데이터를 효과적으로 분석할 수 있는 방법론들이 제시되고 있음

- 일반적인 통계분석을 넘어, 비정형데이터(텍스트)로 작성된 대량의 정보들을 정량화하여 해석하고, 의사결정에 활용할 수 있는 기법들이 빠르게 발전하고 있음

1) 관계부처 합동(2019), 바이오헬스 산업 혁신전략
2) 유거송 외(2019), 바이오 · 의료분야 지능형 연구개발정보데이터 분석시스템의 예산배분 · 조정 활용기법 연구, 한국과학기술기획평가원

○ 지능형 R&D정보데이터 분석시스템은 이러한 기법(알고리즘)들을 도입한 R&D 동향분석 도구로서, 방대한 양의 정부R&D 과제 및 논문 데이터를 자동으로 분류하고 의미를 추출할 수 있음

※ KISTEP 업무의 수요에 입각하여 설계되었으며, '19년 예산배분 · 조정 활용기법 연구를 수행함

□ R&D 정책수립 및 투자의사결정의 객관적 근거 확보를 위해서는 이러한 도구를 적극 활용하여 국내외 연구개발 현황을 분석하고 글로벌 트렌드를 포착할 필요

○ 특히 해외 정부R&D의 경우 기존의 언론 또는 인터넷 검색을 통해 조사 가능한 일부 대형사업(프로그램)에 한정된 조사에서 진화된 동향분석이 요구됨

○ 과제 단위의 해외 정부부처 R&D 정보를 수집하고 국내 조사·분석 과제데이터와 연계한 비교분석을 수행하여 시사점을 도출할 필요

- 국내 조사·분석 데이터 기반 분석과 상응한 수준의 분석을 해외 R&D에도 적용함으로써 보다 확실한 비교 기준을 수립할 필요

○ 동 연구에서는 지능형 R&D정보데이터 분석시스템*을 이용하여 이러한 작업을 수행하고, KISTEP 내 활용 활성화를 위한 온라인화 작업을 추진하고자 함

* 이후 "분석시스템"으로 약칭

○ 분석시스템에 내재된 자연어처리, 클러스터링 등의 알고리즘 등을 이용하여 국내외 R&D동향을 객관적으로 비교하고 투자의사결정의 근거로 제시하고자 함

- 이러한 분석 방법론 확립은 향후 타 기술분야의 동향분석에 적용할 수 있으며, 향후 분석시스템 고도화를 위한 기초자료로도 활용 가능

○ 분석시스템은 원내 다양한 업무에 활용될 잠재력이 높으며, 이용 활성화를 위해 현재 오프라인 PC*에서만 사용 가능한 분석시스템을 온라인 사용이 가능하도록 전환할 필요가 있음

* 현재는 생명기초사업센터에 비치된 인터넷에 연결되지 않은 PC에서만 사용 가능함

제2절 연구의 목표 및 내용

□ 본 연구과제는 지능형 R&D정보분석시스템의 온라인화와 분석시스템에 탑재된 알고리즘을 해외 부처 R&D 데이터에 적용하여 국내외 바이오헬스 분야 연구개발 동향을 비교해보는 두 가지의 연구로 구성됨

❍ (분석시스템 온라인화) 기존에 오프라인 PC에서만 사용 가능하도록 개발되어있는 분석시스템을 KISTEP 원내 서버로 이관하여 직원들이 자유롭게 접속하여 사용할 수 있도록 조치하는 작업

❍ (국내외 바이오헬스 연구개발 동향 비교) 분석시스템 상에서 현재 국가연구개발과제 조사분석 데이터(NTIS)를 대상으로 적용되어있는 분석 알고리즘을 해외 부처*를 대상으로 동일하게 적용하여 국내외 연구개발 동향을 비교

* 미국 NIH, 미국 NSF, 영국 UKRI

□ 분석시스템 온라인화 작업은 KISTEP에서 운영하는 과학기술연구개발 정보 포털인 K2Base*와 연계하여 다양한 부서에서 활용이 가능하게 함으로써, 분석시스템에 대한 피드백 및 개선 수요의 지속 발굴을 촉진하고자 함

* http://www.k2base.re.kr

❍ K2Base에 원내 직원 ID로 로그인할 시, 지능형 분석시스템으로 접속할 수 있는 링크가 노출될 예정

※ 2021년 상반기 중 서비스 개시를 목표로 하고 있으며, K2Base의 'KISTEP 업무지원' 메뉴 안에 포함 예정

❍ 동 시스템은 KISTEP 내 서버에서 자체 운영할 계획이며, 높은 컴퓨팅 자원 사용량 등을 고려하여 내부에서 시범 운영하고 외부 공개는 추후 고려

- 높은 메모리 사용량 등으로 인해 동시 접속자 수 제한 등의 운영 방안 마련 및 최적화 등의 유지보수 작업이 지속적으로 필요

○ 분석시스템의 원내 활용을 활성화하기 위해 원내에서 인터넷 접속이 가능한 형태로 분석시스템을 이관하고 향후 지속적 고도화 및 유지보수가 가능한 체계를 마련

□ 국내외 바이오헬스 연구개발 동향 비교는 웹크롤링 등을 이용해 바이오헬스 분야 해외 정부R&D 과제정보 데이터를 수집하고, 국내 정부R&D 조사·분석 데이터와 연계하여 분석시스템 기반 인공지능(머신러닝) 분석을 수행

○ 현재 기존에 개발되어 있는 PubMed 데이터베이스 검색 기능은 논문 초록을 검색한다는 측면에서 국내 연구개발과제와 동등하게 비교하기가 어렵고, 논문은 과제의 산출물이라는 점에서 더욱 동일 선상에서 비교하기에 한계가 있음

- 논문은 일반적으로 연구과제에 비하여 범위가 좁고, 과제 수행 중 또는 종료 이후에 출판되므로 논문이 검색될 시에는 이미 연구과제 수행이 이루어졌다는 의미로 받아들일 수 있음

○ 우리나라가 NTIS를 통하여 국가연구개발과제 정보를 공개하는 것과 유사하게, 해외 일부 부처들도 인터넷 사이트를 통하여 부처에서 지원하고 있는 연구개발과제 정보를 공개하고 있음

○ 해외 연구과제의 분석이 가능하다면 국내 연구과제와 보다 동등한 비교가 가능하므로 국내와 해외의 연구개발 동향을 보다 정확하게 분석할 수 있음

- 기존에 주로 사업(프로그램) 단위로 거시적으로 분석하던 해외 R&D 동향에 대한 상세한 파악이 가능할 것으로 기대함

○ 본 연구진은 조사 결과 연구과제명, 요약문(abstract) 및 연구비 데이터를 제공하는 3개 부처(NIH, NSF, UKRI)를 대상으로 데이터를 수집하여 현재 분석시스템에 적용되어있는 분석 방법론을 동일하게 적용하였음

- 수집된 데이터는 향후 기술적 검토를 통해 분석시스템에 탑재하여 원내 구성원이 상시적으로 이용할 수 있도록 추진하고자 함

❍ 국내와 해외 데이터에 동일한 검색어를 입력하여 출력되는 과제들을 군집화(클러스터링)하여 투자현황 등을 비교하는 방법론을 제시하고자 함

- 국내와 해외의 정부R&D 과제에 해당 분석기법을 적용하여 바이오헬스 분야의 세부분야별 연도별 투자현황, 주제 분포 등을 비교분석하고, 이에 기반한 시사점을 도출하고자 함

제3절 연구 추진전략 및 방법

□ 전문가 자문, 강의 수강 등을 통해 바이오헬스 R&D 국내외 비교분석 연구를 수행하고, 혁신정보분석센터 및 정보관리팀과 협조하여 분석 시스템 온라인화 추진(표 1)

❍ 외국 정부R&D 데이터 수집·분석을 위해 데이터분석, 웹크롤링 등 해당 분야 전문가 자문, 관련강의 및 자료를 활용하고, 분석결과 해석 및 검증을 위한 바이오헬스 분야 전문가 자문을 추진

❍ 분석시스템 온라인화 개발작업은 혁신정보분석센터의 협조 하에 K2Base 용역개발 사업과 연계하여 추진하며, 원내 전산장비 자원을 활용할 예정

- 향후 K2Base의 기능으로 편입하여 지속적인 기능 고도화 및 유지보수를 가능하게 하고, KISTEP 구성원이 안정적으로 업무에 활용할 수 있도록 논의

※ 추후 원내 정보관리운영위원회 등에서 논의 진행

❍ 분석시스템의 향후 개선 및 고도화를 위한 자연어처리 최신 동향 관련 전문가 강의 및 외부교육을 수강하고, 해외 부처 대상 자연어처리 알고리즘 적용 시 기술자문 추진

- 동 시스템에 적용된 doc2vec 알고리즘 이후 발표된 최신 언어모델 알고리즘에 대해 학습하고 동 시스템에의 고도화 방안을 도출
- 해외 부처 데이터를 동 시스템에 직접 연동할 수 없으므로, 별도 프로그래밍을 위한 기술자문 추진

〈표 1-1〉 연구과제 추진체계

목표	지능형 R&D정보데이터 분석시스템 활용 활성화	
내용	**국내외 바이오헬스 R&D 비교분석** (분석시스템 활용 연구)	**분석시스템 온라인화** (KISTEP 구성원 활용도 제고)
수행 주체별 역할	**동 연구진** • 웹크롤링 활용 해외 정부 R&D 과제데이터 수집 • 분석시스템을 활용한 국내외 R&D 현황 분석 수행	**동 연구진** • 분석시스템 온라인화 용역 추진을 위한 RFP 작성 및 개발 관리
		↑ (협력) ↓ • 향후 K2Base로 분석 시스템 편입을 위한 인터페이스 · 기능 보완 및 고도화 관련 아이디어 발굴
	외부 전문가 • 웹크롤링, 데이터셋 구축 관련 자문 • 분석결과 검증(해외R&D 관련 정성적 의견 제시) • 분석시스템 개선 관련 자문 및 의견 제시	**혁신정보분석센터** • 용역발주 관련 행정 지원
		외부 개발진 • 분석시스템의 서버 이관(Migration) 및 적용 작업 • 테스트 및 수정(Debug)
		정보관리팀 • 원내 온라인 활용을 위한 전산장비(서버) 자원 제공

제2장 자연어처리 최신 동향

제1절 총론

제2절 전이학습 개요

제3절 딥러닝 기반 언어모델 및 연구 동향

제4절 텍스트 클러스터링 활용

제1절 총론[3)]

□ 동 연구과제에서 개발한 지능형 분석시스템의 기반이 되는 자연어처리 기술의 최신 동향 조사를 통해 내용의 이해를 돕고자 함

○ 지능형 분석시스템은 doc2vec이라는 텍스트임베딩 기술에 대부분의 기능을 의존하고 있으며, 본 장에서는 관련된 언어모델, 다른 텍스트임베딩 알고리즘 등에 대해 상세히 알아보고자 함

□ 전자 문서 작성이 용이해지면서 많은 수의 텍스트 문서들이 생산되고 있는데, 이를 효과적으로 관리하여 목적에 맞게 사용하는 것에 대한 수요가 증대

○ 문서를 효과적으로 다루는 기법으로는 텍스트 정보 검색(Information Retrieval), 텍스트 분류(classification), 텍스트 기반 정보 추출(Information Extraction), 텍스트 군집화(clustering) 등이 있음.

○ 텍스트 군집화란 많은 수의 문서들을 유사한 문서 집합으로 군집화하는 것을 말함

- 미리 주제를 정하고 이에 맞춰 문서를 분류하는 텍스트 분류와 다르게 주제를 미리 알 수 없는 경우 유용함
- 텍스트 분류는 학습을 위해 많은 양의 학습 데이터가 필요하지만, 미리 주제를 알 수 없는 텍스트 군집화는 학습 데이터가 필요 없음
- 학습 데이터 구축에 대한 시간과 비용은 절감할 수 있지만, 수집된 문서 집합을 몇 개의 군집으로 나눠서 만족한 결과를 얻을 수 있을지에 대해서는 많은 노력이 필요한 것도 사실임.

□ 텍스트 문서를 잘 다루기 위해서는 문서에 대한 이해를 기반으로 하여야 하는데 해당하는 기술이 자연어처리(Natural Language Processing)이며 최근에는 언어지능이라 불리움

3) 서면자문 : 임수종(한국전자통신연구원 언어지능연구소 책임연구원)

❍ 자연어 처리 기술은 크게 자연어 분석과 자연어 생성으로 구분할 수 있으며, 그동안은 분석 기술에 초점이 맞춰져 있었으며 인공 지능 기술의 발전으로 인해 점차 자연어 생성에 대한 연구도 증대되고 있음.

❍ 여러 분야에 딥러닝이 도입되어 성능이 비약적으로 개선된 것처럼 자연어 처리 분야에도 활발하게 딥러닝이 도입되고 있음

- 자연언어는 한국어, 영어, 중국어와 같이 고유한 성격을 갖고 있어서 각 언어별로 다양한 특성이 반영되어야 하기 때문에 기존 방법으로는 사용가능한 수준의 기술을 개발하는 것이 힘듦
- 딥러닝으로 인해 전이학습 등 기존 방법을 뛰어넘을 수 있는 다양한 방법이 제시됨
- 모든 데이터를 벡터 공간으로 사상(mapping) 하는 딥러닝 특성으로 인해 각 언어간 특성이 희석되어 언어 독립적인 알고리즘 개발이 용이해짐

❍ 이러한 기술 변화 흐름에 맞춰 기존 자연어 처리라는 용어 대신 딥러닝을 이용한 언어지능(Language Intelligence)이라는 용어로 대체되는 추세임

- 인공지능을 사람의 단위 지능에 매핑하기 위해 시각, 청각, 언어 지능으로 구분하여 각각의 기술 분야를 분류함.

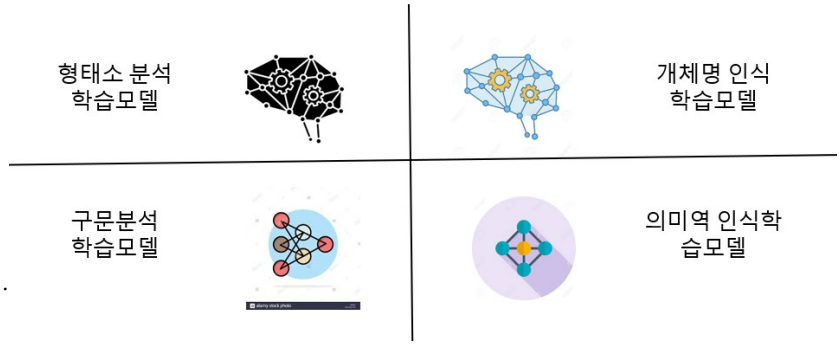
  ※ 시각 지능은 사물의 종류, 위치, 동작 등을 이해하고, 청각 지능은 소리를 이해하며, 언어 지능은 글을 읽고 이해하고 말하는 것을 뜻함

- 기술적 용어로 언어지능이란 단어 및 주변 문맥을 신경망을 이용하여 표현하고 학습하여 언어 이해 태스크를 수행하거나 언어를 생성하는 기술

❍ 언어지능 기술에 딥러닝이 적용된 초창기에는 시각지능, 청각 지능에 적용된 기술이 그대로 적용되었으나 한계점이 노출됨

- 형태소 분석, 개체명 인식, 구문분석과 같은 전형적인 태스크와 기계독해, 질의응답 문제를 풀기 위해 각 태스크에 적합한 다양한 딥러닝 모델(Feed Forward Network, RNN, CNN, Bi-directional LSTM)이 연구되고 적용
- 초기 딥러닝 기법은 전통적인 기계학습에 비해 좋은 성능을 보였지만 다양한 언어지능 태스크에 특화된 알고리즘과 학습 데이터가 필요하여 한계 상황에 다다름

- 전통적인 기계학습은 각 task 별로 서로 다른 학습데이터, 서로 다른 알고리즘, 서로 다른 학습 모델이 생성되었지만(그림 2-1), 초기 디러닝 기법은 전통적인 기계학습에 비해 공통된 알고리즘으로 서로 다른 task에 적용 가능하다는 점만 개선됨.



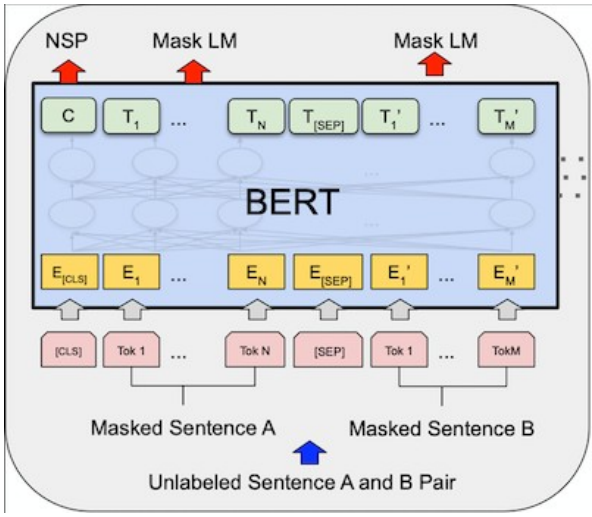
[그림 2-1] 각 task에 적용된 전통적인 기계학습 결과

□ 초기 딥러닝이 방법의 한계를 극복하기 위해서 사전학습(pre-training) 을 이용한 언어 모델 기반의 전이 학습 방법이 주목을 받기 시작함.

❍ Word2Vec 과 같은 딥러닝 방법을 적용한 언어 모델은 초기에는 성능 개선이 미미하였음

❍ 딥러닝 방법이 고도화됨에 따라 이를 이용한 언어 모델 구축 방법도 개선되었고, 기존 방법을 개선한 Doc2Vect(Google, 2014), Global Word Vectors(Glove, 스탠포드대, 2014), FastText(facebook, 2017), GPT-2 등에 이어 구글에서 발표한 BERT는 기존 방법에 비해 큰 성능 향상을 보임

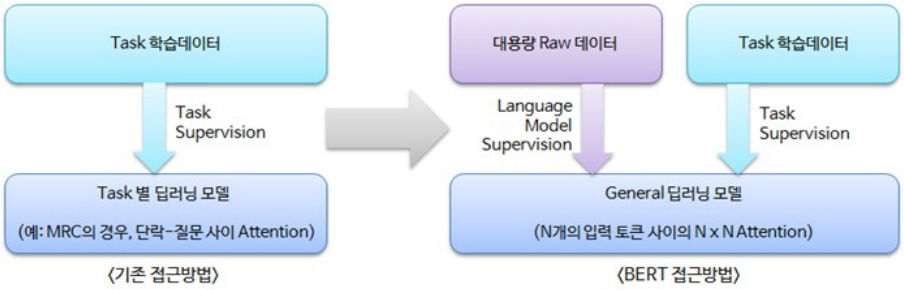
- 구글이 2018년 11월에 공개한 BERT는 양방향 기반의 어휘 표현을 학습하여 (그림 2-2) 다양한 태스크에서 종래 기술보다 크게 우수한 성능을 보임에 따라 많은 후속 연구가 BERT를 기반으로 연구됨



[그림 2-2] BERT 동작 개념

❍ 사전 학습을 통해 언어 모델을 구축하고 이를 각 응용 태스크(ex. 기계독해, 번역, 챗봇 등)에서 fine-tuning 하는 방법이 언어 지능 기술의 새로운 파라다임을 제시함

- 기존 접근 방법은 각 task 별로 많은 양의 학습 데이터를 구축하고 task에 가장 적합한 알고리즘 방법을 이용하여 독립적으로 학습
- 사전학습 기반의 전이 학습은 수집이 용이한 대용량 데이터를 이용하여 공통적인 부분을 미리 학습한 후에 상대적으로 소량의 task 학습 데이터를 이용하여 미세조정하는 방식을 택함(그림 2-3)



[그림 2-3] 기존 딥러닝 학습 방법과 전이학습 방법 개념 비교

- 성능 기준선과 성능은 개선된 전이학습 기법에 의해 상향조정 되고 있으며 새롭게 제시되는 방법은 성능 기준선에 기존에 도달했던 시간을 단축하며 도달하거나 이를 능가하게 됨
  ※ 성능 기준선은 human base line으로 인간이 직접 특정 task를 수행했을 때의 성능을 말함
- 텍스트 언어인식 AI 벤치마킹 툴인 GLUE(General Language Understanding Evaluation) 및 후속 버전인 SuperGLUE의 다양한 자연어 태스크에 대해 쉽게 인간 수준인 성능 기준선에 도달함

제2절 전이학습 개요

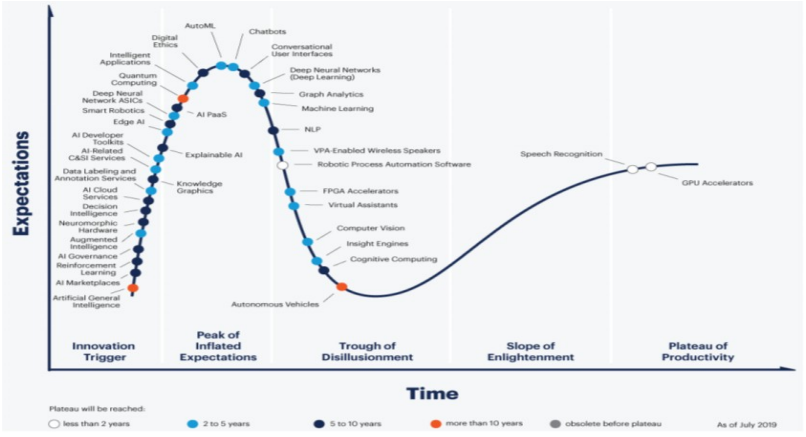
□ 전이학습(Transfer Learning)이란 학습 데이터가 부족한 분야의 인공지능 모델을 구축을 위해 상대적으로 데이터가 많으며 비슷한 분야에서 사전 훈련된 모델을 이용하는 기계학습 기법

○ 딥러닝 시대에 많은 알고리즘이 공개되어 성능 결정은 주로 학습 데이터의 양과 질에 따라 좌우됨

- 학습 데이터를 각 태스크 별로 구축하는 것은 수집, 구축, 품질 관리등 시간과 비용의 문제와 연관됨.

○ 딥러닝 기반 알고리즘들이 대부분 빅데이터와 많은 양의 컴퓨팅 파워가 필요하기 때문에 실제 적용에 장벽이 존재하기 때문에 적은 데이터와 좀더 적은 자원으로 효율성을 중시하는 기술에 대한 관심이 증대됨

- 이러한 환경에 대응하기 위한 기술 중 하나가 전이학습으로 한 분야에서 학습한 결과를 학습한 적이 없는 다른 분야에 적용하는 개념임
- IT 전문 리서치 기관인 가트너(Gartner)는 2019년 새롭게 부상하는 기술 중 하나로 전이학습을 언급하였으며, 하나의 문제 또는 데이터셋에서 학습한 것을 다른 영역에 적용하는 기술로 소개함(그림 2-4)

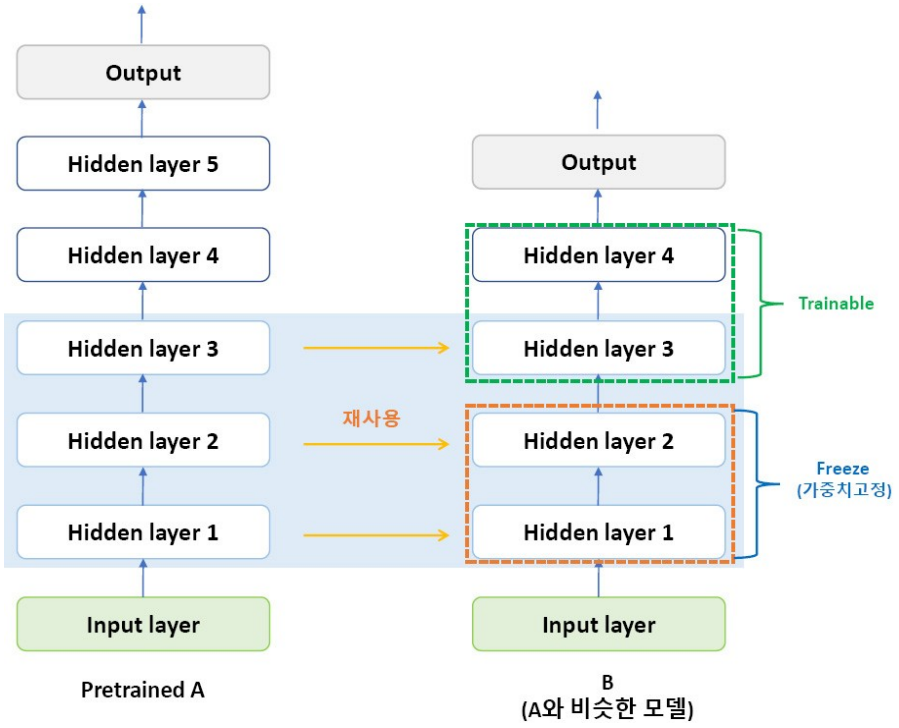


출처 : Forbes, “What’s New In Gartner’s Hype Cycle For AI” 2019. 9. 25.

[그림 2-4] 2019년에 새롭게 부상할 기술(가트너)

❍ 새로운 태스크에 대한 인공지능 모델을 구축할 때 비슷하며 잘 동작하는 기존 모델의 일부를 재사용하며 이 과정을 미세조정(fine-tuning) 이라고 함(그림 2-5)

- 많은 학습 데이터와 파라미터를 사용하여 딥러닝을 학습할 때 처음부터 새롭게 학습을 하면 학습 속도가 매우 느린데, 이럴때는 기존에 비슷하게 학습된 모델을 사용하여 하위 계층은 재사용하는 것이 훨씬 효율적임
- 아래 그림에서 기존 A에서 은닉층 1, 2는 그대로 사용하고, 새로운 태스크 B를 위해서 은닉층 3, 4는 새롭게 학습해야 함
- 태스크 B를 위해서 은닉층 1, 2, 3, 4를 모두 학습하지 않고, 3, 4만을 학습하기 때문에 필요한 데이터와 컴퓨팅 파워가 상대적으로 적게 필요함



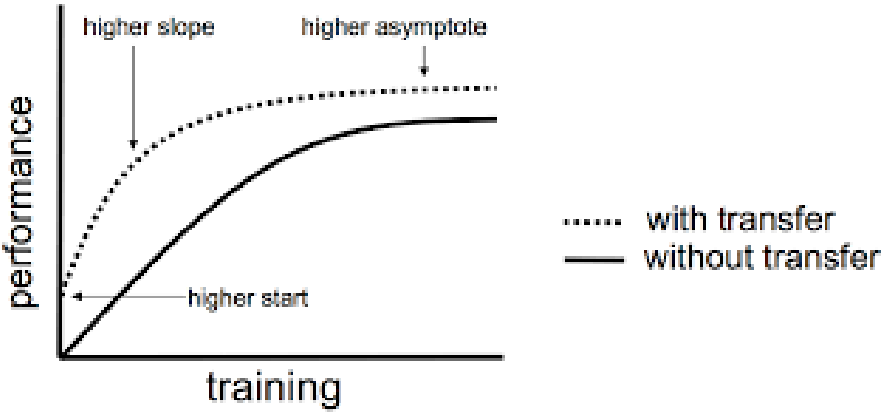
출처 : Hands-On Machine Learning with Scikit-Learn and Tensorflow, Chapter 11, 2017)

[그림 2-5] 전이학습의 개념

- 미세조정(fine-tuning)은 feature extraction, pre-trained model을 모델 구조로 사용, 다른 은닉층을 고정시키고 일부분 층을 학습하는 방법이 있음

□ 전이학습은 다른 영역에서 미리 학습된 모델의 효과적인 활용이 가장 중요하며, 기존 모델을 재사용할 수 있기 때문에 다른 학습 방법론에 비해 초기 학습속도가 빠른 장점이 있음

○ 기존 방법에 비해 전이 학습은 학습의 향상도가 좀더 가파르게 상승할 뿐 아니라, 좀더 나은 성능을 얻을 수 있음(그림 2-6)



출처 : Handbook of Research on Machine Learning Applications and Trends, Chapter 11

[그림 2-6] 전이학습 사용 시 기존 방법과 학습 속도와 성능 비교

□ 전이학습은 딥러닝이 적용된 여러 분야에서 시도되고 있으며, 자율주행차 분야나 언어지능 분야에서는 의미있는 진전을 보이고 있음

○ 전이학습 모델이 활발히 적용되고 있는 대표적인 분야는 자율주행차 분야

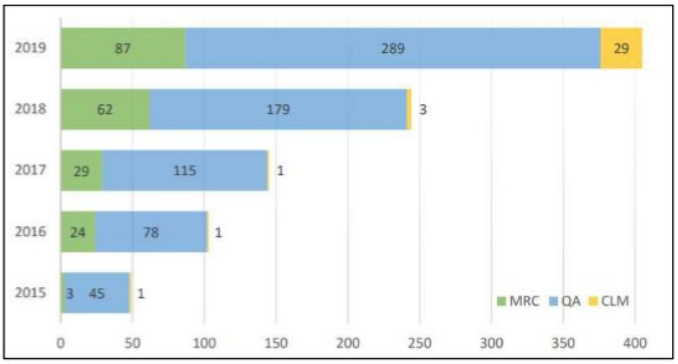
- 충분히 학습이 되지 않은 상태의 자율주행 자동차가 학습을 위해 실제 도로에서 주행하는 것은 주변에 위험을 끼치므로, 사전에 최대한 시뮬레이션 환경에서 학습한 후 이 변수들을 이용하여 현실 데이터를 이용해 전이학습을 수행하는 것이 더 안전함

- 도메인 전이학습(Domain Transfer Learning)을 활용한 현대자동차의 자율주행 기술 연구는 실생활이 아닌 인터넷 사이트나 영상물(영화 등)에서 관련 데이터를 수집하고 사전 모델 학습에 활용하고, 무단횡단이나 역주행 등 실제 도로에서 발생할 수 있는 돌발상황에 관해서 따로 학습

❍ 자율주행차 분야만큼이나 데이터 수집 등 여러 난관이 있는 자연어처리 분야도 전이 학습을 도입하여 성능 개선을 보이고 있음

- 자연어처리의 경우 특정 언어에 우수한 기술을 개발하더라도 한국어, 영어, 중국어와 같이 다양한 특성을 갖는 모든 언어에 적용하는 것은 쉽지 않기 때문에 일반적으로 언어 전반적인 학습 모델 기반 위에 특정 언어, 특정 태스크에 대한 미세조정만으로 원하는 기술을 개발하는 방향을 선호하게 됨
- 최근에는 BERT나 GPT3와 같이 자연어처리에도 사전 훈련된 딥러닝 언어모델을 이용하여 전이학습을 활발히 적용

※ BERT 이후 딥러닝 사전학습 모델 연구는 Arxiv 공개 기준 '18년 3편에서 '19년 29편으로 비약적으로 증가함(그림 2-7)



<Arxiv에 공개된 딥러닝 사전학습 언어모델 논문 출판 동향 (출처>

[그림 2-7] Arxiv에 공개된 딥러닝 사전학습 언어모델 논문 출판 동향

- 어떠한 방식으로 사용될지 모르는 상태에서도, 미리 해당 언어의 방대한 양의 텍스트를 수집하고 이를 이용하여 언어에 대한 사전 지식으로 구성된 모델을 구축

- 구축된 해당 언어 모델을 사용하여 특정 목적에 맞게 사용하기 위해서는 미세조정 과정과 해당 목적에 맞는 상대적으로 소량의 학습 데이터만 이용하면 됨

❍ 전학습은 만능은 아니며 적용하려는 범위에 따라 유용하지 않을 수 있으나 인공지능 도입을 위한 비용과 시간 문제를 해결할 수 있는 현실적인 대안으로 각광받고 있음

제3절 딥러닝 기반 언어모델 및 연구 동향

□ 언어 모델이란 한국어, 영어, 중국어와 같은 일반적인 자연 언어라는 현상을 모델링하고자 단어 열(sequence)이 발생할 확률 분포로 구성된 모델을 말함.

❍ n개의 단어가 존재할 때 이 문자열이 실제 사람이 사용하는 자연언어일 확률을 모델링함

- 어떤 문장(단어의 sequence)가 주어졌을 때 문장이 실제 자연언어로 사용된 확률을 계산
- 아래의 예처럼 한국어로 문법에 맞고 자연스러운 문장 A의 확률은 어순이 바뀌어 약간 어색한 문장 B보다 언어모델에서 확률이 높고, 전혀 쓰이지 않을법한 문장 C보다는 문장 B의 확률이 높다.

A= 내가 학교에 간다. B=학교에 내가 간다. C=간다 내가 학교에. P(A) = 0.1, P(B) = 0.001, P(C) = 0.000000001

- 언어모델의 활용은 확률적 언어모델을 기반으로 다음 단어를 예측하는 것으로 n개 단어가 주어졌을 때 n+1 번째 단어가 발생할 확률을 이용함
- 음성인식에서 문장을 순차적으로 인식할 때 다음 단어의 후보가 많을 경우 그동안 인식된 단어들 다음에 와서 가장 확률이 높은 단어를 선택함

기존 인식 단어열(S): 나는, 강당에서, 영어로 인식 후보 단어: 1) 발표합니다 2) 발포합니다. P(S|발표합니다) 〉 P(S|발포합니다) 일 경우 선택 단어는 '발표합니다'

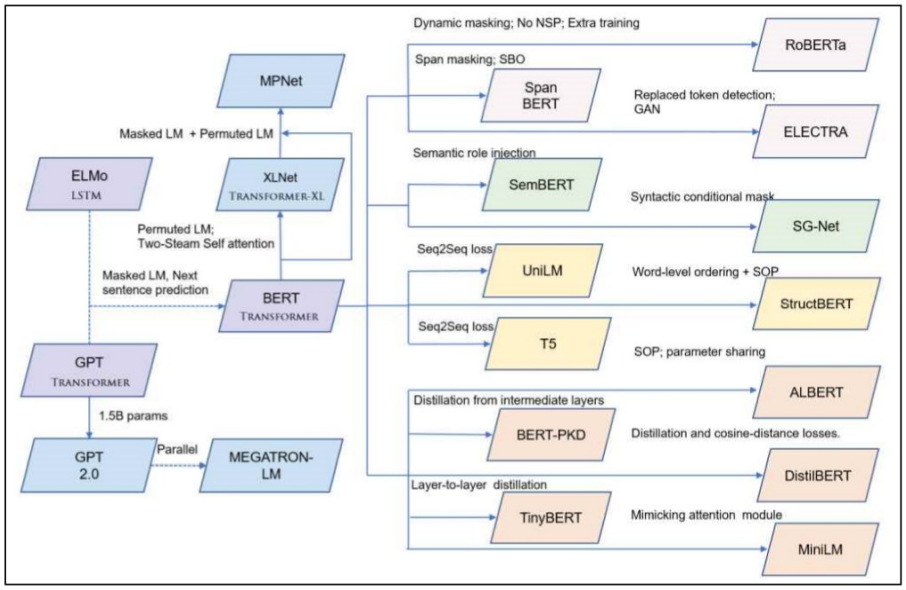
□ 기존에는 확률 모델이나 통계 방법 등 고전적인 방법을 사용하였으나, 최근에는 언어모델을 구축하기 위해서 BERT나 GPT-3 같은 전이 학습을 활용

○ 미리 방대한 양의 학습 데이터로 사전 학습을 시키고, 이 모델에 전이 학습 기법을 활용하여 미세조정을 통해 소량의 학습데이터와 적은 학습 시간 만으로도 목적 태스크에 대한 인공지능 모델을 구축

○ BERT는 각 단어에 대해서 주변 단어들과의 자기 집중(self-attention) 연산을 거친 결과(벡터)에 해당 단어의 문맥이 표현되어 있다고 보고, 이를 응용 태스크에 적용하는 방법을 제안

- BERT 언어모델의 입력과 출력은 N개 단어를 입력하고, 각 토큰에 해당하는 벡터를 출력하도록 구성됨
- 사전 학습 태스크는 Masked Language Model(MLM)과 Next Sentence Prediction(NPS)로 구성됨
- MLM은 주변 단어를 이용하여 해당 단어를 예측하는 태스크이고, NSP는 두 개의 문장이 선/후 관계인지 여부를 판단
- BERT는 base 모델의 경우 12 layer, 12 multi-head, 768 차원의 단어 벡터로 구성, large 모델의 경우 24 layer, 16 muli-head, 1024 차원의 단어 벡터로 구성됨
- BERT를 학습하기 위해 Books Corpus(800M 단어), 영어 위키백과(2,500M 단어)를 사용하였고 사용한 총 용량은 16GM, 사전 학습은 256 batch, 1M step 횟수 학습하였으며, 30,000개 word piece 부분 어휘 단어로 문장을 분할하여 학습

□ BERT 이후의 언어모델 연구 동향은 아래 그림과 같으며, 주요 모델 설명은 아래와 같음(그림 2-8)



[그림 2-8] BERT 이후 딥러닝 사전학습 언어모델 연구 동향

❍ 워싱턴대, 프린스턴대, AllenAI 연구소 페이스북에서 2019년에 SpanBERT를 공개

- BERT 대비 주요 개선 내용은 Span Masking, Span Boundary Objective, single-sequence training
- Span masking은 독립된 단어에 masking 하는 BERT와 달리 연속된 단어 span을 masking 하여 좀더 실질적인 언어 현상을 반영
- Span Boundary Objective는 masking한 단어의 원 단어를 예측할 때 해당 단어의 출력 벡터의 손실함수(loss function)에 span의 가장자리에 있는 단어의 손실함수를 추가하여 학습(그림 2-9)

  ※ (예시) 그림 9에서 문장의 7번째 단어가 "football" 이라면, x7 의 출력이 football인지를 예측하는 loss와 span의 경계에 해당하는 x4, x9와 7번째 단어의 위치 임베딩인 p7을 이용한 결과가 football인지에 대한 loss를 결합

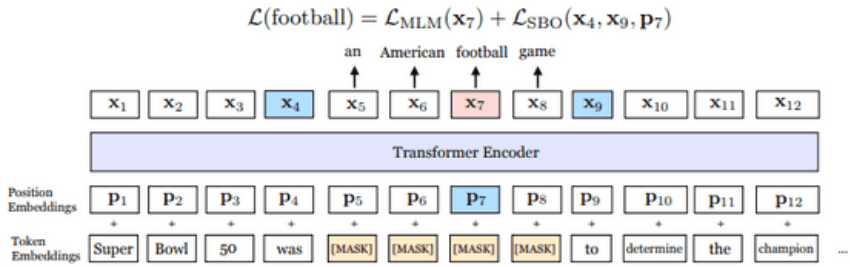


Figure 1: An illustration of SpanBERT training. In this example, the span *an American football game* is masked. The span boundary objective then uses the boundary tokens *was* and *to* to predict each token in the masked span.

출처 : Mandar Joshi 외(2019) SpanBERT: Improving Pre-training by Representing and Predicting Spans

[그림 2-9] SpanBERT Span Boundary Objective 예시

- BERT의 nsp를 제거하고 단일 segment로 입력을 구성하여 학습을 하는 방법을 채택하였고 nsp를 제거할 경우 성능이 개선됨
- 기계독해 데이터셋 대상 실험 결과, SQuAD v1.1 평가셋에서 BERT 보다 3.3%p 우수한 94.6%의 성능, SQuAD v2.0 평가셋에서 BERT보다 5.4%p 우수한 88.7%의 성능

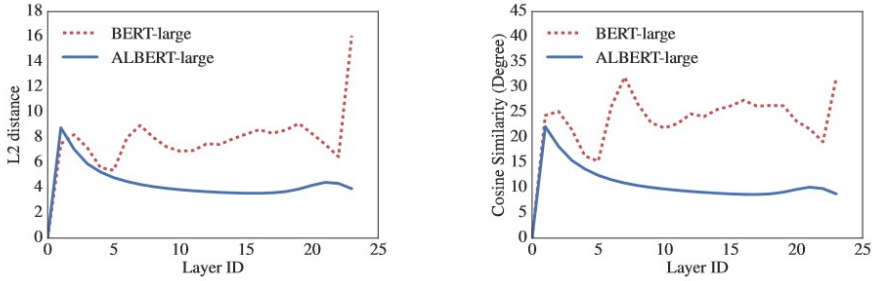
❍ CMU, Google Brain은 BERT 모델의 masked LM의 한계 극복을 위하여 permutated LM과 이를 위한 two-stream self-attention 기반 XLNet ('19.06) 모델을 발표하였으며, SQuAD 2.0에서 BERT 대비 7% 가량 성능을 개선함

❍ Facebook은 BERT 모델의 사전학습 방법을 개선하여, dynamic masking, no next sentence prediction, large batch의 방법으로 BERT 모델보다 SQuAD 2.0 에서 7.6% 가량 개선된 RoBERTa('19.07) 모델 발표

- 매번 같은 마스킹된 부분을 이용하여 학습하는 것이 아니라 매 학습 단계마다 마스킹된 부분을 새롭게 할당하여 적용하는 dynamic masking을 채택
- SpanBERT와 마찬가지로 nsp 태스크를 적용하지 않았으며, 기존 256 배치보다 훨씬 큰 2K 또는 8K 배치를 적용함

❍ 구글에서는 BERT large 보다 큰 모델을 효과적으로 학습하고 성능을 개선하기 위하여 ALBERT(A Lite BERT for self-supervised learning of language representations)를 공개

- BERT와 비교하여 개선된 점은 factorized embedding parameterization, cross-layer parameter sharing, inter-sentence coherence loss 임
- 주변 문맥에 독립적인 토큰 임베딩 파라미터의 차원을 문맥을 유지해야 하는 트랜스포머 레이어 내의 히든 벡터와 같은 크기를 유지하지 않고 1024 차원에서 128차원으로 대폭 줄이는 방법을 도입(factorized embedding parameterization)
- 트랜스포머의 각 레이어에 포함된 학습 파라미터를 모든 레이어에 동일하게 적용하는 cross-layer parameter sharing을 도입하였으며, 각 레이어의 입력 벡터와 출력 벡터 사이의 L2 distance 및 코사인유사도를 보면 도입하지 않은 경우(그림 2-10 좌측)보다 도입한 경우(그림 2-10 우측)이 안정적인 변화를 보임



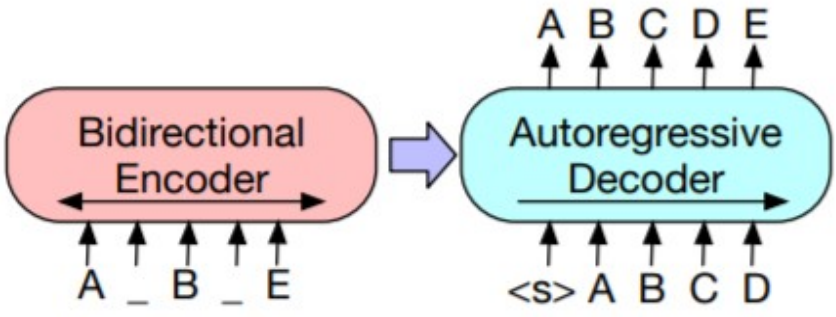
출처 : Zhenzhong Lan 외(2019) ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations

[그림 2-10] ALBERT cross-layer parameter sharing 도입 효과

- BERT의 nsp 문제를 극복하기 위해서 동일 문서에서 연속적은 추출된 부분의 순서가 바뀌었는지 여부를 인식하는 SOP(Setence Order Prediction) 태스크로 변경하여 학습
- 160G 학습 데이터를 사용하고, 4K 배치 크기에, 1.5M step 학습한 결과, SQuAD v2.0 테스트셋 평가 결과, Single 모델은 90.9% F1, ensemble 모델은 92.2% F1 성능을 보며 기존 방법을 뛰어넘음

❍ 페이스북에서 공개한 BART(Bidirectional and Auto-Regressive Transformer)는 BERT와 후속 연구가 트랜스포머의 인코더만 사용한 것과 달리 인코더, 디코더를 같이 사용하여 언어 모델을 학습

- BART에서 디코더는 이전 단계의 출력 결과를 다음 단계의 입력으로 사용하는 Auto regressive 방식을 채택(그림 2-11)
- 트랜스포머 인코더만 사용하면, 입력과 출력 단어 수가 같아야 한다는 제약이 있지만, 인코더와 디코더를 함께 사용하면 다른 경우에도 학습이 가능
- 디코더를 함께 채택한 이외에도 token masking, token deletion, text infilling, sentence permutation, document rotation의 사전 학습 태스크를 제안
- text infilling은 임의 길이 텍스트 span을 하나의 단위로 간주하여 masking하고 이를 학습하며, 특이한 것은 아래 그림과 같이 대상 단어가 없는 경우에도 masking을 하여 단어가 없음을 인식하는 학습까지 포함함
- 그림 좌측에서 A와 B 사이 masking('_')은 A와 B 사이에 단어가 없음을 학습하기 위함이고, B와 E 사이 masking은 단어 C D를 동시에 인식하기 위해 학습



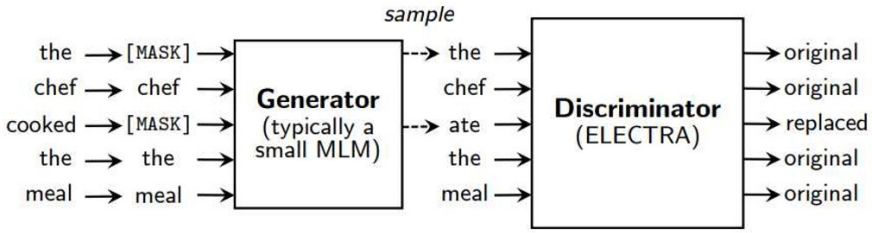
출처 : Mike Lewis 외(2019) BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension

[그림 2-11] BART text infilling 예제

- BART는 RoBERTa와 비슷한 수준의 언어이해 태스크 성능을 가지고 있으며, 기존의 BERT 및 RoBERTa를 적용하기 어려운 언어 생성(generation) 업무도 우수하게 수행할 수 있음

❍ MS는 MT-DNN(Multi-Task Deep Neural Network) 방법으로 BERT와 같이 2단계 학습(Pre-training, Fine-tuning)을 수행하나, 재학습 단계에서 응용 태스크 별 신경망을 추가한 멀티 태스크 학습방법으로 BERT 대비 약 1.8% 성능개선을 이룸

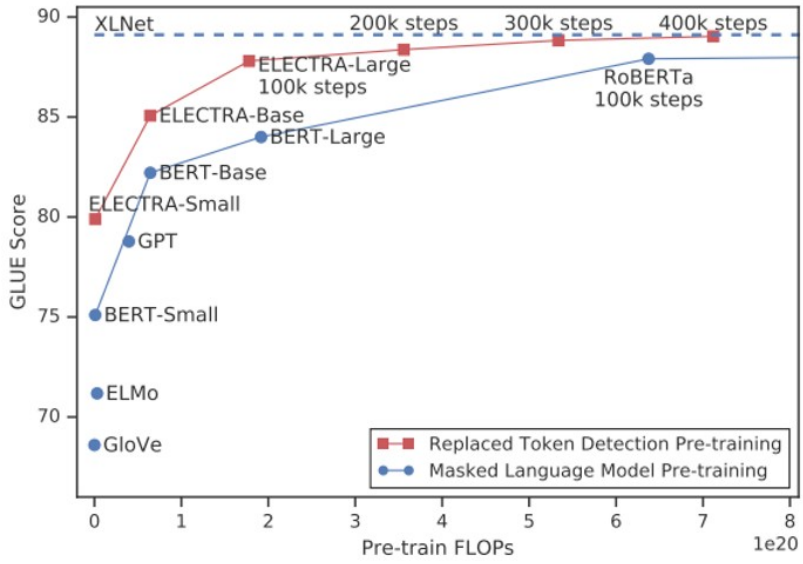
❍ Stanford Univ.는 언어모델 사전학습을 위하여 GAN과 유사한 생성-분류 구조를 가지는 ELECTRA 모델을 공개하였으며(그림 2-12), 이는 기존 BERT 모델 대비 8.4%, RoBERTa 모델 대비 1.6% 우수한 성능을 보임



출처 : Kevin Clark 외(2020) ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators

[그림 2-12] GAN 구조 개념을 언어모델에 도입한 ELECTRA 모델

- ELECTRA에서 해결하고자 한 문제는 N개의 입력 단어 중 15%만 masking 하여 loss 값이 계산되기 때문에 언어모델의 학습 효율성이 떨어진다는 점
- 학습 효율을 향상시키기 위해 N개 입력 단어 전체에서 loss를 계산하기 위한 방법을 제안
- 전체 텍스트로부터 loss를 계산하기 때문에 초기 학습속도가 빠르며, 동일한 사전학습 FLOPs를 처리할 경우, BERTBase, BERT-Large, RoBERAa보다 우수한 성능을 보임(그림 2-13)



출처 : Kevin Clark 외(2020) ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators

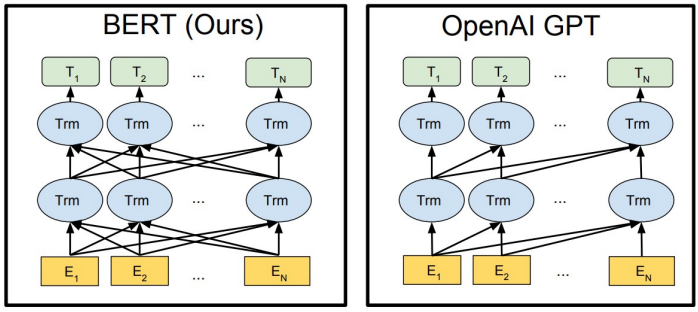
[그림 2-13] 다른 딥러닝 언어모델과 ELECTRA 학습 속도 비교

- 약 142GB의 학습 데이터를 사용하였으며, 같은 용량의 학습 데이터로 BERT large 모델 대비 4.4배, RoBERTa 모델 대비 4.5 배의 학습을 수행한 결과 같음. 전체 학습 데이터를 모두 사용하여 같은 학습 데이터를 4.5배 정도 더 활용 가능
- SQuAD 2.0 테스트 셋 기준, RoBERTa 모델 89.8% F1, ALBERT 90.9% F1보다 우수한 91.4% F1 성능

❍ 딥러닝 언어 모델은 크게 언어를 이해(Encoder)를 위한 모델과 언어를 생성(Decoder)하는 모델로 구분이 되는데 앞에서 언급한 BERT를 비롯한 후속 모델은 모두 언어 이해 모델이며 다음에 설명할 GPT(Generative Pre-trainned Transformet) 모델은 언어 생성 모델임

❍ OpenAI에서 최근 공개한 딥러닝 언어 모델 GPT-3는 1750억개(175Billion) 파 라미터로 3000억(300B)개 데이터를 학습하였으며, 사후학습(fine-tuning) 없이 응용 태스크에서 우수한 성능을 내는 퓨샷(few-shot) 학습이 가능함을 보임.

- GPT 언어생성 모델은 입력 텍스트의 다음 단어를 예측하는 방식으로 동작하여 앞에서 설명한 확률적 언어모델과 유사한 방식으로 작동하며 양방향으로 연결된 BERT(그림 2-14 좌측)와 다르게 GPT(그림 2-14 우측)는 진행 방향(forward)으로만 연결된다는 점이 특징임



출처 : Jacob Devlin 외(2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

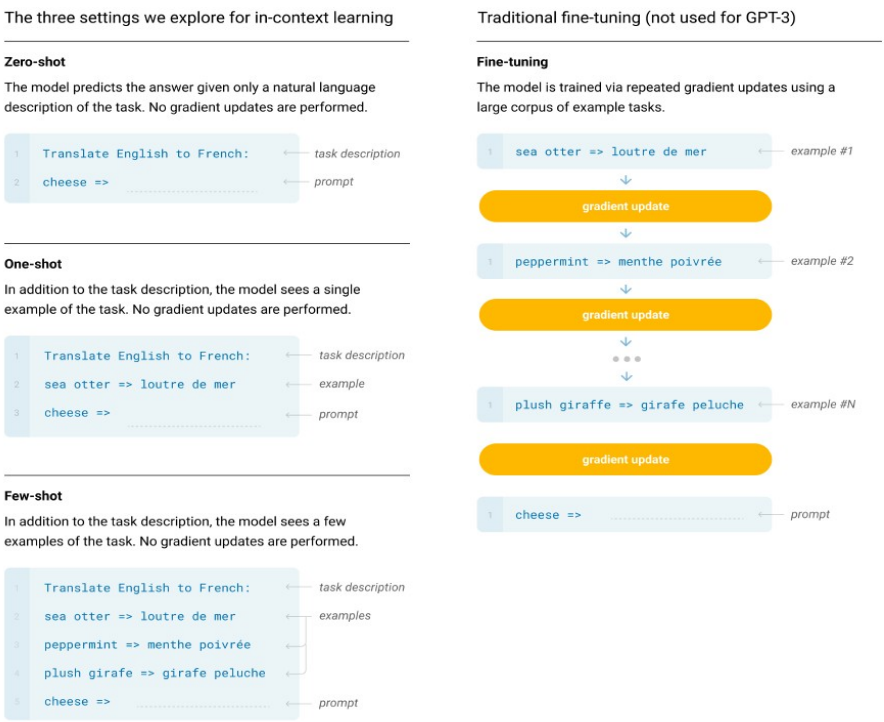
[그림 2-14] 언어 이해모델 BERT와 언어 생성 모델 GPT 비교

- OpenAI API인 GPT-3는 미세조정이 없는 퓨샷(few-shot) 학습을 통해 언어 모델의 학습 효율을 개선함

  ※ 퓨샷 학습: 상대적으로 적은 데이터로 인공 신경망을 재학습하는 방법으로 퓨샷보다 적은 데이터를 사용하는 개념으로, 하나의 데이터는 원샷, 데이터 없이새로운 태스크에 언어모델만으로 바로 적용하는 방법은 제로샷이라고 하며 퓨샷 학습은 추후 AGI(Artificial General Intelligence) 달성에 중요한 능력으로 평가됨

- BERT 및 GPT-2 와 같은 기존 딥러닝 언어 모델 기술은 추가학습(미세조정, fine-tuning)을 적용해야만 응용 태스크에 적용가능하지만 GPT-3는 이러한 과정 없이 적용 가능하다고 알려짐

- 일반적인 사후학습(그림 2-15 우측)과 달리 GPT-3는 단순히 학습 예제 입력(그림 2-15 좌측) 만을 통해서 태스크를 달성할 수 있음



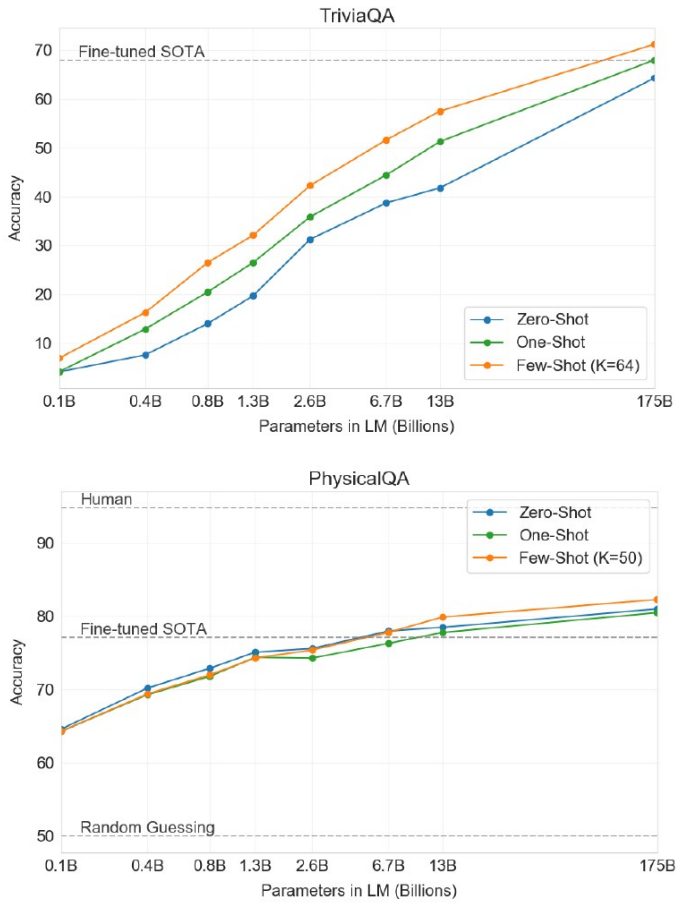
출처 : Tom B. Brown 외(2020) Language Models are Few-Shot Learners

[그림 2-15] GPT-3와 다른 모델의 작동 방식 비교

- LAMBADA 분야를 기준으로 기존의 세계 최고 수준인 68.0 정확도를 뛰어넘어 86.4 정확도 확보

○ 24개 이상의 언어처리 태스크에 적용하여, GPT-3 모델의 퓨샷-학습 능력 평가

- 적용 태스크 예: 문장생성, 기계번역, 상식 추론, 기계독해, 문장 간 관계추론, 산술연산 등
- 특히, GPT-3가 생성한 뉴스 기사 텍스트는 사람이 구분하기 어려운 수준으로 평가됨
- TriviaQA 및 PysicalQA(상식 추론) 태스크는 추가학습을 적용한 기존 최고 성능보다 우수한 성능을 보임(그림 2-16)



출처 : Tom B. Brown 외(2020) Language Models are Few-Shot Learners

[그림 2-16] GPT-3를 적용한 TriviaQA PysicalQA 정확도 비교

- OpenAI는 GPT-3의 놀라운 성능으로 인해 모델 자체를 공개했을 때 악용되는 것을 방지하기 위해 베타 버전의 API 형태로 공개한다고 주장

□ 언어처리 기술 고도화에 따른 서술형, 복합 근거형, 상식 추론형 등 고난이도 문제 도전 진행 중

❍ 단일 단락에서 사용자 질문의 단답형 정답을 추론하는 문제는 사람보다 우수한 성능으로 평가됨에 따라, 문서로부터 정답을 찾는 문제, 서술형

정답을 찾는 문제, 복합 근거형 및 상식 추론형 등의 고난이도 문제에 대한 연구 및 챌린지 진행 중

- Stanford Univ.는 위키피디아(Wikipedia)에서 추출한 질문과 단락에서 단답형 정답을 맞추는 기계독해 챌린지(MRC: Machine Reading Comprehension) SQuAD를 '15년부터 추진하여 15만 건의 학습 데이터셋을 확보하고 언어처리 분야를 선도하고 있음
- 구글은 Natural Questions(NQ) 챌린지를 '19년에 시작하여, SQuAD와 같이 단락에서 정답을 찾는 것이 아니라, 문서에서 정답을 찾아 야하는 고난이도의 챌린지를 운영하고 있음(30만 건의 학습 데이터셋)
- 구글, 스탠포드대, 카네기멜론대학은 '18년 HoppotQA 챌린지를 통해 위키피디아 전체 문서의 여러 문장에서 문제의 답을 선정한 근거를 추출하는 설명 가능한 질의응답(explainable QA) 기술인 Multi-hop reasoning 기술 개발을 추진하고 있으며 약 11만 건의 학습 데이터셋을 제공하고 있음
- 앨런 연구소는 질문과 정답문장의 어휘 매핑 관계를 이용하여 정답을 추출하는 현재 기술의 한계를 극복하기 위하여 문장의 의미이해(초등 수준)와 상식적 추론이 가능한 차세대 질의응답 챌린지 시작

  ※ 앨런 연구소는 AI2 Reasoning Challenge(ARC)를 '18년에 시작하여, 단순 정보검색 기반의 답변 추출이 어려운 초등수준 과학 분야 객관식 7,787건 학습 셋을 제공함

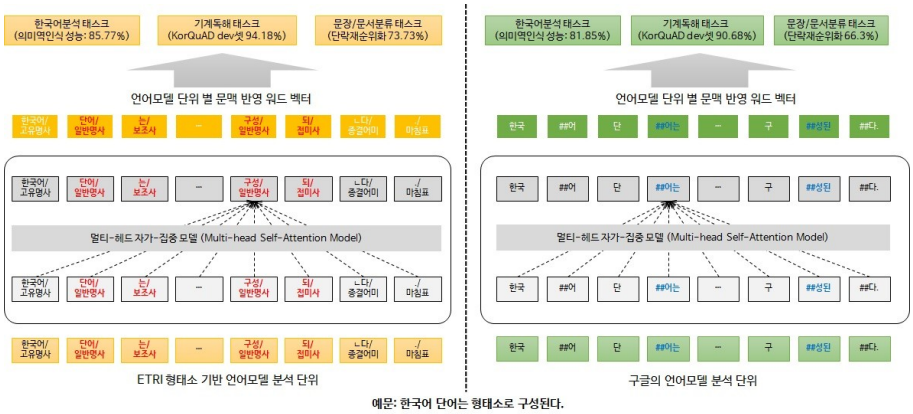
  ※ 앨런 연구소는 CommonsenseQA를'19년에 시작하여, 상식적 추론이 필요한 다지선다형 문제 12,102건 학습 셋을 제공함

❍ 향후 언어지능 기술은 하나의 단락 및 문서로부터 정답을 추출하는 현 단계를 넘어서 여러 가지 소스에서 정답을 추출하고 내용을 종합하는 추론 단계로 발전하여 귀납적 · 연역적 추론을 통해 논리적으로 지식을 자동생성할 것으로 전망됨

□ 딥러닝 사전학습 언어모델 기반 한국어 처리 기술 고도화

❍ 한국어처리 분야도 영어권 연구와 유사하게 딥러닝 언어모델 및 질의응답 기술이 연구 중이며, 한국어와 영어의 언어적 차이로 인하여 한국어에 적합한 딥러닝 언어모델 기술이 제안됨

- 한국어는 내용어(명사, 동사)와 기능어(조사, 어미)가 결합하는 교착어이나, 영어는 어미가 활용에 따라 변화하는 굴절어라는 특징을 가짐
- 딥러닝 언어모델 기술로 엑소브레인(Exobrain) 과제에서 형태소분석 결과에 기반한 KorBERT 언어모델을 개발 및 공개하여, 학계 및 산업계의 주요 언어 모델로 활용됨
- KorBERT 언어 모델은 백과사전, 신문 기사 등 방대한 텍스트(23GB)를 대상으로 47억 개의 형태소를 학습했으며, BERT 모델의 MLM 방법론을 개량하여 성능을 개선함(그림 2-17)



출처 : 공공 인공지능 오픈 API · DATA 서비스 포털

[그림 2-17] KorBERT와 구글 BERT 비교

- 형태소 분석 적용 여부에 따른 응용 태스크별 성능 개선 정도를 측정하기 위하여 2가지 모델을 공개함(그림 2-18)

  ※ KorBERT-Morphology 모델은 형태소 분석 결과를 기반으로 하고 있으며, KorBERT-Wordpiece 모델은 형태소 분석을 사용하지 않아 손쉽게 사용 가능한 모델임

배포 모델	세부 모델	세부 내용	모델 파라미터
KorBERT	Korean_BERT_Morphology	▪ 학습데이터: 23GB 원시 말뭉치 (47억개 형태소) ▪ 형태소분석기: 본 OpenAPI 언어분석 중, 형태소분석 API ▪ 딥러닝 라이브러리: pytorch, tensorflow ▪ 소스코드: tokenizer 및 기계독해(MRC), 문서분류 예제 ▪ Latin alphabets: Cased	30349 vocabs, 12 layer, 768 hidden, 12 heads,
	Korean_BERT_WordPiece	▪ 학습데이터: 23GB 원시 말뭉치 ▪ 딥러닝 라이브러리: pytorch, tensorflow ▪ 소스코드: tokenizer ▪ Latin alphabets: Cased	30797 vocabs, 12 layer, 768 hidden, 12 heads,

출처 : 공공 인공지능 오픈 API · DATA 서비스 포털

[그림 2-18] KorBERT 세부 모델 설명

❍ KorBERT 언어모델 평가는 다음 5개를 대상으로 수행하였으며 구글 다국어 BERT 모델이나 word piece 기반의 한국어 모델에 비해 좋은 성능을 보임(그림 2-19)

- 의미역 인식(Semantic Role Labeling) : 문장 내에서 술어에 의해 기술되는 사건에 대한 개체들의 역할을 인식
- 기계 독해(Machine Reading Comprehension) : 주어진 단락에서 질문이 요구하는 정답을 찾음
- 단락 순위화(Passage Ranking) : 검색 결과 집합에서 질문에 찾는 정답이 들어 있는 단락 순위화
- 문장 유사도 추론(Natural Language Inference) : 2개 문장 간 의미가 동일한지 여부를 분류
- 문서 주제분류 : 문서의 주제를 기존에 사람이 정의한 54개의 클래스 중 하나로 분류

구분	의미역인식	기계독해	단락순위화	문장유사도추론	문서주제분류
평가데이터 및 규격	Korean Propbank, 학습: 19,302 문장 평가: 3,773 문장	KorQuAD 데이터, 학습: 60,406건 평가: 5,773건 (dev셋)	학습: 45,521 질문 평가: 1,000 질문 (질문당 평균 8.7개 단락)	학습: 10,874문장쌍 평가: 1,209문장쌍 (이진 분류체계: 유사, 무관)	학습: 9,301건 평가: 1,035건 (54개 분류체계)
평가 방법	F1[2]	Exact Match[3] / F1	Precision@Top1	Accuracy	Accuracy
(Google) Word Piece[4]기반 한국어 언어모델	81.85%	80.82% / 90.68% (정답 경계 구분을 위해 후처리 수행)	66.3%	79.4%	91.1%
(엑소브레인) Word Piece 기반 한국어 언어모델	85.10%	80.70% / 91.94% (정답 경계 구분을 위해 후처리 수행)	70.5%	82.7%	93.4%
(엑소브레인) 형태소 기반 한국어 언어모델	85.77%	86.40% / 94.18%	73.7%	83.4%	93.7%

[2] F1: 정확률(Precision, 시스템이 결과가 정답인 비율)과 재현률(Recall, 실제 정답을 시스템이 맞춘 비율)의 조화평균
[3] Exact Match: 시스템이 제시한 결과와 정답이 완전히 일치하는 비율
[4] Word Piece: 하나의 단어를 내부 단어(Subword Unit)들로 분리하는 단어 분리 모델

출처 : 공공 인공지능 오픈 API · DATA 서비스 포털

[그림 2-19] 한국어 5개 태스크에 대한 KorBERT 평가 결과

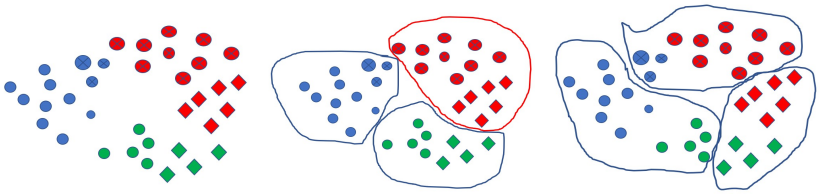
❍ 한국어 기계독해를 위한 챌린지로 LG CNS 주관의 KorQuAD 챌린지가 있으며, V1.0 은 SQuAD 1.0과 같이 주어진 단락에서 단답형 정답을 찾는 태스크이며, V2.0은 구글 Natural Questions와 유사하게 문서 단위에서 정답을 찾는 태스크 수행

❍ 한국어 질의응답 기술은 위키백과 기반 일반상식 질의응답 기술 뿐 아니라, 법률, 특허, 금융 등 전문분야 대상 질의응답 기술 연구 진행 중

제4절 텍스트 클러스터링 활용

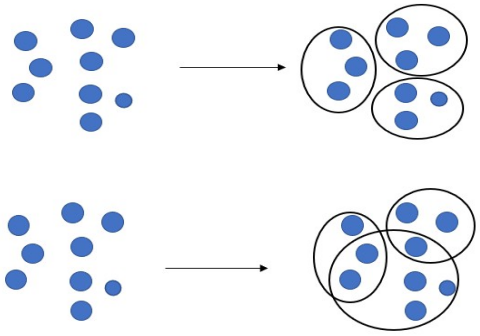
□ 텍스트 클러스터링은 미리 주제를 알 수 없을 경우 다양한 클러스터링 기법을 활용하여 문서 집합을 군집화하여 군집화 단위별로 유용한 정보를 얻을 수 있다.

○ 클러스터링은 미리 정해진 주제가 없기 때문에, 아래 그림과 같이 왼쪽의 데이터를 군집화할 때 똑같은 데이터 집합에 대해서, 가운데처럼 색깔별로 군집화도 가능하고, 오른쪽처럼 모양별로 군집화도 가능함.(그림 2-20)



[그림 2-20] 클러스터링 도식화 예제

- 위와 같은 기본적인 군집화는 1개의 문서가 1개 군집에만 속할 수 있는 hard clustering, 1개 문서가 1개 이상 군집에 속할 수 있는 soft clustering으로 구분됨(그림 2-21)



[그림 2-21] Hard clustering(위), soft clustering(아래) 도식화

○ 방대한 양의 비정형 텍스트 데이터를 유사한 주제의 문서로 군집화하면 필요한 정보를 보다 효과적으로 파악할 수 있음

- 자동으로 군집화된 결과를 이용하여 문서 군집별 통계, 트렌드 분석에 등을 활용하여 기업 및 공공 기관의 빠르고 정확한 의사 결정에 많은 기여가 가능함

□ 클러스터링 기법에도 다른 자연어 처리 분야와 유사하게 딥러닝을 적용하여 기존 방법보다 고품질로 군집화된 문서 집합을 얻을 수 있음

○ 전형적인 클러스터링 기법의 절차는 문서를 대표하는 주제어를 추출하고, 각 문서별로 대표 주제어의 유사도를 측정하여 특정 값 이상의 유사도를 보이는 문서를 같은 군집으로 묶는 방식

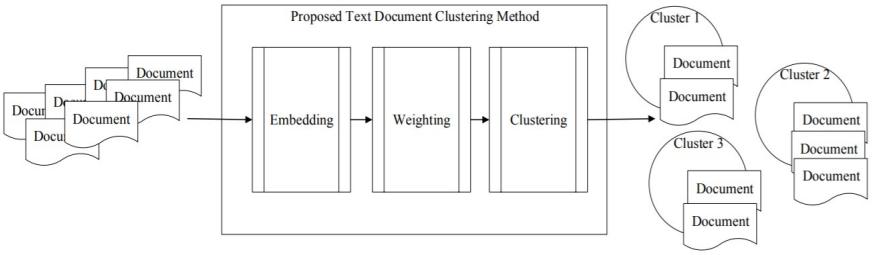
- 문서 군집화의 결과에 주로 영향을 주는 요소는 주제어 추출과 유사도 측정 방법임
- 주제어 추출은 일반적이지 않으면서도 문서를 대표하는 주제어를 추출하고 이를 문서를 대표하는 의미로 설정하는 것으로, 기본적으로 자연어 분석 기술에 좌우됨
- 유사도 측정 기술은 자연어의 의미를 고려해야 하기 때문에 자연어 단어를 그대로 이용하는 것은 쉽지 않기 때문에 주로 단어를 숫자화 하여 코사인 유사도와 같은 기법을 사용함

○ 기존 클러스터링 기법들의 대부분은 문서의 특징 정보의 파악이 미흡하고 군집 내 문서들 간의 유사도가 낮아 실무 활용이 어려웠음

- 대표적인 기존 클러스터링 기법으로 주어진 문서 집합을 k개의 군집으로 묶는 K-Means 알고리즘은 각 군집의 거리 차이의 분산을 최소화 하는 방식으로 동작함

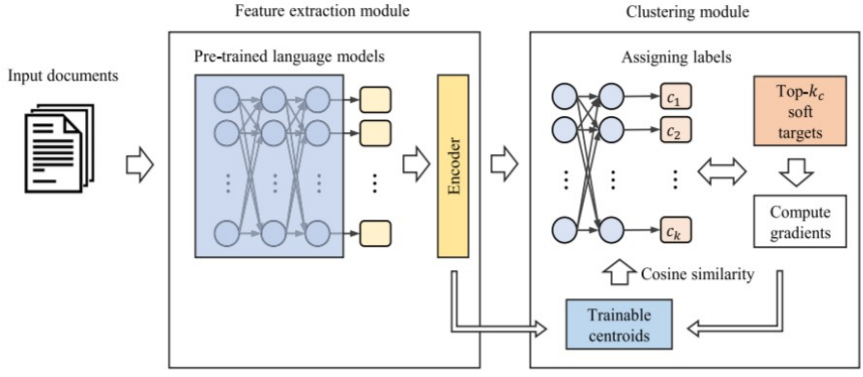
○ 최근에는 딥러닝에 기반한 클러스터링 기법이 소개되고 있는데 다른 딥러닝 기법과 유사하게 문서를 벡터로 임베딩하고 이를 이용하여 각 문서간 유사도를 측정하는 방식으로 기존 방법보다 향상된 결과를 보여줌(그림 2-22, 23)

- 기존 방법에 비해 딥러닝을 적용하였을 때 개선되는 부분은 1) 딥러닝 언어모델과 전이 학습을 이용하여 주제어를 추출하는 단계에서 충분히 문맥을 고려할 수 있다는 점과 2) 이렇게 추출된 대표 주제어를 임베딩 기법을 사용하여 벡터 공간의 적절한 위치에 사상하여 다른 문서 혹은 군집과 유사도를 계산할 때 좀더 정확해진다는 점임



출처 : Yutong Li 외(2020) A Text Document Clustering Method Based on Weighted BERT Model

[그림 2-22] 딥러닝을 이용한 클러스터링 처리 예



출처 : 박진욱 외(2019) ADC: Advanced document clustering using contextualized representations

[그림 2-23] 딥러닝을 이용한 클러스터링 시스템 구조

□ 설명가능한 AI 와 같은 최신 딥러닝과 결합된 클러스터링 기법은 새로운 편리함을 제시해줄 수 있음

❍ 딥러닝 기반 클러스터링된 군집이 어떤 주제에 관한 기준으로 군집이 된 것인지를 인간 사용자가 읽고 판단해야 함

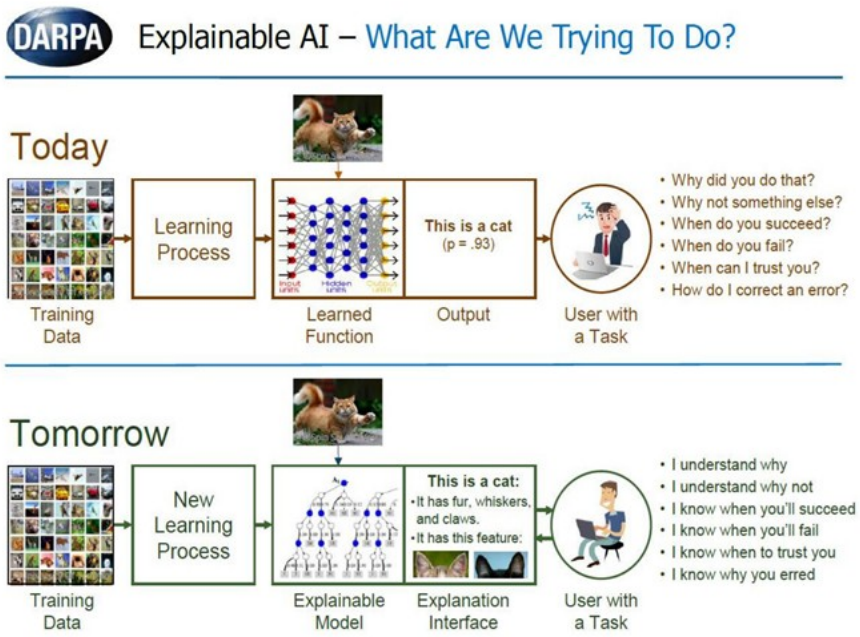
- 각 문서에서 추출된 대표 주제어를 보고 판단을 하는 방법이 있으나 단편적인 키워드만으로 판단하는 것은 쉽지 않은 작업임

❍ 이러한 불편한 점을 개선하기 위해서 설명가능한 AI와 클러스터링 결합을 고려해 볼 수 있음.

- 설명가능한 AI(Explainable AI, XAI)란 인공 지능이 예측한 결과에 대한 판단 근거를 사람이 이해할 수 있는 방식으로 제시하는 인공지능 형태를 말함
- 클러스터링 결과 군집의 예와 같이 딥러닝을 적용하여 클러스터링 결과 성능은 개선이 되었지만, 딥러닝의 모델구조 및 학습, 판단 과정이 매우 복잡하여 어떻게 학습하고 어떻게 판단하여 군집이 되었는지 사람이 이해하기는 불가능
- XAI는 판단 이유 및 근거 제시를 통하여 인공지능 결과의 신뢰성을 확보하여 실세계 인간-인공지능 협업을 위한 필수 요소로 주목 받고 있음.

❍ 시각 분야는 설명가능한 AI 연구가 가장 활발한 분야로 설명 가능성을 위한 많은 기법이 제안됨

- 아래 그림과 같이 고양이 그림을 찾아주는 인공지능 엔진에서 기존 학습 방법은 고양이로 예측한 결과가 맞을 확률이 93%라는 것을 제시해 주는데 그치지만 XAI를 적용할 경우 그에 대한 근거(털, 수염이 존재하고 귀 모양을 이미지로 제시)를 제시
- 이렇게 제시된 근거와 설명을 통해 인공지능 엔진 사용자는 예측 결과에 대해 신뢰를 갖거나, 혹은 예측 결과가 틀렸다는 것을 판단 가능함



출처 : AI Magazine, 2019

[그림 2-24] 설명가능한 인공지능 예

❍ 설명 가능한 AI 기술은 단일 접근방법이 아닌 다양한 유형의 접근방법이 연구되고 있으며, 주요 접근방법의 예는 아래와 같음

- 입력 요인(input attributaion) 분석: 관련성(relevance) 점수 또는 딥러닝 모 델의 경도(gradient) 분석을 통한 설명가능성 제시 방법으로, LRP, RAP, DeepLIFT, Guided Backprop, GradCAM 등의 방법 제안
- 내부(Internal) 분석: 딥러닝 모델 내부 뉴런의 활성화 조건을 분석하는 접근 방법으로, Network Dissection, GAN Dissection 등의 방법 제안
- 집중(Attention) 분석: Attention을 통한 설명가능성 제시 접근방법으로, RETAIN, Saliency Maps 등의 방법 제안
- 대리(Surrogate) 모델 분석: 설명가능성을 제시하는 대리 모델(linear, tree, 규칙 기반 등)을 학습하는 접근방법으로, LIME, SHAP, DeepRED, RULEX 등 의 방법 제안

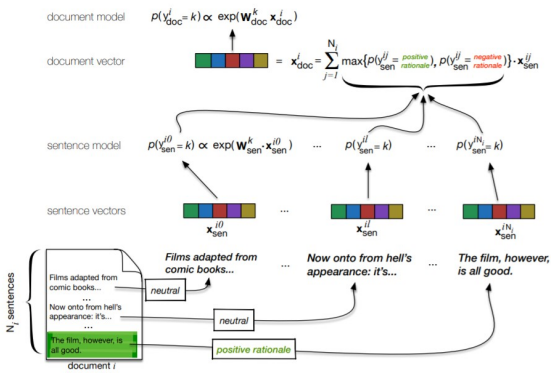
❍ 모델 설명 가능성 연구는 단일 접근 방법보다 주어진 환경 및 목적에 맞는 접근 방법을 선별하여 사용하는 추세이며, 크게 ante-hoc 접근 방법과 post-hoc 접근 방법으로 구분

- 모델 생성 단계부터 설명 가능한 특화 모델을 구축하는 ante-hoc 접근 방법은 설명 가능성을 제공하는 학습 모델이라는 장점을 가지나, 최근 딥러닝 모델 수준의 정확도를 제공하기 어렵다는 단점
- 별도의 딥러닝 모델을 학습한 후, 테스트 케이스 실행을 통해 설명을 제시하는 post-hoc 접근 방법은 높은 정확도를 제공하는 딥러닝 모델에 설명 가능성을 부여한다는 장점으로 최근 주요 흐름으로 연구 중

❍ 다른 분야와 다르게 자연언어 분야 설명가능 AI 연구는 상대적으로 단순한 텍스트 분류 태스크에 적용되어, 텍스트 분류 결과에 대한 설명을 제시하는 연구가 많이 수행됨

- (UT-Austin) rationale 정보를 CNN에 병합한 분류 모델(Rational augmented CNN model)을 제안하여, 텍스트 분류 결과에 대한 설명을 자동으로 생성함

  ※ 문장을 같은 중요도로 생각하는 것이 아니라 입력 단계에서부터 중립(neural), 긍정(positive rationale), 부정(negative rationale)로 분류하여 텍스트 분류 시에 참고하도록 하고, 이를 분류 결과와 함께 제시(그림 2-25)



출처 : Ye Zhang 외(2016) Rationale-Augmented Convolutional Neural Networks for Text Classification

[그림 2-25] Rational augmented CNN Model

- (Univ. of Amsterdam) 딥러닝 기반의 텍스트 분류 결과 해석을 위하여 분류기에서 두각을 보이는 피쳐 단어를 추출하고, 추출된 단어에 기반한 분류 모델 제안하였으며, 앞에서는 문장 단위로 판단한 것을 단어 단위로 판단했다는 점에서 차이점

제3장 지능형 연구개발정보데이터 분석시스템 온라인화

제1절 추진결과

제2절 분석시스템 사용 매뉴얼

제1절 추진결과

□ 지능형 연구개발정보데이터 분석시스템 온라인화는 2020년도 K2Base 개발용역에 포함하여 진행하였으며, 용역사와 지속적인 협의를 통하여 온라인화 및 환경개선을 추진함

❍ 기존에 개발된 분석시스템의 기능과 구조, 사용된 알고리즘 및 라이브러리(패키지) 등에 대해 용역사에 설명하고 개략적인 개발 방향을 논의

❍ 기존에 개발된 분석시스템이 설치된 PC의 환경 구성 및 분석시스템 소스 코드를 분석하여 온라인화를 위해 개발이 필요한 부분을 도출함 (그림 3-1, 2)

- 웹브라우저 기반 온라인 사용을 위한 html, JSP 개발이 필요하며, 다중 사용 시 안정성을 위한 DB 업그레이드 필요
- 기존 데스크탑에서 GPU가 탑재된 서버로 이관
  ※ 연말에 구축 예정인 KISTEP 원내 서버 자원 할당 계획

❍ 기존에 구동 자체에 문제는 없으나, 사용환경(GUI) 측면에서 일부 존재했던 세부적인 불편점에 대한 개선

- 불필요한 기능 제거 및 메뉴 재구성 병행

❍ 온라인화 작업(포팅, porting) 추진 및 2019년 조사·분석 데이터를 추가

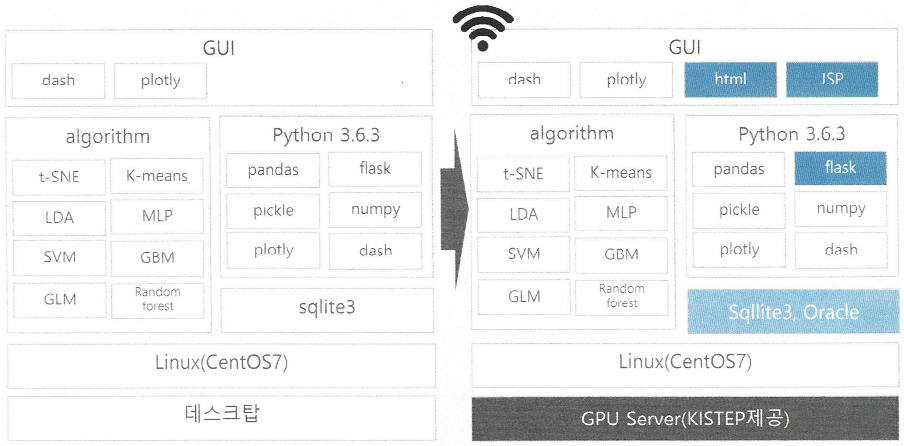
- 기존에 탑재되어있던 데이터에 추가하여 doc2vec 재학습 수행

❍ 향후 유지보수의 용이성을 위한 코드 정리 및 문서화, 매뉴얼 작성

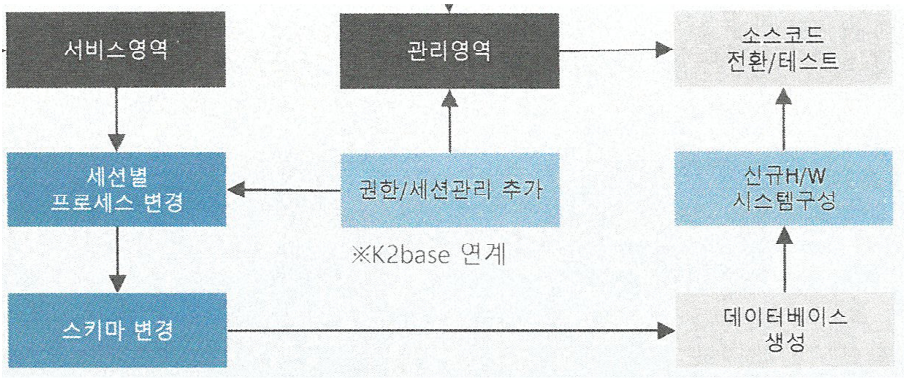
❍ 테스트 서버 운영을 통한 지속적 피드백을 수행하였고, 향후 K2Base 연계 및 개인화·권한관리 방안 결정

- 복수의 사용자가 동시 접속하여 사용할 수 있게 하기 위한 권한·세션 관리 기능이 필요하며, 기존에 합쳐져 있었던 서비스영역과 관리영역의 분리 작업이 필요함

- 추후 서버 이관 후에도 안전성 모니터링 및 유지보수가 필요할 것으로 예상됨



[그림 3-1] 분석시스템 온라인화를 위한 추가 구성요소



[그림 3-2] 분석시스템 온라인화 작업 흐름도

□ 기존에 사용에 불편함이 있었던 세부적 요소들에 대한 개선을 추진함

○ 결과 표의 csv 파일 내보내기(다운로드), 분석 버튼 등 2회 클릭해야지만 실행되었던 버튼들을 1회만 클릭해도 실행되도록 수정

- html, JSP 웹프로그래밍 개선을 통해 부드러운 사용이 가능하도록 개선하였음

□ 완성도 및 필요성이 낮은 메뉴(과학기술표준분류 자동분류기, 보고서 자동 생성기)는 온라인 버전에서는 제거함

○ 기존에 개발되어 있었던 과학기술표준분류 자동분류 기능은 실험적으로 개발했던 사항으로, 분류 정확도가 실제 사용할 수 있을만큼 높지 않고 NTIS에 유사한 기능*이 도입됨에 따라 필요성이 낮아져 삭제함

* 딥러닝 학습 기반 과학기술표준분류 추천 기능으로, 과제명과 요약문을 입력하면 어떤 과학기술표준분류에 해당하는지 추천해주는 기능임

○ 보고서 자동 생성 기능은 신약개발 정부R&D 포트폴리오 통계브리프 보고서만을 위한 기능으로서, 원내 공동 활용 필요성은 없으므로 온라인 버전에서는 삭제함

- 그래프 생성은 데이터에 기반하여 자동으로 이루어지나, 워딩의 경우 사람이 사전에 입력해놓은 템플릿에 숫자만 교체되는 방식임
- 기존에 개발된 기능은 추후 신약개발 통계브리프 작성 시 오프라인 버전에 구현되어 있는 기능을 활용할 계획임
- 단, 이와 같은 자동 리포팅 기능은 실무적 활용도가 잠재적으로 높아 추후 유사한 기능 필요시 개발을 위해 소스 코드 활용이 가능

□ KISTEP 내 온라인 서비스를 위해 중앙 서버의 가상화 작업 및 도메인 부여가 필요하며, 이후 K2Base 의 메뉴에 편입할 예정임

○ K2Base 메뉴 중에서 KISTEP 직원에게만 노출되는 “KISTEP 업무지원” 메뉴의 세부 메뉴로 추가될 계획

○ 온라인화 서비스 시작 초기 모니터링을 통해 시스템 부하 등을 파악하고, 최대 동시 사용자 수를 결정할 계획

제2절 분석시스템 사용 매뉴얼

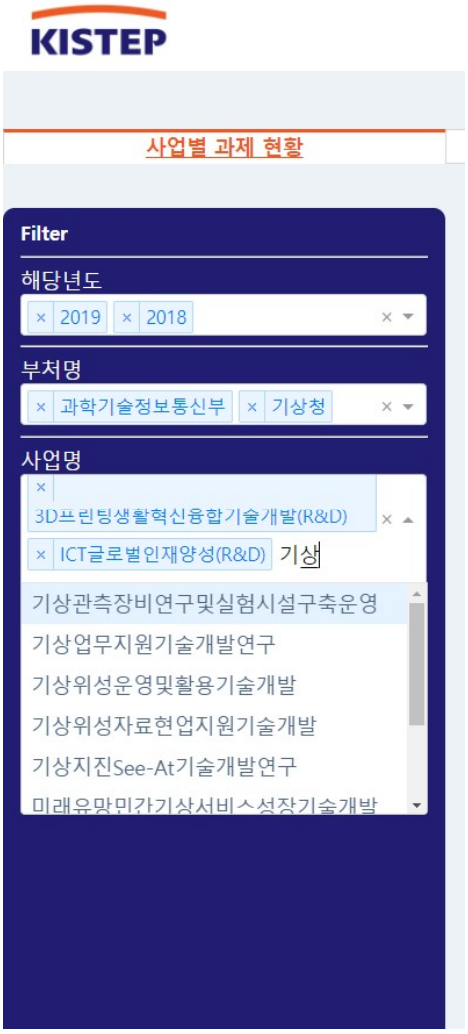
□ 동 절에서는 온라인화된 지능형 R&D정보데이터 분석 시스템의 기능별 사용법을 정리하여 추후 활용 시 도움이 될 수 있도록 하고자 함

○ 동 시스템에서 제공하는 메뉴는 다음과 같으며, 기능별로 사용법을 설명하고자 함

1. 기본 분석
1.1. 사업별 과제 현황
1.2. Word Cloud
1.3. 과제 검색 및 연구비분석
1.4. 연도별 사업내용 변화 분석
1.5. 사업간 연관성 계층분석
2. 국내/해외 클러스터링 분석
3. 과제 자동분류 학습기

□ (1.1. 사업별 과제 현황) 연도/세부사업 별 과제 목록을 보여주는 기능으로, 좌측의 Filter 메뉴에서 ① 해당년도를 선택하고, ② 부처명을 선택하면, ③ 사업명에서 선택 가능한 사업 리스트가 나타나며 사업명의 앞글자들을 입력하면 자동완성으로 사업명이 제시됨(그림 3-3)

○ 해당년도나 부처명은 여러개를 선택할 수 있으며, 입력된 내용을 삭제하려면 각 항목 왼쪽의 x표시를 클릭하면 됨



[그림 3-3] 연도, 부처명, 사업명 선택 방법

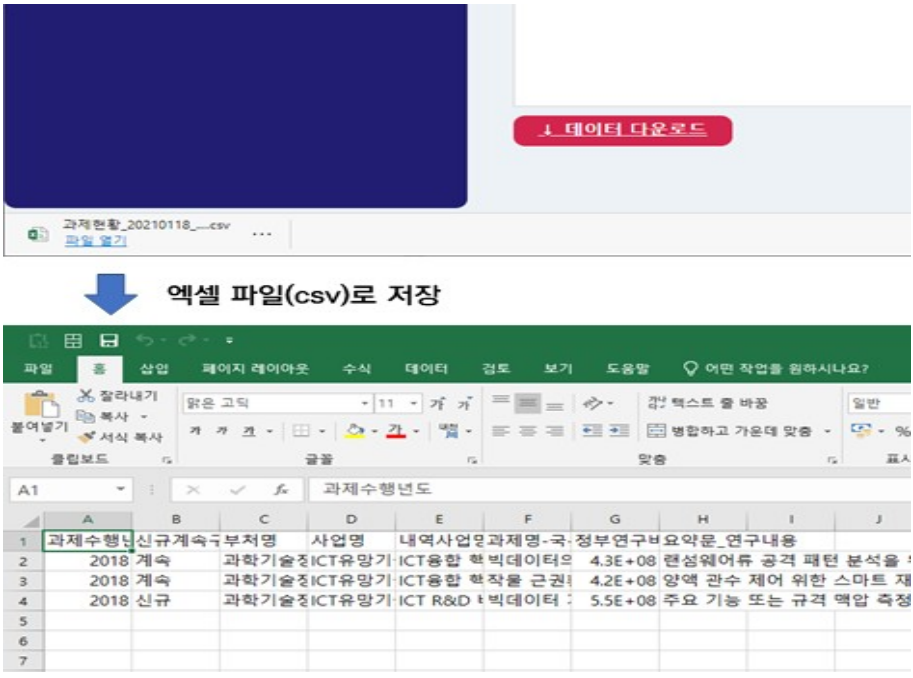
❍ 사업명까지 선택하면 우측 표에 과제 목록이 출력되며, 열 이름을 클릭하면 클릭한 열을 기준으로 과제들이 정렬되며, “Filter Rows” 버튼을 이용하면 엑셀과 유사하게 원하는 열 기준으로 필터링이 가능함

- 필터링 시, 수식(부등호)을 입력하거나, 텍스트를 입력하면 해당하는 수식을 만족하는 과제 또는 입력한 문자열을 포함하는 과제들이 출력됨

[그림 3-4] 출력 과제의 필터링 및 정렬 예시

○ 표 형식으로 출력된 과제는 화면 하단의 "데이터 다운로드" 버튼을 클릭하여 csv 형식으로 다운로드할 수 있음

- 파일 이름은 저장 시간을 반영하여 자동으로 생성됨



[그림 3-5] 표 데이터 다운로드 방법

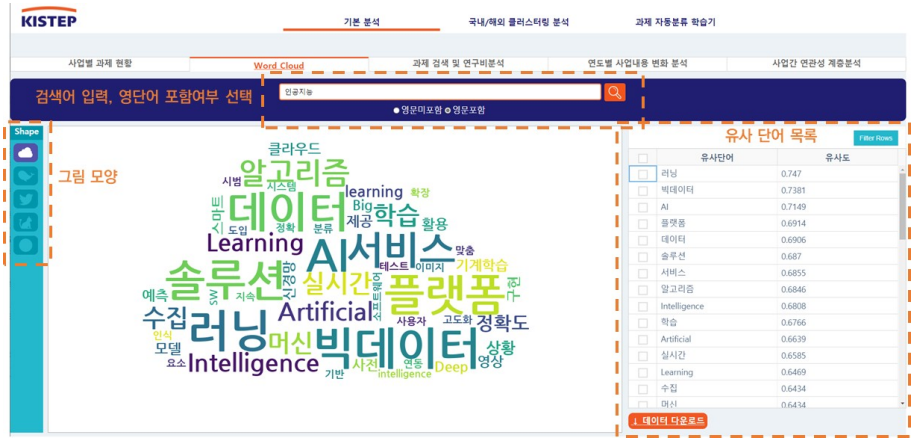
○ 표의 정렬, 다운로드 기능은 다른 메뉴에서도 동일하게 사용 가능함

□ (1.2. Word Cloud) Word2vec 학습 결과를 바탕으로, 임의의 단어와 유사도가 높은 다른 단어들의 목록을 출력하고, 유사도에 따라 글씨 크기가 달라지는 시각화 그림을 출력하는 기능임

○ 유사도가 높을수록 그림에서 글씨의 크기가 크며, 구름, 고래, 새, 고양이, 원 등 다양한 모양의 그림을 선택할 수 있음

○ "단어 혹은 문장을 입력해주세요..."라고 써있는 칸에 원하는 검색어를 입력하고, 우측의 주황색 돋보기 버튼을 클릭하면 워드클라우드 그림 및 단어 목록이 출력됨

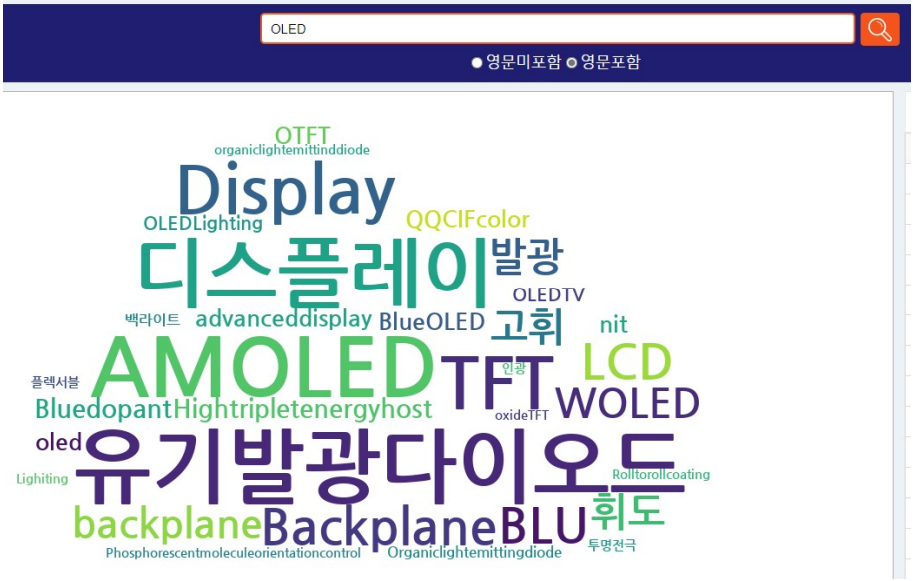
- 그림은 오른쪽 마우스 클릭하여 그림을 다른이름으로 저장하여 다운로드할 수 있으며, 유사 단어 목록은 상기 기술한 방법과 동일하게 정렬 또는 다운로드가 가능함



[그림 3-6] 워드클라우드 기능 설명

○ 워드클라우드 생성 시 영단어를 포함할 수도 있고 제외할 수도 있는데, "영문미포함", "영문포함" 선택 버튼을 클릭하여 동일한 검색어에 대해 워드클라우드의 영단어 포함 여부를 선택할 수 있음

- 조사분석 데이터 상의 단어들만을 학습한 것으로, 한글 단어만을 원할 경우 영문미포함을 선택하고, 영어로 된 용어들을 검색하고싶을 경우 영문포함을 선택



[그림 3-7] 영문포함 워드클라우드 예시

❍ 워드클라우드 분석 시 오타 등 올바르지 않은 단어들이 간혹 등장할 수 있는데, 이는 조사분석 데이터 상에서 연구책임자가 과제 내용 또는 과제명에 오타를 입력했기 때문이며, 향후 동 시스템의 데이터 정제를 통해 해결이 필요

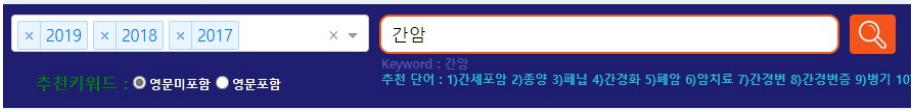
- 전문용어의 비중이 높은 과학기술 과제 정보데이터의 특성 상 일반적인 사전을 이용한 일괄적 오타수정이 불가능하기 때문에, 현재는 오타 및 띄어쓰기 오류 등에 대한 전처리는 수행하지 않은 상태임

□ (1.3. 과제 검색 및 연구비분석) 원하는 연도 범위에서 과제를 검색하고 사업별/부처별/연구단계별/연도별 통계 및 클러스터링 분석을 할 수 있는 기능임

❍ 상단에 연도 범위를 선택할 시, 여러 연도를 선택할 수 있으며 1개년만 선택할 수도 있음

❍ “단어 또는 문장을 입력해주세요...” 칸에 원하는 검색어를 입력하면, 입력칸 하단에 추천 키워드가 출력되는데, 이는 워드클라우드 검색결과(유사도가 높은 단어)를 출력함으로써 검색어를 보다 구체화 하고싶을 때 활용할 수 있음

- 영문미포함, 영문포함 선택 버튼은 워드클라우드 기능과 동일한 기능을 수행함



[그림 3-8] 연도 선택 및 검색어 입력, 추천키워드 출력 예시

❍ 검색어를 입력하고 우측의 돋보기 버튼을 클릭하면 화면 중앙에 로딩 아이콘(원형의 회전하는 그림)이 출력되고, 분석이 완료되면 로딩 아이콘이 사라지고 표에 검색 결과 과제들이 출력됨

- 입력한 검색어와 유사도가 높은 과제를 연도별로 1,000개씩 출력하며, 연번(유사도 등수), 과제수행년도, 유사도, 부처명, 사업명, 내역사업명, 정부연구비, 과제명, 연구목표, 연구내용, 연구개발단계를 출력함
  ※ ‘rowid’ 열은 내부 연산을 위한 값으로 사용자 입장에서는 의미없는 값임
- 출력 결과는 동일한 방법으로 정렬, 필터링이 가능하며, 좌측 하단의 데이터 다운로드 버튼으로 csv 형식으로 내려받을 수 있음
- 행 삭제 기능은 추후 제거 예정인 기능으로 사용하지 않는 것을 권장하며, 데이터 다운로드 후 엑셀에서 삭제 요망

[그림 3-9] 로딩 아이콘

	연번	과제수행년도	유사도	부처명	사업명	내역사업명	정부연구비(원)	과제명-국문	요약문_연구목표	요약문_연구내용	연구개발단계	rowid
	1	2019	0.5081	교육부	개인기초연구(교·	기본연구(1년~3년	50000000	간암 종양미세환	1) 간암 세포와 ?	1차년도 간암 CA	기초연구	492931
	2	2019	0.505	과학기술정보통신	바이오의료기술?	염상의과학자 연	75000000	혈장 엑소좀 오믹	혈장 엑소좀 오믹	1단계 1년차 간암	응용연구	483925
	3	2019	0.502	보건복지부	포스트게놈신산업	인간 유전체 이형	30000000	HLA A 펩티돔 분	한국의 호발암인	연구 내용은 크게	기초연구	511407
	4	2019	0.501	교육부	개인기초연구(교·	기본연구(1년~3년	40000000	간암모델에서 gly	간암 모델에서 m	○ Metformin di	기초연구	491622
	5	2019	0.497	보건복지부	포스트게놈신산업	인간 유전체 이형	37000000	암유전체 분석을	최근 Immune ch	암유전체 분석은	기초연구	511393
	6	2019	0.4939	교육부	개인기초연구(교·	기본연구(4년~5년	37500000	B형 간염 간세포'	B형 간염 유래 간	연구의 필요성 Cl	기초연구	492422
	7	2019	0.4864	과학기술정보통신	개인기초연구(과·	재도약연구(신진·	30000000	간세포암에서 원	발암 및 암전이 ;	연구내용 및 범우	기초연구	480133
	8	2019	0.486	과학기술정보통신	개인기초연구(과·	신진연구	100000000	간암의 면역관문	본 연구에서는 먼	마우스 syngenei	기초연구	477380
	9	2019	0.4772	과학기술정보통신	바이오의료기술?	연구수요기반유전	420000000	eRNomics 기반 :	한국인 간암 발병	1차년도 간암 세:	기초연구	484350
	10	2019	0.4745000000000	과학기술정보통신	개인기초연구(과·	중견연구(총연구·	100000000	간암발생에서 암	암줄기세포는 암	1차년도 HGF ME	기초연구	472049
	11	2019	0.4711	교육부	이공학학술연구?	지역대학우수과학	50000000	대장암에서 현미	대장암에서 현미	본 연구의 실험군	기초연구	498427
	12	2019	0.4710000000000	과학기술정보통신	개인기초연구(과·	중견연구(총연구·	25001000	크로마틴 구조조	크로마틴 3차원 '	1차년도 간암 발·	기초연구	475551

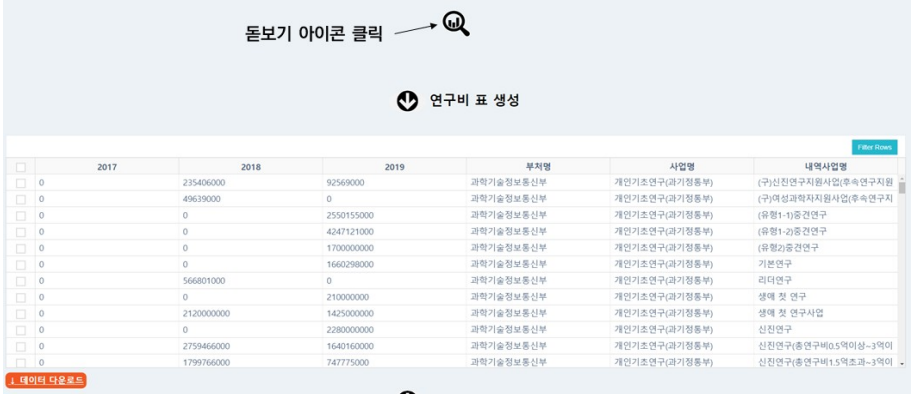
↓ 데이터 다운로드

[그림 3-10] 과제 검색 및 데이터분석 결과 예시(검색어 : 간암)

❍ 검색이 이루어진 상태에서 화면 하단으로 스크롤을 조금 내려 돋보기 모양 아이콘을 클릭하면, 연도 및 내역사업 단위로 검색 결과 과제들의 정부연구비 통계를 출력함

- 엑셀의 피벗테이블과 유사한 기능이며, 연도에 따른 사업명 변화는 반영하였지만 부처명이 변경된 경우에는 다른 부처로 인식하는 점을 주의할 필요

※ 연도가 달라지면서 조사분석 데이터 상에서 동일한 사업의 이름이 조금씩 바뀌는 경우는 통일시켰으나, 부처명은 조직체계의 변화 등 단순히 이름만 바뀌는 경우가 아닐 수 있으며, 당시의 부처명 자체가 의미가 있으므로 조정하지 않음(예: 미래창조과학부와 과학기술정보통신부는 동 시스템 내에서 다른 부처로 인식함)

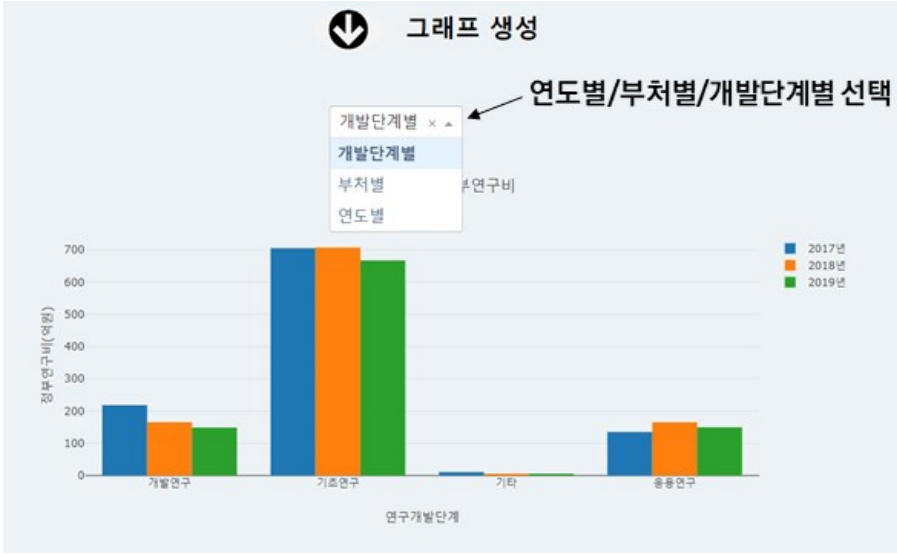


2017	2018	2019	부처명	사업명	내역사업명
0	235406000	92569000	과학기술정보통신부	개인기초연구(과기정통부)	(구)신진연구지원사업(후속연구지원
0	49639000	0	과학기술정보통신부	개인기초연구(과기정통부)	(구)여성과학자지원사업(후속연구지
0	0	2550155000	과학기술정보통신부	개인기초연구(과기정통부)	(유형1-1)중견연구
0	0	4247121000	과학기술정보통신부	개인기초연구(과기정통부)	(유형1-2)중견연구
0	0	1700000000	과학기술정보통신부	개인기초연구(과기정통부)	(유형2)중견연구
0	0	1660298000	과학기술정보통신부	개인기초연구(과기정통부)	기본연구
0	566801000	0	과학기술정보통신부	개인기초연구(과기정통부)	리더연구
0	0	210000000	과학기술정보통신부	개인기초연구(과기정통부)	생애 첫 연구
0	2120000000	1425000000	과학기술정보통신부	개인기초연구(과기정통부)	생애 첫 연구사업
0	0	2280000000	과학기술정보통신부	개인기초연구(과기정통부)	신진연구
0	2759466000	1640160000	과학기술정보통신부	개인기초연구(과기정통부)	신진연구(총연구비0.5억이상~3억이
0	1799766000	747775000	과학기술정보통신부	개인기초연구(과기정통부)	신진연구(총연구비1.5억초과~3억이

↓ 데이터 다운로드

[그림 3-11] 연구비 표 생성 예시

❍ 스크롤을 더 내리면 개발단계별, 부처별, 연도별 연구비 통계를 출력하는 화면이 나오며, 메뉴에서 원하는 기준을 선택하면 아래에 자동으로 그래프가 생성됨



[그림 3-12] 그래프 생성 예시

❍ 화면 최하단부에는 클러스터링 분석을 할 수 있는 기능이 구성되어 있으며, 위쪽의 작은 상자에 원하는 클러스터(군집) 수를 입력하고 돋보기 버튼을 클릭하면 연도별 클러스터링 분석이 수행됨

- 분석에 시간이 걸리는 과정으로, 화면 중앙의 로딩 아이콘이 사라질 때까지(10~20초 내외) 기다려야 함
- 분석 과정에서는 위쪽에서 검색한 연도별 1,000개씩의 과제들을 서로 유사성이 높은 클러스터들로 분류하는 작업이 수행되며, 사용자가 입력한 군집 수는 각 연도별로 적용됨

  ※ 예를 들어, 2개 연도에 5개의 군집 수를 입력하면 총 2×5=10개의 군집으로 검색 결과들이 분류되는 것임

❍ 분석이 완료되면 "학습이 완료되었습니다. 군집들을 선택하고 아래 버튼을 눌러주세요."라는 문구가 출력되며, 그 아래의 흰색 상자를 클릭하면 연도 또는 연도(클러스터번호) 형식의 선택 항목들이 나타남

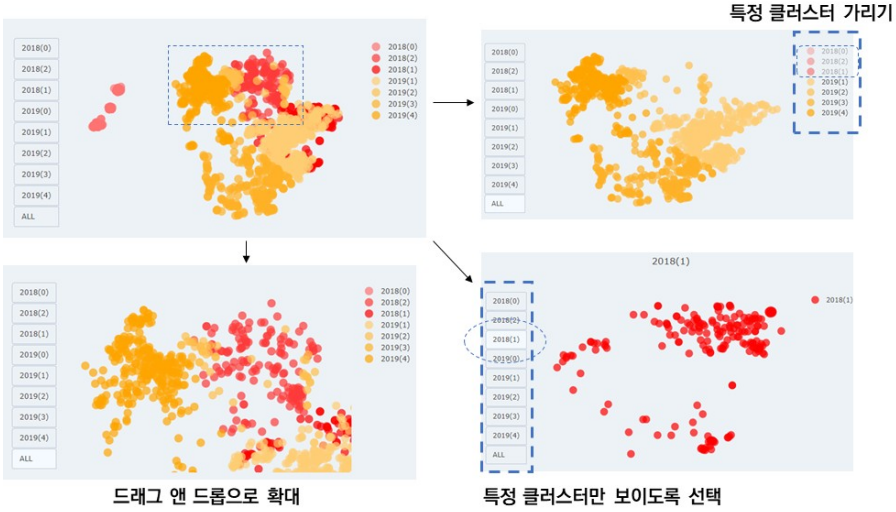
- 연도를 선택하면 그 연도의 클러스터 전체를 하단에 시각화하며, 특정 연도의 특정 클러스터만을 선택할 수도 있음

- 연도 또는 클러스터를 선택할 후, 2개가 나란히 위치하는 돋보기 버튼 중 왼쪽 버튼을 클릭하면 아래에 과제들이 클러스터별로 시각화됨
- 각 연도는 색깔로 구분되며, 같은 연도 내에서 클러스터들은 명도로 구분됨
- 각 점들은 과제들을 나타내며, 각 점에 마우스 커서를 올릴 시 어떤 과제인지 표시됨

[그림 3-13] 연도별 클러스터링 분석 수행 예시

❍ 그림을 확대하고 싶을 경우, 확대하고자 하는 부분을 마우스 드래그 앤 드롭으로 선택하면 되며, 원상태로 돌아가고 싶을 때는 그래프 위 아무 지점을 더블클릭하면 됨

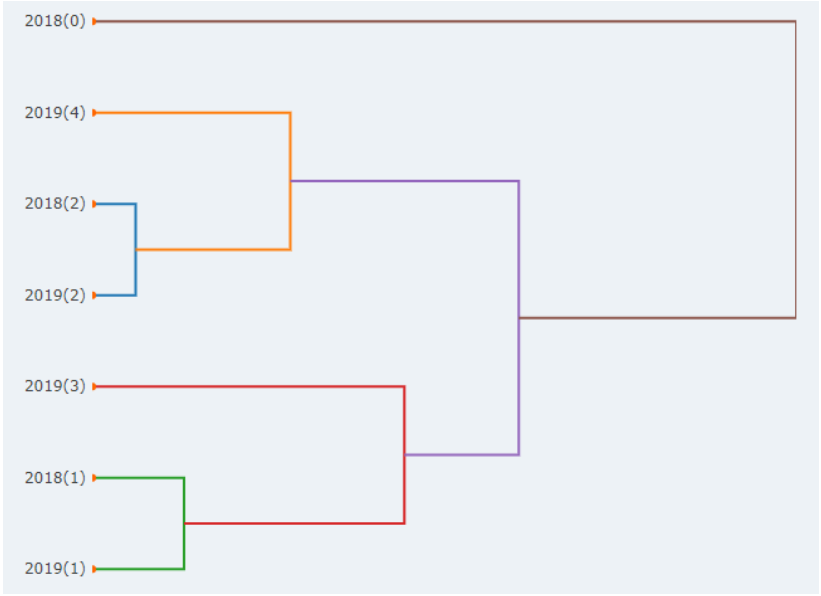
- 특정 클러스터만을 보고싶을 때는, 좌측의 버튼 중 하나를 클릭하면 되며, 특정 클러스터를 가리고 싶을 때는 우측의 클러스터 이름을 클릭하면 됨



[그림 3-14] 클러스터링 그림 확대, 특정 클러스터 선택 방법

❍ 오른쪽 돋보기 버튼을 클릭하면, 여러 클러스터들을 서로 가까이 있는 (유사도가 높은) 클러스터끼리 묶어주는 계층도(dendrogram)를 생성함

- 아래 그림의 예시에서는, 2018년 2번 클러스터와 2019년 2번 클러스터, 2018년 1번 클러스터와 2019년 1번 클러스터가 서로 가장 가깝고, 2019년 4번 클러스터와 2018/2019년 2번 클러스터, 2019년 3번 클러스터와 2018/2019년 1번 클러스터가 그 다음으로 서로 가까운 것을 보여줌

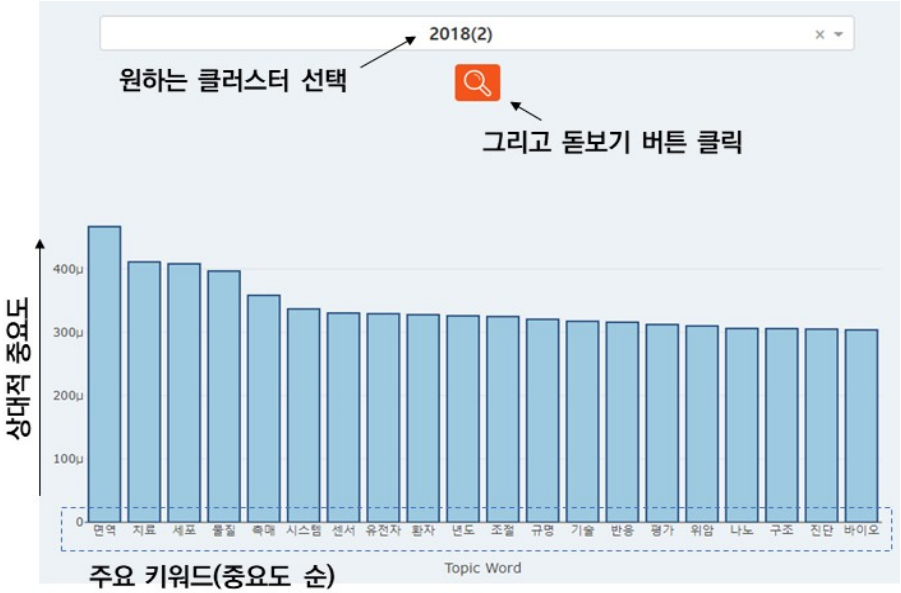


[그림 3-15] 클러스터 계층도 예시

❍ 시각화가 완료되면, 클러스터별로 주요 키워드를 분석할 수 있으며, 스크롤을 더 하단으로 내려 “위에서 선택한 군집과 반경을 정해주세요” 문구가 써있는 칸을 클릭하여 원하는 클러스터를 선택하고 그 아래 돋보기 버튼을 클릭하면 주요 키워드가 출력됨

※ (주의) 시각화를 위한 연도 또는 클러스터 선택 시, 1개 연도만을 선택할 경우 오류로 인해 이 단계에서 선택할 수 있는 클러스터가 드롭다운 메뉴에 나타나지 않으므로, 시각화 시 2개 이상의 연도를 선택하거나, 개별 클러스터들을 지정해주어야 함

- 주요 키워드 출력 시, 클러스터링 그래프에서 특정 부분을 확대하여 돋보기 버튼을 다시 클릭하면 현재 확대되어 표시되는 부분에 포함된 과제들(점들)에 한해서만 주요 키워드를 출력함

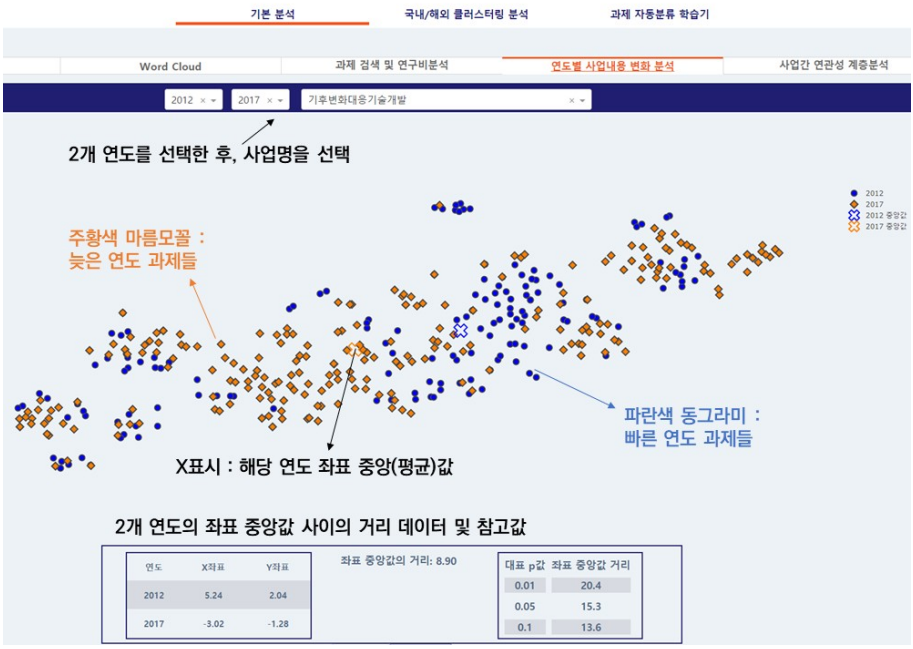


[그림 3-16] 클러스터별 주요 키워드 출력 예시

□ (1.4. 연도별 사업내용 변화 분석) 특정 사업의 과제들이 연도 간에 얼마나 변화했는지 분석하는 기능으로, 2개의 연도가 선택되어 있는 칸에 원하는 2개의 연도를 선택하면 사업을 선택할 수 있음

○ 선택한 2개의 연도에 수행과제가 둘 다 존재하는 사업만을 선택할 수 있으며, 사업을 선택하면 잠깐의 로딩 후 아래에 과제 시각화 및 각 연도별 좌표 평균값, 좌표 평균값 간의 거리가 출력됨

- 만약 2개 연도 수행과제 개수의 합이 50 미만이면 2차원 차원축소 과정에서 과제들이 서로 매우 멀게 분산되는 현상이 있어 분석 결과가 왜곡되므로 “과제 수가 너무 적어 분석이 불가합니다. 다른 사업을 선택해주세요.” 라는 메시지가 출력되며 분석이 수행되지 않음



[그림 3-17] 연도별 사업내용 변화 분석 예시

❍ 하단에 출력되는 좌표 중앙값의 거리는 선택한 사업의 2개 연도 간의 내용 차이를 나타내는 대푯값으로, 거리가 멀수록 내용 차이가 크다고 볼 수 있음

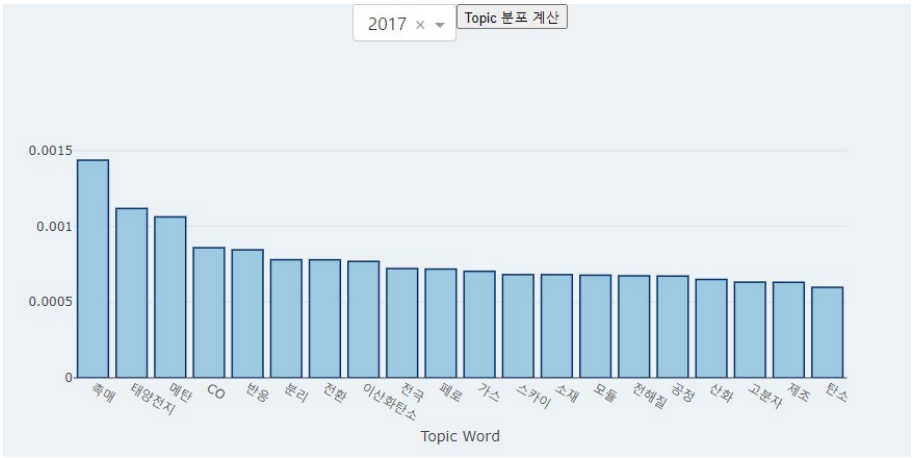
- 오른쪽의 좌표 중앙값 거리에 따른 대표 p값은 중심극한정리에 따른 통계적 유의미성으로, 좌표 중앙값 거리가 13.6(p=0.1) 이상이면 2개 연도 간 통계적으로 내용이 유의미하게 다를 가능성이 높고, 20.4(p=0.01) 이상이면 그 가능성이 매우 높다고 볼 수 있음

※ 해당 내용의 상세 이론은 전년도 연구보고서(유거송 외(2020), 바이오·의료분야 지능형 연구개발정보데이터 분석시스템의 예산배분·조정 활용기법 연구)를 참조

❍ 좌표 중앙값 거리 박스 하단에는 주요 키워드 출력 기능이 구성되어 있으며, “Select…”라고 써있는 박스를 클릭하여 2개 연도 중 하나를 선택한 후 “Topic 분포 계산” 버튼을 클릭하면 해당 연도 과제들의 주요 키워드가 출력됨

- 상단의 과제 시각화 그래프의 특정 부분을 확대(드래그 앤 드롭)하거나 축소(그래프 위 아무 위치나 더블클릭)할 수 있으며, 확대된 상태에서 “Topic 분포 계산” 버튼을 클릭할 시 확대되어 현재 보이는 과제(점)들의 주요 키워드가 분석됨

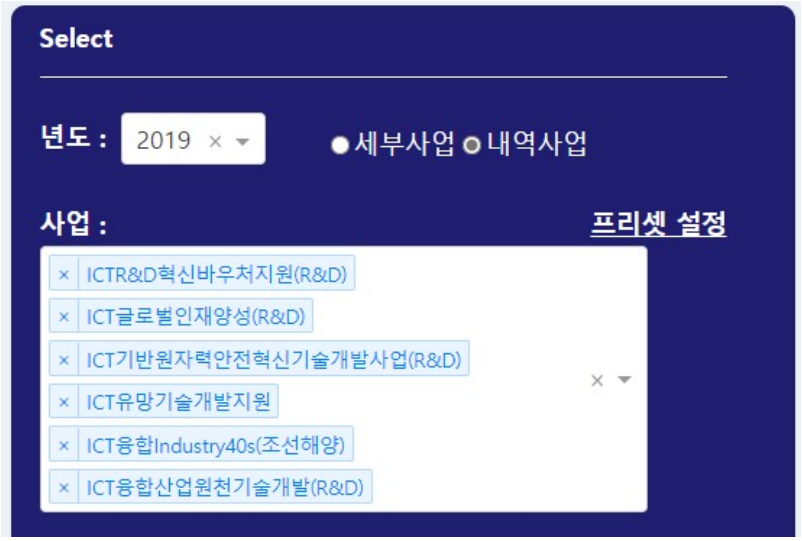
  ※ (주의) 전체 과제의 주요 키워드를 분석하고 싶을 경우 전체 과제들이 포함되도록 드래그 앤 드롭을 한 상태에서 “Topic 분포 계산” 버튼을 클릭해야 분석이 실행됨



[그림 3-18] 연도별 주요 키워드 분석 예시

□ (1.5 사업간 연관성 계층분석) 세부사업 또는 내역사업들 간의 내용적 유사성을 분석하여 사업군을 분류하고 각 사업군의 주제를 볼 수 있는 기능으로, 특정 연도 사업들을 분석할 수 있음

❍ 좌측 짙은 파란색으로 구성된 메뉴에서 연도를 선택하고, 세부사업 단위로 분석할지 내역사업 단위로 분석할지 여부를 선택하면, 해당 연도에 분석 가능한 사업들의 목록이 표시되며 마우스 클릭 또는 타이핑으로 여러 사업들을 선택할 수 있음



[그림 3-19] 사업 선택 예시

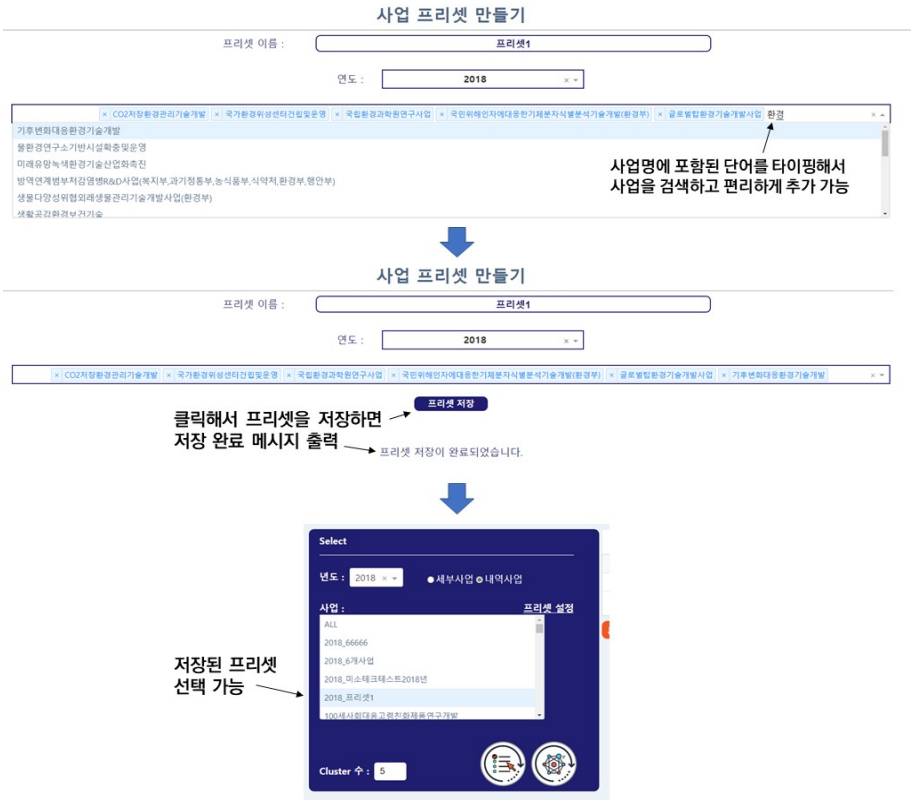
❍ 다수의 사업을 반복적으로 분석하고자 할 경우를 대비하여 동 시스템에는 프리셋 저장 기능을 지원하여 번거로운 사업 선택 과정을 간소화할 수 있음

- 사업 선택 칸 위의 “프리셋 설정” 링크를 누른 후 프리셋 이름을 입력하고, 연도 및 사업을 선택한 후 프리셋 저장 버튼을 클릭하면 “프리셋 저장이 완료되었습니다.”라고 출력되며,

  ※ (주의) 프리셋 이름에는 특수문자나 “_”(언더스코어) 문자가 포함될 경우 오류를 일으키므로 포함시키지 말 것

- 그 상태에서 다시 사업간 연관성 계층분석 메뉴로 들어가면 저장한 프리셋을 사업 선택 칸에서 불러올 수 있으며, 프리셋을 지정한 연도를 선택하면 아래 사업 목록에서 “연도_프리셋이름” 형식으로 저장되어 있는 것을 확인할 수 있음

- 프리셋과 다른 사업 또는 프리셋을 추가로 선택하면 프리셋에 포함된 사업들과 추가로 선택한 사업들이 함께 분석됨



[그림 3-20] 사업 프리셋 지정 방법

❍ 분리할 사업군(클러스터)의 개수를 “Cluster 수” 칸에 입력하고, 두 개의 동그란 버튼 중 왼쪽 버튼을 누르면 각 사업군의 주요 키워드가 출력되며, 그 다음 오른쪽 버튼을 누르면 사업군 별로 계층도(덴드로그램)가 생성됨

- 계층도 바로 위의 흰색 박스에서 클러스터 번호를 선택하면 해당 클러스터의 계층도가 그려짐
- 계층도의 사업 이름들을 보고 클러스터에 포함된 사업들을 알수 있으며, 해당 사업들의 주요 키워드를 상단부의 표에서 확인하면 됨
- 선택한 사업들을 처음으로 분석할 때는 사업군 개수를 1로 입력하여 전체의 계층도를 관찰하고 몇 개의 사업군으로 나누는 것이 적당할 지를 판단한 후, 사업군을 입력하는 것을 추천

- 내역사업명으로 분석할 시 계층도에서 사업명은 “내역사업명(세부사업명)” 형식으로 표시됨

[그림 3-21] 사업군별 주요 키워드 출력 예시

[그림 3-22] 사업간 연관성 계층분석 예시

□ (2. 국내/해외 클러스터링 분석) 조사분석 과제와 PubMed 등재 초록을 검색하고 클러스터링 분석을 수행할 수 있는 기능으로, 좌측은 조사분석 과제, 우측은 PubMed 논문을 분석하여 상호 비교가 가능하도록 구성됨

○ 조사분석 과제와 PubMed 논문 분석 방법은 동일하므로 조사분석 과제를 기준으로 매뉴얼을 작성함

❍ 상단에서 연도를 선택하고 원하는 클러스터 수, 검색 결과 수(검색어와 유사도가 높은 몇 개의 과제를 분석할 것인지) 및 검색어를 입력한 후, “분석 시작” 버튼을 클릭하면 분석이 수행됨

❍ T-SNE 파라미터 설정 기능은 검색 결과 과제(또는 논문)를 시각화하고 클러스터링 할 시 2차원으로 차원 축소하는 알고리즘(T-SNE)의 파라미터를 변경할 수 있는 기능임

- 일반적으로 사용하는 값으로 설정되어 있으나, 분석 결과의 품질(클러스터링이 잘 되는지)을 높이기 위해서 변경이 필요하다고 사용자가 판단할 시 변경할 수 있도록 구성함
- Perplexity 파라미터는 5~50 사이의 값을 사용하도록 권고되며, 이 값이 작을수록 알고리즘은 보다 근시안적으로 차원축소를 수행하게 됨[4)]

  ※ 수백차원의 고차원 데이터를 2차원 또는 3차원의 저차원으로 변환하면서, T-SNE 알고리즘은 어떠한 벡터값들이 서로 가까이 있는지의 정보를 보존하면서 새로운 저차원의 벡터값으로 변환하는데, perplexity 값이 작으면 가까움의 기준을 짧게 잡으므로, 만약 데이터가 커다란 2개의 그룹으로 느슨하게 나눠지는 형태라면 그 형태가 무시되고 작은 여러개의 그룹으로 데이터가 잘못 나눠지게 됨
- Perplexity 값을 선정하는 일률적인 법칙은 존재하지 않으므로 통상적인 값(30 정도)을 사용하는 것이 대부분의 경우에 적절하고, 데이터의 특성에 대해 알고있을 경우 조정하는 것이 타당함
- Iteration 파라미터는 일종의 기계학습인 T-SNE 알고리즘의 반복 횟수로, 충분히 수렴할 수 있도록 1,000 이상의 값으로 설정하는 것이 좋음
- Learning Rate 파라미터는 T-SNE 알고리즘이 반복을 수행할 때마다 결과값들을 얼마나 많이 조정할지를 결정하는 값으로, 너무 크면 결과값이 수렴하지 않고 진동하거나 발산하게 되며, 너무 작으면 수렴하기 위해서 매우 큰 iteration 값을 필요로 하게 되므로, 이를 조정하고자 할 경우 iteration 파라미터를 같이 조정하는 것을 권장

4) How to Use t-SNE Effectively(https://distill.pub/2016/misread-tsne/)

[그림 3-23] 클러스터링 분석을 위한 과제 검색어 입력 예시

❍ "분석 시작" 버튼을 클릭하면 위에서 입력한 검색어와 코사인유사도가 높은 지정한 개수의 과제들이 출력되며, 이는 동일한 방법으로 정렬·필터링하거나 엑셀파일 형식(csv)으로 내려받을 수 있음

- 동 기능에서는 검색 결과 과제들의 과제고유번호와 클러스터 번호를 표시해 주기 때문에, 연구수행단계나 연구개발단계 등 동 시스템에서 제공하지 않는 추가적인 과제정보(6T, 과학기술표준분류 등)를 포함하여 분석하고자 할 경우 NTIS 또는 K2Base에서 해당 데이터를 획득하고 연계하여 분석하면 됨

[그림 3-24] 과제 검색결과 예시

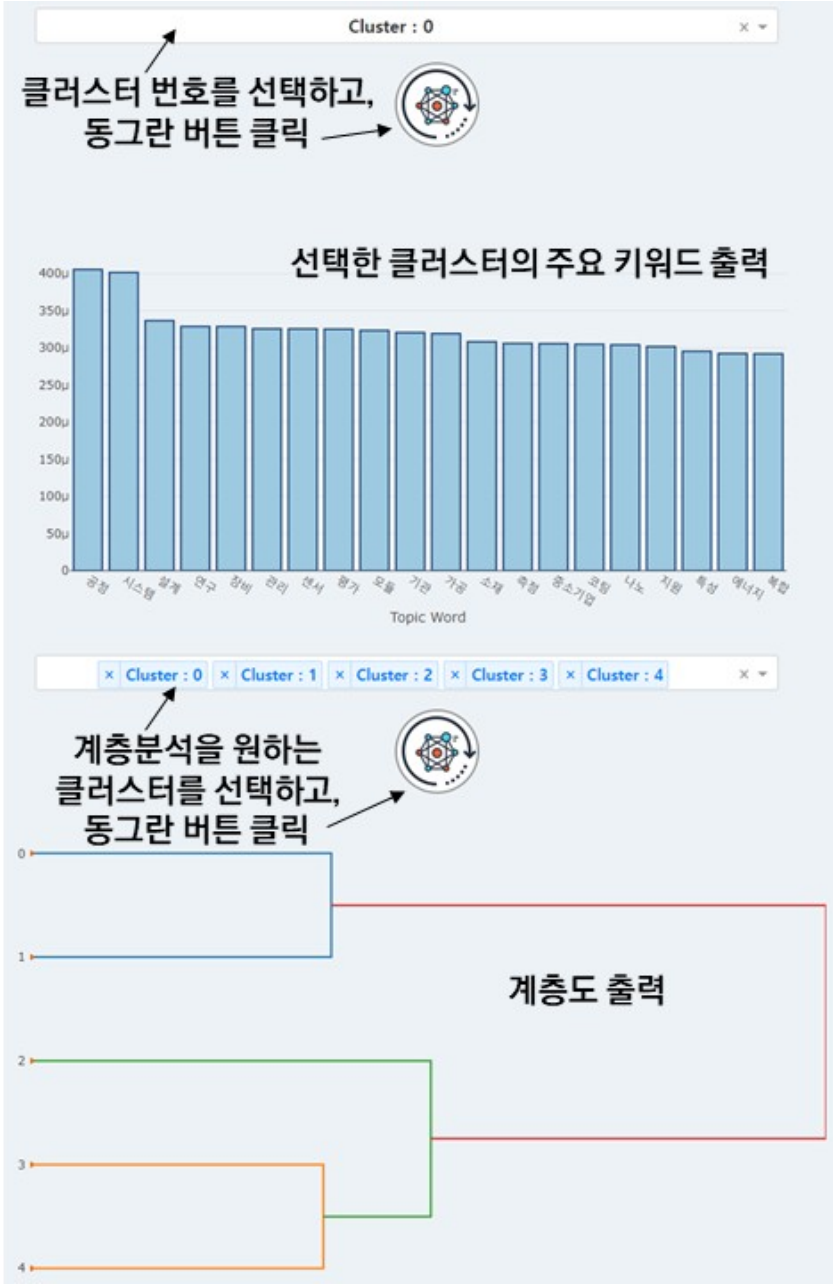
❍ 검색결과 표 아래에는 검색결과의 클러스터링 결과가 시각화되어 표시되는데, '1.3. 과제 검색 및 연구비분석' 기능과 동일한 조작법으로 특정 클러스터만 표시하거나 확대/축소가 가능함

- 마우스 커서를 그림 위에 올리면 마우스 커서의 x좌표가 같은 점들의 과제 내용을 표시해주며, 이를 이용해 각 클러스터 과제들의 내용이 무엇인지 살펴볼 수 있음

[그림 3-25] 클러스터링 그림 예시

❍ 클러스터링 그래프 아래의 흰색 바에서 클러스터 번호를 선택하고, 동그란 버튼을 클릭하면 각 클러스터의 주요 키워드가 출력됨

❍ 주요 키워드 그래프 아래의 흰색 바에서 클러스터들을 선택한 후 동그란 버튼을 클릭하면 클러스터 간의 거리에 따른 계층도가 생성됨



[그림 3-26] 클러스터별 주요 키워드 및 클러스터간 계층도 생성 예시

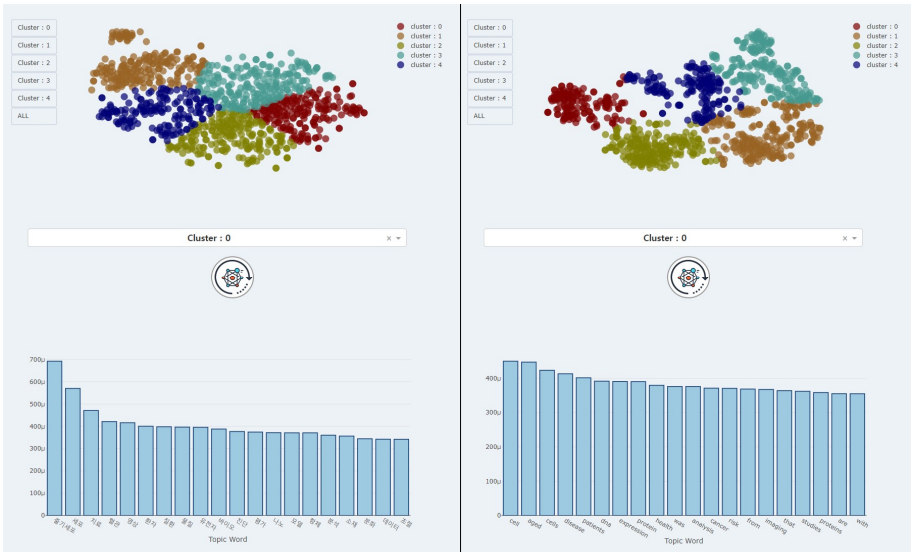
❍ 동 분석은 PubMed 논문초록 데이터에도 동일한 절차를 거쳐 수행할 수 있으며, 이를 통해 동일한 주제의 조사분석과제와 국제논문 검색 결과를 비교해볼 수 있음

- 예를 들면, “줄기세포”와 “stem cell”을 각각 조사분석과 PubMed에서 검색하여 경향을 비교할 수 있음

  ※ PubMed는 바이오 분야의 논문 데이터베이스이므로 국내외 비교분석은 바이오 분야에 한정하여 가능함

- PubMed 초록 검색 결과에서는 발간 연월, 저자 소속기관, 제목, MeSH heading, 초록, 태그값, 클러스터 번호를 제공하는데, 태그값은 내부 연산을 위한 값으로 필요한 정보는 아님

  ※ MeSH heading은 PubMed에서 제공하는 키워드임



[그림 3-27] 조사분석, PubMed 클러스터링 분석 예시

□ (3. 과제 자동분류 학습기) 동 기능은 사용자가 기존에 직접 분류해놓은 과제 데이터의 분류기준을 인공지능이 학습하여 새로운 과제를 입력하였을 때 자동으로 분류해줄 수 있도록 학습 모델을 구축하는 기능으로, 학습 데이터와 테스트 데이터를 업로드하여 사용하도록 구성되어 있음

❍ 좌측 상단의 “Train Set” 칸에 학습 데이터를 업로드하거나 좌측 메뉴 중 “Train data”에서 기존에 학습한 적 있는 데이터를 불러와야 함

- 학습 데이터는 아래와 같은 형식으로 엑셀로 작성하여 csv 형식으로 저장되어 있어야 함

※ (주의) 학습 데이터 및 테스트 데이터에는 과제수행년도, 과제명-국문, 과제고유번호, 사업명, 내역사업명, 정부투자연구비, 요약문_연구목표, 요약문_연구내용 컬럼(열)이 포함되어 있어야 하며(순서는 상관 없음), 컬럼이 누락되거나 컬럼명이 틀렸을 경우 에러가 나므로 주의해야 함

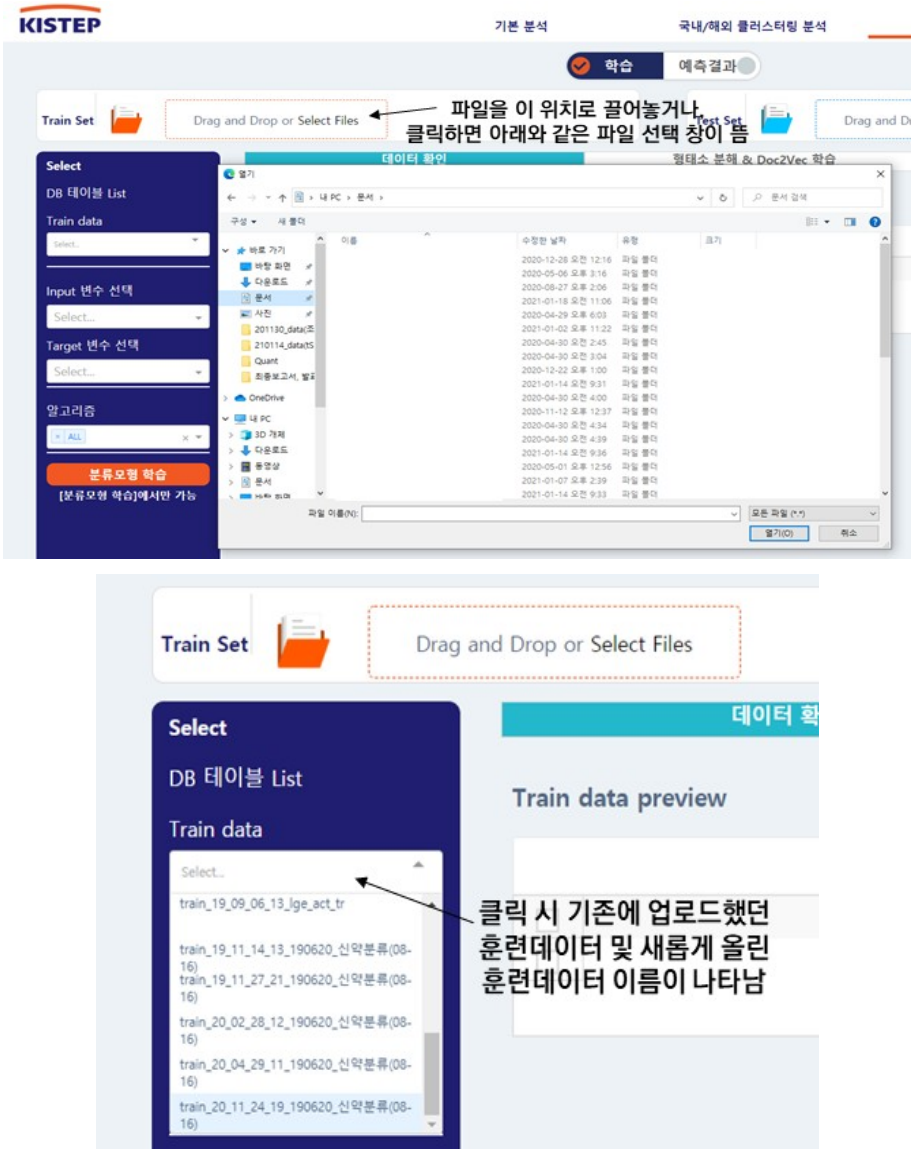
	정보1	정보2	정보3	…	정보m	분류
과제1						
과제2						
과제3						
:						
과제n						

정보들은 분류를 위해 학습하는 내용(과제명, 연구내용, 연구목적 등), 분류는 항목들이 기존에 분류되어있는 답(태그) 값임

[그림 3-28] 학습 데이터 예시

- 업로드한 학습 데이터는 Train data 아래의 칸에서 선택할 수 있으며, 이 때 과거에 학습을 위해 업로드한 적이 있는 데이터를 선택하여 불러올 수도 있음

※ 학습데이터 업로드 시 “train_업로드시간_파일이름” 형식으로 저장됨



[그림 3-29] 학습데이터 업로드 및 선택 방법

❍ 학습 데이터를 선택하면 우측에 학습 데이터의 내용 중 일부가 표시되며 (Train data preview), Input 변수와 Output 변수, 그리고 분류모형에 사용할 알고리즘을 선택할 수 있음

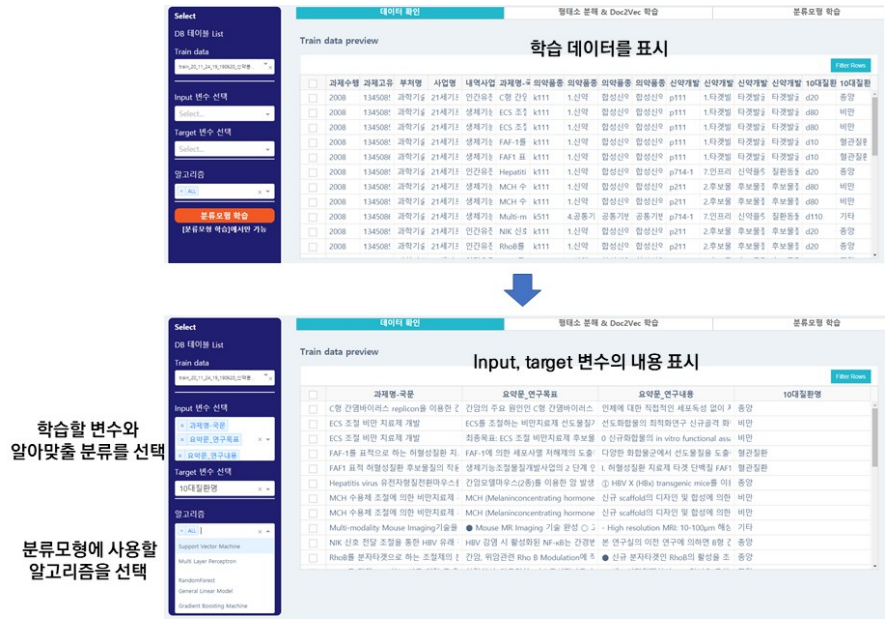
- Input 변수는 인공지능이 어떤 항목의 내용으로 학습하기 위한 피쳐(feature)이며, target 변수는 인공지능이 알아맞히는 항목으로 여기서는 기존에 사용자가 분류해놓은 분류 값이 됨

  ※ Input 변수는 여러개 지정할 수 있으며, target 변수는 하나만 지정할 수 있음

- Input과 target 변수를 지정하면 우측에 학습 데이터 내용을 일부 표시하던 부분 중에서 사용자가 input과 target 변수로 지정한 컬럼(열)들만 표시됨

- 분류모형에 사용할 알고리즘은 총 5가지*가 있으며, 원하는 알고리즘을 선택하거나, 기본값인 ALL을 선택하여 5가지 알고리즘을 모두 사용하여 결과를 비교할 수 있음

  * Support vector machine(SVM), Mulilayer perceptron(MLP), Gradient boosting machine(GBM), Generalized linear model(GLM), Random forest(RF)



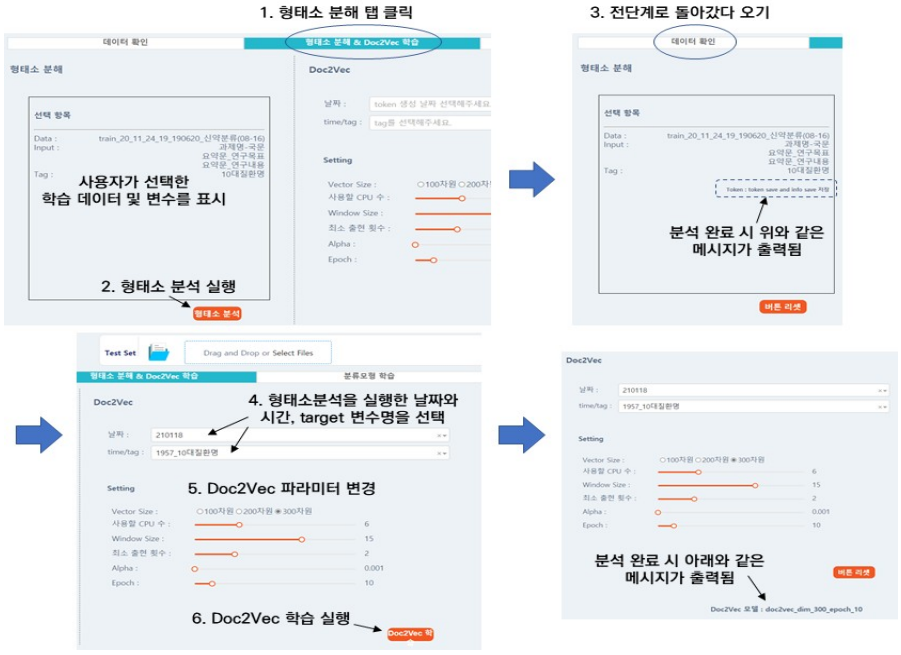
[그림 3-30] Input, target 변수 선택 예시

❍ 이 단계에서 “형태소 분해 & Doc2Vec 학습” 탭을 클릭하여 다음 단계로 넘어가 형태소 분석을 실행하면 로딩 아이콘이 나타나며, 아이콘이 사라질 때까지 기다린 후 Doc2Vec 학습을 실행함

- 연산량이 많은 단계로 로딩은 수 분 정도 소요될 수 있음
- 형태소 분석이 완료되면 “Token : token save and info save 저장” 이라는 메시지가 출력됨

❍ Doc2Vec 학습은 형태소 분석이 완료된 학습 데이터(input, target 변수)를 벡터화하여 분류 알고리즘이 학습할 수 있는 형태로 전환하는 과정임

- 날짜와 time/tag는 사용자가 형태소 분석을 실행한 날짜와 시간, target 변수명으로 기록되며, 날짜를 먼저 선택하면 그 날짜에 형태소 분석을 실행한 time/tag들이 나타나므로 사용자가 원하는 항목을 선택하면 됨
- (주의) 새롭게 형태소 분석을 실행했을 경우, 화면 좌측에서 형태소 분석을 실행한 결과가 우측 날짜, time/tag 목록에 바로 반영이 안되는 문제가 있으므로, “데이터 확인” 탭을 클릭하여 전단계로 갔다가 다시 “형태소 분해 & Doc2Vec 학습” 탭을 클릭하여 되돌아오면 날짜와 time/tag 목록에서 새롭게 형태소 분석을 실행한 데이터가 추가된 것을 확인할 수 있음



[그림 3-31] 형태소 분석 및 Doc2Vec 학습 방법

❍ 날짜와 시간, target 변수명을 올바르게 선택하였으면 아래 Doc2Vec 파라미터를 조정한 후 "Doc2Vec 학습" 버튼을 클릭하면 Doc2Vec 학습이 실행됨

- Doc2Vec 파라미터는 현재 일반적인 값으로 설정되어 있어 사용자가 조정할 필요성은 낮지만, 벡터의 크기를 더 작게 바꾸는 등 조정할 수 있는 기능이 구현되어있음

※ Vector Size : 차원수가 높을수록 더 큰 크기의 벡터에 학습 데이터를 대응시키므로, 더 많은 정보가 저장되지만 분석의 속도는 더 느려짐

※ 사용할 CPU 수 : 멀티쓰레딩(multi-threading)을 조절하는 기능으로, 사용자가 조정할 필요는 없음

※ Window Size : 학습 시 매 단어 좌우 몇 개까지의 단어를 학습에 활용할 것인지를 정하는 것으로, 너무 적으면 문맥의 정보가 학습되지 않고, 너무 크면 다른 문장의 단어도 학습에 반영할 가능성이 커지므로 오히려 문맥이 왜곡될 수 있음

※ 최소 출현 횟수 : 데이터에서 최소 출현 횟수 미만으로 등장하는 단어는 제외시키는 파라미터로, 데이터에서 너무 적게 등장하는 단어는 충분히 학습되기도 어렵고 데이터 내에서 가지는 의미도 적기 때문에 적절한 값을 설정할 필요

※ Alpha : learning rate(학습 속도)의 시작값으로, 이 값이 크면 학습이 더 빨리 될 수도 있지만 오히려 수렴하지 않을 가능성도 커지므로 기본값에서 변경하지 않는 것을 추천

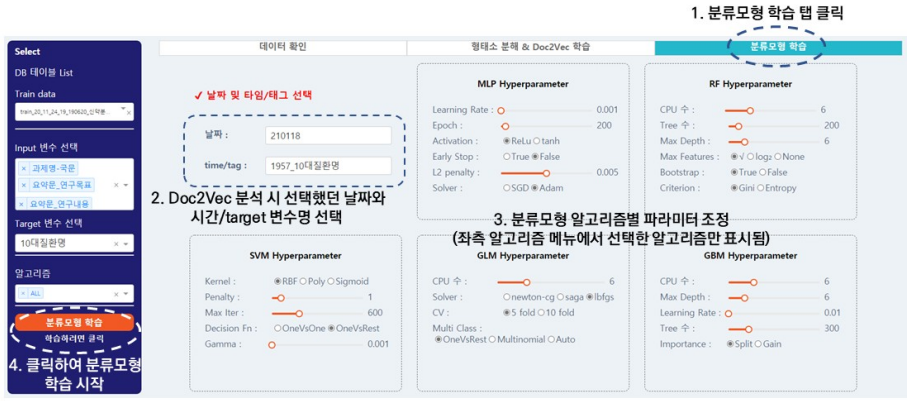
※ Epoch : 학습 데이터를 총 몇바퀴 돌면서 학습을 할 것인지를 지정하는 것으로, 클수록 더 반복학습이 되어 학습 데이터에 대한 정답률이 높아지나, 과적합(overfitting)의 가능성도 생기므로 기본값에서 크게 증가시키지 않는 것을 추천

- Doc2Vec 학습이 완료되면 "Doc2Vec 모델 : doc2vec_dim_300_epoch_10" 과 같은 메시지가 출력됨

❍ 다음 단계는 분류모형 학습으로, "분류모형 학습" 탭을 클릭한 후, Doc2Vec 학습 시 선택하였던 형태소 분석이 이루어진 날짜와 시간, target 변수명을 올바르게 선택하고 "분류모형 학습" 버튼을 클릭하면 됨

- 좌측 짙은 파란색 메뉴의 하단 "알고리즘" 부분에서 지정한 알고리즘들의 파라미터를 조정할 수 있는 메뉴가 우측 화면에 표시됨

- 각 알고리즘의 파라미터는 일반적으로 사용되는 값으로 설정되어 있으며, 학습모형의 분류성능을 높이기 위해 파라미터를 조정할 수 있지만, 각 분류모형 알고리즘에 대해 다루는 것은 본 보고서의 범위를 벗어나므로 기술하지 않음



[그림 3-32] 분류모형 학습 방법

❍ 회전하는 원형 로딩 아이콘이 사라지고 분류모형 학습이 완료되면, 테스트 데이터를 업로드하고, 화면 상단의 “예측결과” 버튼을 클릭하여 학습한 모형의 성능평가 및 실제 신규항목 분류 기능이 구현된 메뉴로 진입

- 새로운 테스트 데이터를 업로드하고자 하면, 우측 상단의 “Test Set” 부분에 학습 데이터를 업로드할 때와 동일한 방법으로 테스트 데이터를 업로드하면 됨
- 테스트 데이터 업로드 시 주의사항은, 학습 데이터와 같은 형식으로 저장되어 있어야 하며, 과제수행년도, 과제명-국문, 과제고유번호, 사업명, 내역사업명, 정부투자연구비, 요약문_연구목표, 요약문_연구내용 컬럼(열)이 포함되어 있어야 함(순서는 상관 없음)
- 테스트 데이터로 분류모형의 정확도를 평가하고자 할 경우에는 해당하는 target 변수 값(컬럼)이 테스트 데이터에 들어있어야 하며, 분류모형이 제공하는 예측치만 필요할 경우에는 없어도 무관함

❍ 화면 상단에서 테스트 데이터와 분류모형을 선택한 후, 이전에 성능평가 또는 분류예측을 한 적이 없는 조합일 경우 “클릭” 버튼을 클릭하여 테스트 데이터의 형태소 분석을 실행함

- 테스트 데이터는 “학습” 메뉴 화면에서 업로드했던 테스트 데이터 뿐만 아니라 이전에 사용했던 모든 테스트 데이터를 활용할 수 있으나, 적용하고자 하는 분류모형에서 사용된 input, target 변수가 포함된 데이터만 사용 가능

- 날짜와 시분/tag는 Doc2Vec 학습 또는 분류모형 학습 시 선택하였던 형태소 분석 날짜 및 시간, target 변수명을 선택하면 해당 학습모형이 불러와짐

[그림 3-33] 테스트 데이터 및 분류모형 불러오기

❍ 분류모형 칸에서는 해당 분류모형 학습 시 선택하였던 알고리즘을 선택할 수 있는데, 선택한 모든 알고리즘의 성능을 비교하고자 할 경우 “ALL”을 선택하고(ROC curve와 accuracy만 표시됨), 각 알고리즘이 예측한 과제별 분류 예측치를 보고싶을 경우 개별 알고리즘을 선택하여야 함

- ROC curve란 예측모형의 false positive와 true positive가 얼만큼인지 그래프화하여 보여주는 것으로, 곡선 아래의 면적이 1에 가까울수록(즉, 모양이 정사각형에 가까울수록) 정확히 예측한 것임

  ※ 동 시스템에서는 분류별로 ROC curve를 따로 출력하여, 어떠한 분류를 더 잘 예측하고 어떤 분류가 오답률이 높았는지 분석 가능

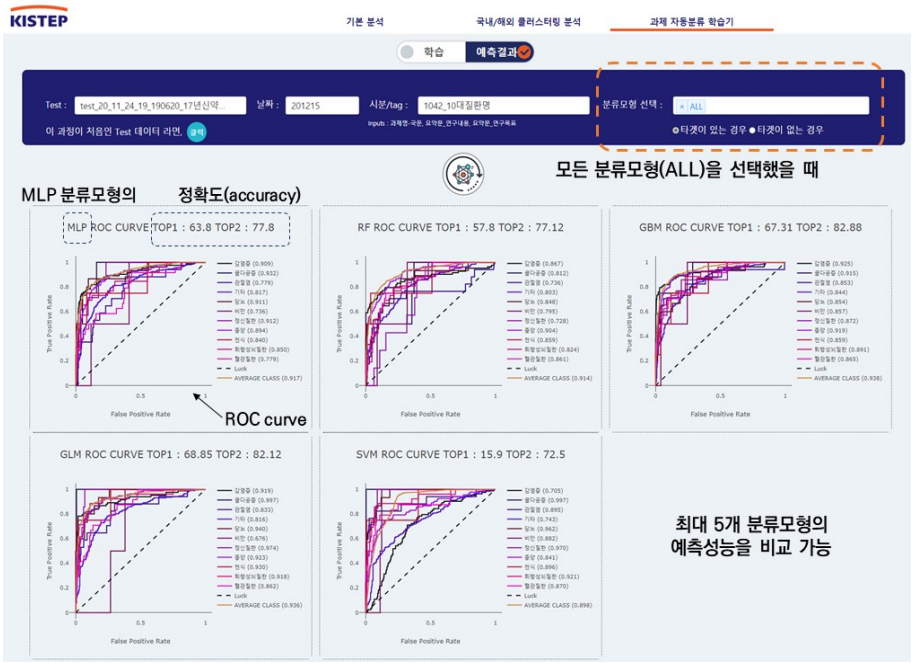
- Accuracy란 테스트 데이터의 총 과제 중에서 몇 개의 과제의 분류를 정확히 예측했는지의 비율로서, 0~100 사이의 값임(%라고 보면 됨)

  ※ TOP1과 TOP2 accuracy를 제시하는데, TOP1은 예측모형이 가장 가능성이 높다고 제시한 분류의 정/오답 여부만을 판단하며, TOP2는 예측모형이 가능성이 높다고 제시한 2개 분류가 둘 다 정답이 아닐 경우만 오답으로 판단하는(더 관대한) 기준임

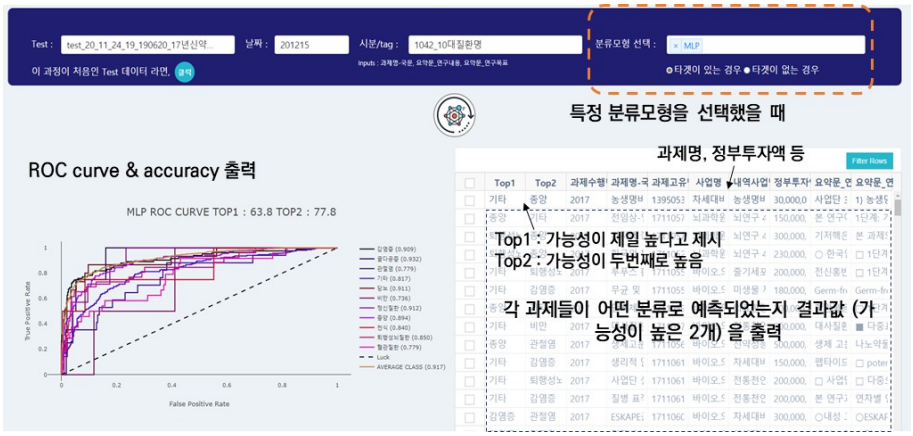
- “ALL”을 통해 모든 분류모형의 성능을 비교하는 것이 아니라, 각 분류모형이 각 과제에 어떤 분류값을 예측했는지 결과를 보고자 할 때는, 분류모형 칸에서 개별 알고리즘을 한 개씩 선택해야 하며, 이 때 좌측에는 ROC curve와 accuracy, 우측에는 예측 결과 표가 출력됨

❍ 분류모형을 선택한 후 동그란 아이콘을 클릭하면 결과값이 출력됨

- 결과값이 모두 출력되지 않을 경우, 동그란 아이콘을 한 번 더 클릭



[그림 3-34] 모든 분류모형의 분류 성능을 비교하는 예시



[그림 3-35] 특정 분류모형의 분류 성능 및 예측 결과값을 출력하는 예시

❍ 분류모형 선택 칸 아래 “타겟이 있는 경우”와 “타겟이 없는 경우”를 선택해야 함

- 타겟이 있는 경우는 테스트 데이터에 실제 분류(답) 정보가 들어있는 경우로, 분류모형이 몇 %나 정확히 분류를 예측하는지 평가할 때에 해당함
- 타겟이 없는 경우는 테스트 데이터에 실제 분류(답) 정보가 없고, 어떤 분류인지 모르는 새로운 데이터를 순수히 분류모형을 이용해 무슨 분류인지 예측하는 것만을 원할 때 선택

※ 타겟이 없는 경우에는 ROC curve나 accuracy를 구할 수 없으므로, 분류모형 선택 시 “ALL”을 선택하면 아무런 결과도 출력되지 않고, 반드시 개별 분류모형(알고리즘)을 선택해야 함

제4장 자연어처리 기반 국내외 바이오헬스 기술개발 동향 비교분석

제1절 데이터 획득 및 분석방법

□ 동 연구에서는 해외 바이오헬스 관련 정부부처 중에서 연구과제 데이터(연구비, 요약문) 확보가 용이한 부처를 대상으로 데이터를 수집하고 지능형 분석시스템에서 차용하고 있는 알고리즘을 유사하게 적용하였음

❍ 해외의 바이오헬스 관련 정부부처는 미국의 NIH(National Institute of Health), NSF(National Science Foundation), 영국의 UKRI(UK Research and Innovation) 등이 있음

- (NIH) 미국 국립보건원으로, HHS(Department of Health and Human Services) 산하 기관 중 보건의료 관련 연구를 총괄하는 기관
- (NSF) 미국 과학재단으로, 우리나라의 한국연구재단과 유사하게 다양한 분야의 기초연구자들을 대상으로 연구비를 지원하는 기관
- (UKRI) 영국 BEIS(Department of Business, Energy and Industrial Strategy) 산하의 영국 연구혁신기구로, 다양한 분야의 연구를 지원하고 있으며 영국 내 정부R&D 중 가장 큰 비중을 차지

❍ 본 연구에서 이용하는 분석방법을 적용하기 위해서는 요약문(abstract) 데이터를 구할 수 있어야 하므로 이를 공개하고 있는 NIH, NSF, UKRI 3개 부처에 대해 연구를 진행하였음

- doc2vec 학습을 위해서는 해당 문서를 설명하는 충분한 양의 텍스트, 즉 요약문 데이터가 필수적임
- 각 부처들은 자체적으로 운영하고 있는 웹사이트에서 원시 데이터(raw data)를 바로 제공하거나, 조회 및 웹크롤링을 허용하여 데이터를 수집할 수 있음

※ 웹크롤링(Web crawling)은 웹브라우저에 표시되는 내용을 자동으로 수집하는 작업으로, 사람이 직접 수집하기 위해서는 방대한 내용을 반복적으로 복사&붙여넣기 해야 하는 인터넷 상의 데이터를 손쉽게 수집할 수 있게 해주는 여러 가지 기법을 칭함

※ 웹사이트의 형태에 따라 다양한 웹크롤링 기법들이 공개되어 있으며, 수집한 데이터 중 사용자가 필요로 하는 데이터만 선별 · 가공하는 코드를 추가하여 효과적으로 활용이 가능

❍ 수집한 데이터 항목은 각 부처 · 기관에서 발주한 최근 5년간(2015~2019) 과제들의 과제번호, 수행연도, 과제명, 연구비, 요약문이며, 이 중 과제명과 요약문을 합쳐 doc2vec 임베딩을 위한 입력값으로 사용하였음

❍ doc2vec 학습 이후 과제 검색 및 군집화를 수행하는 알고리즘은 지능형 분석시스템과 동일하게 진행됨

□ NIH의 데이터 수집은 NIH의 ExPORTER 플랫폼*에서 직접 다운로드를 제공하는 기본데이터(과제번호, 수행연도, 과제명, 연구비, 수행기관, 분류 등) 및 요약문 데이터를 내려받아 결합하였음

* https://exporter.nih.gov/

❍ ExPORTER 플랫폼은 NIH 과제들의 메타분석을 수행하는 다양한 연구자들을 위해 연도별 과제 리스트를 엑셀(csv) 스프레드시트 형식으로 제공하고 있는 사이트임

❍ 기본데이터와 요약문 데이터는 과제번호(Application ID) 값을 이용하여 결합함

❍ 수집된 데이터는 다음 조건을 적용하여 전처리하였음

- 요약문이 없거나, 150자 미만인 과제는 삭제
- 요약문이 여러 셀에 중복적으로 들어있는 경우* 가장 앞쪽에 있는 셀을 취함

  * 요약문(abstract) 외 public health statement 등의 부가정보가 있는 경우로 추정
- 요약문 맨 앞에 "ABSTRACT :" 와 같이 필요없는 문구가 있는 경우 패턴인식을 통하여 최대한 제거함

  ※ “ABSTRACT”, “SUMMARY”, “INFORMATION”, “DESCRIPTION” 등
- 원시 데이터의 오류로 칸이 밀려있거나, 인코딩 문제 등으로 깨진 글자가 있는 경우(“잟”, “쟲” 등의 깨진 문자) 제거

□ NSF의 데이터 수집은 NSF의 과제 검색 사이트*의 Download Awards 메뉴를 이용하여 연도별 xml 원시 데이터를 다운로드하여 엑셀(csv) 형태로 결합함

* https://www.nsf.gov/awardsearch/download.jsp

○ NSF에서는 각 과제 정보를 담고 있는 개별 xml 파일들을 연도별로 압축하여 제공하고 있으며, 해당 xml 파일의 정보는 파이썬(Python)의 xml 모듈(패키지)을 활용하여 읽어들임

※ xml(Extensive markup language)은 정형화된 항목에 대한 정보들을 저장하기 위한 형식 언어 중 하나임

- 각 과제가 개별 xml 파일로 저장되어 있어 각 파일들을 읽어서 데이터프레임에 추가하고 연도별로 스프레드시트에 저장

○ 요약문이 없거나 150자 미만인 과제는 동일하게 제거함

□ UKRI의 데이터 수집은 UKRI 과제 검색 사이트*에서 "*"를 검색하여 출력되는 모든 과제의 목록을 다운로드한 후, 각 과제의 요약문을 웹크롤링을 통해 수집하였음

* https://gtr.ukri.org/

○ UKRI에서 검색결과를 csv 형식으로 내려받은 파일에는 요약문 정보가 없는 대신 요약문을 조회할 수 있는 링크가 제공되므로, 해당 링크로 접속하여 요약문을 수집하는 과정을 자동화할 수 있음

- 해당 웹크롤링 작업은 파이썬의 requests, beautifulsoup 모듈(패키지)을 활용하여 수행함

※ 웹사이트의 robots.txt 및 이용약관 등 웹크롤링 허용 여부 등을 확인 후 진행함

○ UKRI는 두 가지의 요약문(Abstract, Technical Summary)을 제공하고 있어, 두 가지의 요약문을 하나의 글로 합쳐 doc2vec 학습을 위한 요약문으로 사용함

- 단, UKRI는 종료되지 않은 과제는 요약문을 제공하지 않고 있어 종료과제에 대한 데이터만 수집하고 요약문이 미제공되는 과제는 목록에서 제거함

○ UKRI는 연차별 연구비 데이터를 제공하지 않고 총연구비(AwardPounds 또는 ExpenditurePounds)만 제공하므로, 각 과제의 총연구비를 수행기간으로 나누어 연차별 연구비를 산출함(R 활용)

- 수행기간은 시작연도와 종료연도를 기준으로 정하였으며, 시작일 차이에 따른 편차는 무시하고 총연구비를 수행기간으로 나누어 연차별 연구비를 계산함
  ※ 수행기간(년) = 종료년도 − 시작년도 + 1

□ 수집된 데이터는 일괄적으로 필요한 정보(일련번호, 수행년도(회계연도), 과제, 연구비, 요약문)만을 추출하여 doc2vec 학습 및 결과 분석을 위한 정보로 사용함

○ 연도·기관별 과제 수는 NIH가 가장 많았으며, UKRI, NSF 순이었음(표 4-1)

〈표 4-1〉 부처(기관)별 과제 수

연도	NIH	NSF	UKRI
2015	69,551	13,064	20,896
2016	70,076	12,588	19,509
2017	70,600	12,283	19,285
2018	77,728	12,210	18,934
2019	76,104	12,626	18,310
합계	**364,059**	**62,771**	**96,934**

□ 부처별 분석 알고리즘은 다음과 같은 절차를 거치며, 파이썬을 이용하여 프로그래밍 하였음

※ 지능형 분석시스템의 온라인화 용역개발과 별도로 시스템 고도화를 위해 추진하는 사항으로, 분석시스템과 별개로 프로그래밍함

○ 각 과제의 텍스트 데이터(제목+요약문)에 대해 형태소 분석을 실시하고 doc2vec으로 학습하여 200차원 벡터로 임베딩

○ (입력값) 검색어와 결과 개수(n), 클러스터 수(k)를 설정한 후, 검색어를 앞 단계에서 학습된 doc2vec 모델에 통과시켜 벡터를 얻어냄

○ 검색어와 가장 코사인 유사도(cosine similarity)가 높은 n개 과제를 가져오기

※ 코사인 유사도는 두 벡터가 가리키는 방향이 얼마나 비슷한가를 측정하는 지표로서, 1에 가까울수록 유사도가 높고 −1에 가까울수록 정반대의 의미를 지님

※ 코사인 유사도 계산식 : 두 벡터의 내적(dot product) ÷ 두 벡터의 절대값(norm) 곱

❍ 위 단계에서 가져온 n개 벡터를 t-SNE 알고리즘으로 2차원으로 축소함

※ t-SNE(t-Distributed Stochastic Neighbor Embedding) 알고리즘은 높은 차원 수를 가지는 데이터들의 상호 거리(배치) 정보를 유지하면서 시각화가 가능한 수준(흔히 2차원)으로 차원 수를 낮추는 알고리즘

❍ 2차원으로 축소된 상태의 벡터들을 k-평균 군집화(k-means clustering) 알고리즘으로 k개 클러스터로 군집화

※ k-means 알고리즘은 n개의 벡터들을 k개의 집합으로 나누되, 각 집합별 중심점과 각 점 간의 거리의 제곱합을 최소로 하도록 나누는 알고리즘임

❍ 각 클러스터의 주요 키워드는 LDA, TF-IDF, TF 3개의 방식으로 가중치 점수가 가장 높은 30개의 단어를 추출하여 종합적으로 활용함

- LDA는 각 클러스터에 적용했으며, 흔히 중복적으로 등장하는 단어(study, research 등)는 자동적으로 제외하고 특징적인 단어만 제시해주는 장점이 있으나, 그러한 특성이 클러스터의 내용을 왜곡할 위험성도 같이 존재하여 다른 방법을 병용함

※ 일부 소수 과제에서 특별히 등장하는 단어가 주요 키워드로 제시될 수가 있음

※ LDA(잠재 디리클레 할당, Latent Dirichlet allocation)는 토픽 모델링에 일반적으로 사용되는 알고리즘으로, 문서의 토픽을 파악할 수 있는 키워드를 제시해줌

- TF-IDF와 TF는 편의상 각 클러스터를 하나의 문서로 취급하여 계산하였음

※ TF(Term Frequency)는 클러스터에 등장하는 각 단어들이 몇 번 등장하는지 빈도 수를 측정한 것임

※ TF-IDF(Term Frequency - Inverse Document Frequency)는 각 단어의 TF 값에 그 단어가 몇 개의 클러스터에 공통적으로 등장했는지(Document frequency)의 역수의 로그값(IDF)을 곱한 값으로, 특정 단어가 클러스터 내에서 얼마나 많이 등장했는지와, 다른 클러스터보다 해당 클러스터에서만 등장했는지를 같이 평가하는 지표임

- TF-IDF의 경우 분석 결과 동일한 TF-IDF 값을 가지는 단어가 매우 많이 나타남에 따라 변별력이 없어 사용하지 않음

※ 모든 클러스터에 등장하면서 해당 클러스터 에서의 등장 횟수(TF)가 같은 단어들의 경우 TF-IDF 값이 동일해져서, 분석 결과 많게는 1,000개 이상의 동일한 TF-IDF 값을 가지는 단어들이 나타났음

- TF 역시 연구(research), 개발(development)과 같은 클러스터 특유의 의미를 담지 않고 일반적으로 등장하는 단어들 위주로 키워드를 제시하기 때문에 사용하지 않고, LDA 기반의 키워드에 의존함

제2절 부처별 분석결과

가. NIH(미국 국립보건원) 과제 분석

□ 바이오의료 분야 미국 R&D 투자의 가장 큰 비중을 차지하는 NIH를 대상으로 분석을 수행하였으며, 전년도 연구에서 국내 조사분석 과제 대상으로 분석한 내용과 비교를 위하여 유전체(genomics) 분야를 대상으로 과제 검색 및 클러스터링을 수행함

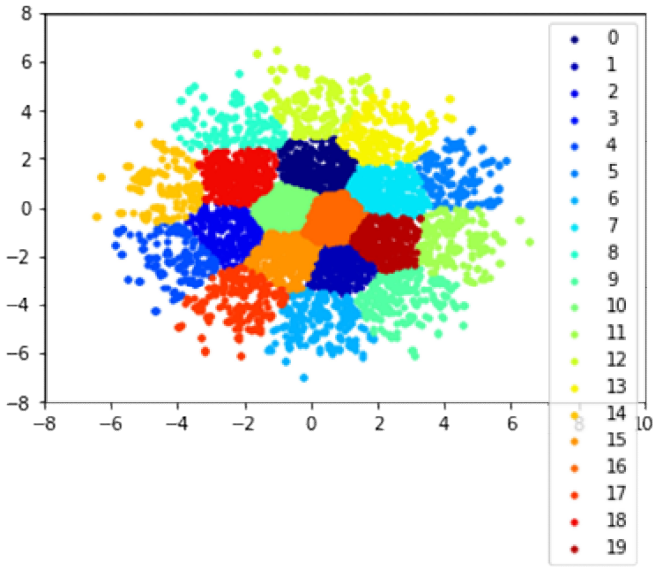
○ 분석 조건은 다음과 같음

- (검색어) genomics
- (검색 과제 수) 5,000개
- (클러스터 수) 20개
- (분석 연도) 2019년

○ 분석 결과, 최근 유전체 관련 기술이 매우 다양한 바이오의료 분야에 범용적으로 적용되고 있음에 따라, 5,000개 과제의 대부분에 해당하는 것으로 드러났으며, 클러스터 간의 내용 차이를 구분하기 어려워 검색어의 구체화가 필요하다고 결론내림(그림 4-1, 표 4-2)

- 세포(cell), 암(cancer), 개발(development), 연구(research), 치료(treatment) 등의 단어들이 여러 클러스터에 공통적으로 등장하고 있어 클러스터 간의 내용 차이 파악이 어려움
- 최근 질병 치료, 질병 관리 등 정밀의료 패러다임에 따라 광범위한 연구에 유전체 분석이 적용되고 있으므로, NIH 지원 과제에서 유전체 해당 여부를 판단하기는 어려울 것으로 보임

○ 따라서 보다 구체적인 검색어 예시로 마이크로바이옴(microbiome)을 테스트함



[그림 4-1] NIH 2019년 'genomics' 5,000개 과제 검색 결과

〈표 4-2〉 NIH 2019년 'genomics' 5,000개 과제 클러스터(20개)별 주요 키워드

클러스터	LDA 키워드
0	['research' 'care' 'health' 'core' 'data' 'treatment' 'cancer' 'center', 'clinical' 'hiv' 'studies' 'training' 'genomics' 'analysis' 'brain', 'resource' 'identify' 'specific' 'program' 'cell']
1	['core' 'research' 'human' 'dna' 'cells' 'cell' 'use' 'disease' 'cancer', 'mechanisms' 'genetic' 'training' 'data' 'risk' 'aim' 'development', 'aging' 'clinical' 'specific' 'infection']
2	['cancer' 'brain' 'research' 'cell' 'risk' 'genetic' 'response' 'clinical', 'study' 'core' 'cells' 'development' 'use' 'disease' 'stroke' 'factors', 'treatment' 'hypertension' 'immune' 'aim']
3	['lifespan' 'superoxide' 'alterations' 'genetic' 'childhood' 'compound', 'aging' 'treatment' 'outcome' 'determine' 'mrd' 'detected' 'ros', 'dependent' 'mitochondrial' 'identify', 'demonstrates' 'increase' 'mild' 'molecular']
4	['aim' 'stroke' 'specific' 'brain' 'research' 'development' 'cancer' 'cell' 'data' 'cells' 'gene', 'clinical' 'rna' 'studies' 'use' 'disease', 'sivd' 'treatment' 'risk' 'aging']
5	['research' 'nfkb' 'care' 'cells' 'cell' 'core' 'cancer' 'clinical' 'gene', 'genetic' 'expression' 'data' 'risk' 'dyrk' 'hiv' 'center' 'treatment', 'training' 'igf' 'alcohol']

클러스터	LDA 키워드
6	['dna' 'research' 'clinical' 'disease' 'cells' 'data' 'cell' 'host' 'lung', 'training' 'core' 'gene' 'aim' 'immune' 'development' 'structure' 'genes', 'th' 'genetic' 'identify']
7	['data' 'research' 'disease' 'dna' 'cells' 'response' 'clinical' 'aim', 'new' 'specific' 'core' 'training' 'ah' 'cell' 'patients' 'treatment', 'health' 'use' 'pol' 'studies']
8	['data' 'cancer' 'research' 'pain' 'cell' 'sleep' 'imaging' 'patients, 'deprivation' 'treatment', 'genomics' 'resource' 'studies' 'tumor', 'genetic' 'pca' 'genes' 'clinical' 'associated' 'analysis']
9	['enzymes' 'cell' 'research' 'cells' 'clinical' 'cancer' 'core' 'disease' 'protein' 'resistance' 'new' 'use' 'data' 'ad' 'human' 'patients' 'signaling' 'studies' 'novel' 'acid']
10	['cancer' 'risk' 'program' 'research' 'diabetes' development' 'disease' 'aim' 'human' 'training' 'health' 'cell' 'clinical' 'cells' 'data' 'social' 'studies' 'outcomes' 'genetic' 'core']
11	['models' 'mutations' 'cancer' 'response' 'cell' 'development' 'pain' 'clinical' 'patients' 'drug' 'treatment' 'signaling' 'research' 'core' 'immune' 'health' 'cells' 'data' 'tumor' 'met']
12	['disorders' 'research' 'genetic' 'data' 'disease' 'cancer' 'sczd' 'clinical' 'gene' 'human' 'genes' 'risk' 'genomic' 'cell' 'variants' 'core' 'based' 'genome' 'psychiatric' 'resource']
13	['research' 'cancer' 'studies' 'core' 'clinical' 'data' 'cell' 'genetic' 'aim' 'genomics' 'dna' 'new' 'risk' 'support' 'genes' 'cells' 'using' 'patients' 'development' 'analysis']
14	['ad' 'research' 'cancer' 'genetic' 'ptsd' 'data' 'risk' 'novel' 'social' 'disease' 'physical' 'imaging' 'tumor' 'development' 'identify' 'specific' 'clinical' 'gene' 'studies' 'patient']
15	['research' 'cells' 'novel' 'disease' 'cell' 'program' 'data' 'project' 'clinical' 'cancer' 'core' 'molecular' 'training' 'development' 'risk' 'patients' 'analysis' 'studies' 'new' 'aim']
16	['genetic' 'research' 'studies' 'treatment' 'clinical' 'cell' 'tumor' 'patients' 'disease' 'care' 'data' 'use' 'core' 'health' 'cells' 'cancer' 'development' 'program' 'mechanisms' 'human']
17	['genetic' 'disease' 'research' 'human' 'genes' 'core' 'cell' 'clinical' 'genome' 'novel' 'gene' 'function' 'studies' 'molecular' 'cells' 'data' 'pdtc' 'mechanisms' 'aim' 'projects']
18	['cancer' 'research' 'tumor' 'cells' 'program' 'imaging' 'cell' 'pain' 'dna' 'breast' 'clinical' 'oral' 'support' 'development' 'biology' 'members' 'tau' 'studies' 'genetic' 'cancers']
19	['core' 'cell' 'research' 'resistance' 'signaling' 'cells' 'cancer' 'treatment' 'human' 'mechanisms' 'program' 'aim' 'risk' 'development' 'studies' 'use' 'identify' 'training' 'receptor' 'drug']

□ 다양한 미생물 군집 내의 상호작용을 연구하는 마이크로바이옴 분야에 대해 연도별로 과제를 검색하고 세부분야에 대한 변화 추이 등에 대하여 최근 5개년(2015~2019) 과제를 분석함

※ 마이크로바이옴은 자연이나 체내에 존재하는 여러 미생물의 군집을 이르는 말로, 최근 유전자 해독 등 분석 기술의 발전에 따라 단일 미생물이 아니라 여러 미생물의 군집적 작용을 활발히 연구하고 있음. 대표적인 분야로 장내 미생물(human gut microbiome) 구성과 사람 건강 · 질병 간의 관계에 대한 연구가 있음

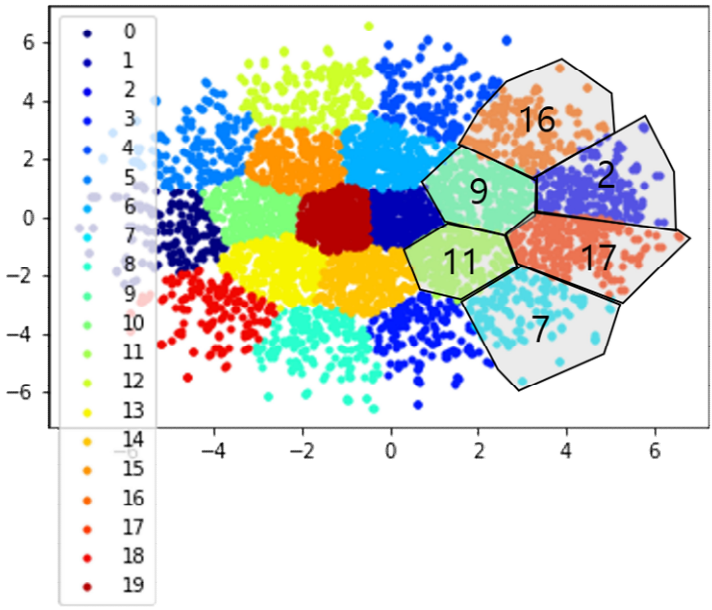
❍ 2019년 과제에 대해 검색하여 분석 조건 최적화를 시도하였으며, 그 결과 연도별 1,500개씩 검색하였음

- 검색어 microbiome으로 5,000개의 과제를 검색한 결과, 마이크로바이옴에 해당하는 것으로 보이는 과제가 약 1,300개였고, 그에 따라 1,500개의 과제를 검색하는 것으로 정함(그림 4-2, 표 4-3)
- 클러스터 별 키워드를 기준으로 판단한 결과, 2, 7, 9, 11, 16, 17번 클러스터가 마이크로바이옴과 직접적 연관이 있는 것으로 보이며, 나머지 과제들은 다른 바이오의료 연구인 것으로 보임
- 이 클러스터들은 전체 분포에서 우측에 몰려있는 것으로 확인되었으며, 총 1,337개 과제임
- 따라서 검색 과제 수를 1,500개로 줄이고, 클러스터 주제 구체화를 위해 15개의 클러스터로 분류

〈표 4-3〉 NIH 2019년 'microbiome' 5,000개 과제 클러스터(20개)별 주요 키워드

클러스터	LDA 키워드
0	['research' 'data' 'disease' 'models' 'patients' 'cell' 'studies' 'clinical' 'cells' 'transmission' 'aim' 'modeling' 'treatment' 'associated' 'development' 'tools' 'risk' 'novel' 'ptn' 'health']
1	['research' 'disease' 'core' 'cells' 'health' 'cell' 'human' 'program' 'mechanisms' 'data' 'il' 'brain' 'aim' 'studies' 'risk' 'cd' 'development' 'cancer' 'specific' 'patients']
2	['gut' 'microbial' 'bacteria' 'intestinal' 'cells' 'microbiome' 'microbiota' 'hiv' 'immune' 'bv' 'human' 'disease' 'gvhd' 'cell' 'mice' 'cd' 'inflammatory' 'research' 'host' 'inflammation']
3	['core' 'data' 'research' 'disease' 'noise' 'cells' 'cell' 'center' 'clinical' 'metabolomics' 'tb' 'provide' 'inner' 'cancer' 'skin' 'risk' 'new' 'project' 'genomics' 'resource']

클러스터	LDA 키워드
4	['cancer' 'aim' 'research' 'bone' 'cells' 'specific' 'exposure' 'crc' 'lung' 'obesity' 'development' 'tumor' 'induced' 'drug' 'human' 'health' 'studies' 'new' 'data' 'signaling']
5	['cancer' 'development' 'clinical' 'data' 'cells' 'intervention' 'brain' 'support' 'pre' 'disease' 'trials' 'signaling' 'cell' 'ad' 'research' 'risk' 'immune' 'social' 'new' 'studies']
6	['cancer' 'research' 'cells' 'clinical' 'cell' 'risk' 'immune' 'human' 'program' 'use' 'tumor' 'data' 'development' 'new' 'center' 'core' 'aim' 'prep' 'specific' 'women']
7	['cell' 'disease' 'gut' 'studies' 'dlx' 'microbiome' 'research' 'cells' 'skin' 'effects' 'metabolomics' 'food' 'associated' 'mechanisms' 'development' 'study' 'risk' 'aim' 'data' 'new']
8	['data' 'research' 'genetic' 'human' 'core' 'bone' 'exposure' 'cell' 'disease' 'community' 'new' 'risk' 'health' 'models' 'cancer' 'center' 'studies' 'environmental' 'methods' 'gene']
9	['disease' 'microbiome' 'cells' 'immune' 'cell' 'host' 'infection' 'gut' 'study' 'development' 'microbiota' 'inflammatory' 'cancer' 'studies' 'research' 'hiv' 'inflammation' 'cd' 'mice' 'genetic']
10	['development' 'aim' 'cells' 'clinical' 'health' 'treatment' 'cancer' 'data' 'cell' 'patients' 'research' 'specific' 'models' 'use' 'gene' 'disease' 'studies' 'novel' 'study' 'care']
11	['research' 'health' 'disease' 'fviii' 'metabolic' 'gut' 'microbiome' 'core' 'cell' 'studies' 'inflammation' 'development' 'role' 'microbiota' 'specific' 'immune' 'cells' 'aim' 'human' 'function']
12	['hiv' 'cell' 'cancer' 'cells' 'immune' 'infection' 'specific' 'research' 'data' 'core' 'use' 'risk' 'studies' 'clinical' 'human' 'dna' 'viral' 'determine' 'disease' 'patients']
13	['development' 'data' 'cells' 'research' 'aim' 'human' 'disease' 'core' 'care' 'studies' 'program' 'center' 'clinical' 'role' 'provide' 'health' 'brain' 'signaling' 'cell' 'hiv']
14	['core' 'research' 'data' 'health' 'studies' 'translational' 'center' 'cancer' 'clinical' 'support' 'provide' 'hiv' 'human' 'investigators' 'models' 'community' 'cells' 'sequencing' 'development' 'pilot']
15	['development' 'clinical' 'cancer' 'data' 'cells' 'cell' 'brain' 'antibiotic' 'hiv' 'based' 'tumor' 'disease' 'risk' 'patients' 'aim' 'infection' 'research' 'immune' 'treatment' 'program']
16	['cells' 'asthma' 'studies' 'lung' 'sepsis' 'immune' 'cell' 'aim' 'microbial' 'hiv' 'research' 'disease' 'airway' 'associated' 'microbiome' 'liver' 'gut' 'function' 'microbiota' 'il']
17	['microbiome' 'microbiota' 'microbial' 'gut' 'oral' 'human' 'bacteria' 'data' 'immune' 'cancer' 'host' 'bacterial' 'cells' 'disease' 'study' 'research' 'milk' 'aim' 'intestinal' 'ahr']
18	['clinical' 'studies' 'research' 'aim' 'burn' 'genetic' 'cell' 'rnas' 'core' 'data' 'patients' 'ad' 'disease' 'long' 'risk' 'consortium' 'cells' 'function' 'rpa' 'use']
19	['research' 'risk' 'core' 'cancer' 'data' 'cell' 'development' 'expression' 'cells' 'disease' 'use' 'function' 'program' 'mechanisms' 'studies' 'health' 'clinical' 'provide' 'specific' 'effects']



[그림 4-2] NIH 2019년 'microbiome' 5,000개 과제 검색 결과 (클러스터 번호 표시)

❍ NIH 2019년 1,500개 과제를 검색한 결과, 총 12개 클러스터 중 8개(712개 과제)가 마이크로바이옴에 직접적으로 연관된 것으로 판명되었으며, 나머지 클러스터는 유전체와 질병의 상관관계 또는 그 외 임상연구에 관련된 것으로 판단됨(표 4-4)

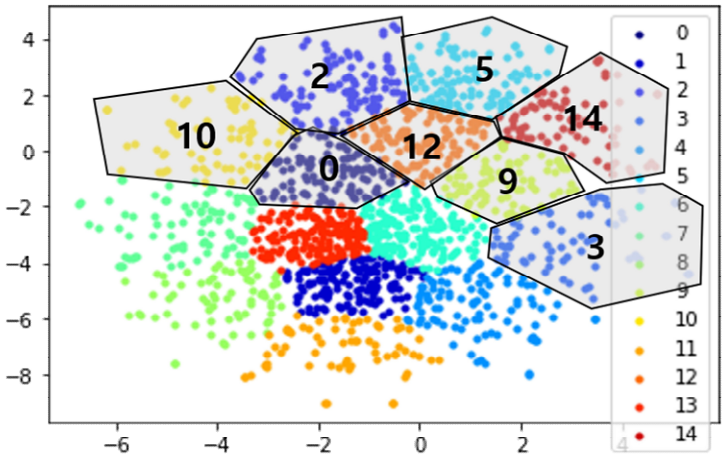
- NIH 2019년 1,500개 과제의 12개 클러스터별 주요 키워드 및 주제는 아래 표와 같으며, 마이크로바이옴 관련 클러스터(0, 2, 3, 5, 9, 10, 12, 14번)들은 대략적으로 위쪽 반원에 모여있음을 알 수 있음(그림 4-3)
- 연구비 기준으로는 약 3억 달러에 해당하는 과제들이 마이크로바이옴 관련 과제에 투자되었음(표 4-5)
- 단, 마이크로바이옴에 연관된 클러스터들의 세부적인 주제를 LDA로 추출한 주요 키워드 및 과제명 리스트를 가지고 정확히 포착하기에는 어려움
- 마이크로바이옴 관련 여부까지가 KISTEP 사업 담당자 수준의 사전 지식(도메인 지식, domain knowledge)을 가지고 판단이 가능한 것으로 보임

- 더욱 세부적인 클러스터링 분석을 위해서는 알고리즘을 통해 더욱 상세하게 클러스터의 주제를 나타내주거나, 관련 전문가의 판단이 요구됨

〈표 4-4〉 NIH 2019년 'microbiome' 1,500개 과제 클러스터(15개)별 주요 키워드 및 주제

클러스터	LDA 키워드	주제
0	['gut' 'mice' 'microbiome' 'research' 'development' 'stress' 'microbiota' 'human' 'cells' 'lung' 'skin' 'ad' 'host' 'inflammatory' 'immune' 'cell' 'disease' 'hiv' 'health' 'fviii']	마이크로바이옴 (질병/면역 연관성)
1	['disease' 'risk' 'data' 'core' 'cell' 'study' 'cells' 'patients' 'research' 'ripk' 'clinical' 'using' 'identify' 'aim' 'development' 'genetic' 'use' 'studies' 'provide' 'health']	유전체 (질병 위험성 연관성)
2	['intestinal' 'gut' 'cells' 'cd' 'human' 'microbiota' 'microbiome' 'hiv' 'lung' 'cell' 'disease' 'carnitine' 'immune' 'microbial' 'research' 'inflammatory' 'ahr' 'inflammation' 'associated' 'metabolic']	마이크로바이옴 (면역/염증 연관성)
3	['cancer' 'host' 'microbiome' 'aim' 'transmission' 'dietary' 'lung' 'bone' 'research' 'ceruloplasmin' 'oral' 'il' 'commensal' 'noise' 'hif' 'disease' 'data' 'metabolites' 'mechanisms' 'risk']	마이크로바이옴 (질병 기전)
4	['cancer' 'cell' 'identify' 'data' 'risk' 'clinical' 'patients' 'genetic' 'disease' 'analysis' 'obesity' 'research' 'models' 'immune' 'new' 'studies' 'gene' 'health' 'core' 'cells']	유전체 (질병 위험성 연관성)
5	['microbiota' 'microbiome' 'gut' 'host' 'bacterial' 'human' 'role' 'microbial' 'mice' 'disease' 'immune' 'bacteria' 'cells' 'studies' 'data' 'intestinal' 'inflammatory' 'mpa' 'research' 'cancer']	장내/구강 마이크로바이옴 이용 질병 치료
6	['cancer' 'clinical' 'patients' 'cells' 'immune' 'ibd' 'research' 'egfr' 'cell' 'disease' 'data' 'core' 'inflammatory' 'training' 'veo' 'health' 'new' 'use' 'associated' 'specific']	면역치료 등
7	['cell' 'rna' 'research' 'immune' 'hiv' 'aim' 'burn' 'health' 'future' 'phage' 'aging' 'use' 'ccr' 'development' 'cd' 'dysfunction' 'risk' 'cells' 'intestinal' 'treatment']	면역학, 에이즈 등

클러스터	LDA 키워드	주제
8	['skin' 'pad' 'data' 'genes' 'research' 'use' 'cocaine' 'brain' 'cells' 'risk' 'part' 'changes' 'replication' 'hiv' 'core' 'dm' 'based' 'human' 'inflammatory' 'sexual']	주제 파악 불가
9	['host' 'immune' 'lung' 'bacterial' 'disease' 'patients' 'uti' 'ptyr' 'ilc' 'ides' 'microbiota' 'ptsd' 'asthma' 'syndrome' 'gut' 'metabolic' 'cells' 'clinical' 'copd' 'domains']	마이크로바이옴 (건강 연관성)
10	['disease' 'asthma' 'hiv' 'immune' 'cells' 'result' 'microbiome' 'host' 'diabetes' 'dlx' 'skin' 'response' 'aim' 'human' 'stumble' 'research' 'data' 'infection' 'microbial' 'cell']	마이크로바이옴 (체질/면역 연관성)
11	['genetic' 'data' 'weight' 'research' 'models' 'cell' 'aim' 'clinical' 'genome' 'cells' 'polarity' 'targets' 'palmitoylation' 'cancer' 'human' 'traits' 'new' 'studies' 'develop' 'sequencing']	주제 파악 불가
12	['microbiome' 'microbial' 'microbiota' 'bacteria' 'oral' 'mg' 'gut' 'cells' 'human' 'intestinal' 'host' 'bv' 'disease' 'immune' 'studies' 'specific' 'vdr' 'hiv' 'data' 'cell']	마이크로바이옴 (질병 연관성)
13	['aging' 'hiv' 'core' 'research' 'use' 'cancer' 'health' 'risk' 'integration' 'aureus' 'clinical' 'development' 'traits' 'program' 'cell' 'administrative' 'investigators' 'function' 'data' 'substance']	주제 파악 불가
14	['azithromycin' 'microbiome' 'gut' 'infection' 'associated' 'sequencing' 'immune' 'inhaled' 'based' 'lung' 'risk' 'cancer' 'patients' 'data' 'sensing' 'gastric' 'bacteria' 'nasal' 'microbial' 'tobramycin']	마이크로바이옴 (건강상태 연관성)



[그림 4-3] NIH 2019년 'microbiome' 1,500개 과제 검색 결과
(클러스터 번호 표시)

〈표 4-5〉 NIH 2019년 'microbiome' 1,500개 과제 클러스터(15개)별
과제 수 및 연구비

클러스터	과제 수	연구비(달러)
0	105	52,469,901
1	146	81,842,420
2	105	47,920,852
3	68	28,751,194
4	97	48,946,694
5	105	41,754,482
6	164	89,428,011
7	82	60,074,573
8	72	29,661,898
9	77	29,432,935
10	66	28,703,893
11	67	33,975,197
12	112	50,051,811
13	159	73,273,036
14	74	26,930,115
총합계	**1,499**	**723,217,012**

※ 과제 수 총합이 정확히 1,500개가 아닌 이유는 데이터 상 오류(결측치 등)로 인해 불러오는데 실패하는 과제가 있는 것으로 추정됨

❍ 최근 5개년(2015~2019) 과제의 비교를 위해 동일한 분석을 2015~2018년 과제에 대해서 시행하였으며, 최근으로 올수록 마이크로바이옴 관련 과제가 대체로 증가하는 경향성이 포착됨

- 마이크로바이옴은 비교적 최근에 각광받고 있는 연구분야로 알려져 있어, 최근으로 올수록 관련 연구과제가 늘어나는 경향성은 사실과 부합함
- 또한 연도별 1,500개의 과제들 중에서 마이크로바이옴과 관련된 클러스터는 연도를 막론하고 위치상 근접하게 모여있는 경향성은 doc2vec을 이용한 클러스터링 분석이 작동함을 보여줌

※ 유사한 텍스트가 공간 상에서 가까운 위치의 벡터로 임베딩되는 것이 doc2vec을 비롯한 텍스트임베딩 알고리즘들의 특성

〈표 4-6〉 NIH 2018년 'microbiome' 1,500개 과제 클러스터(15개)별 주요 키워드

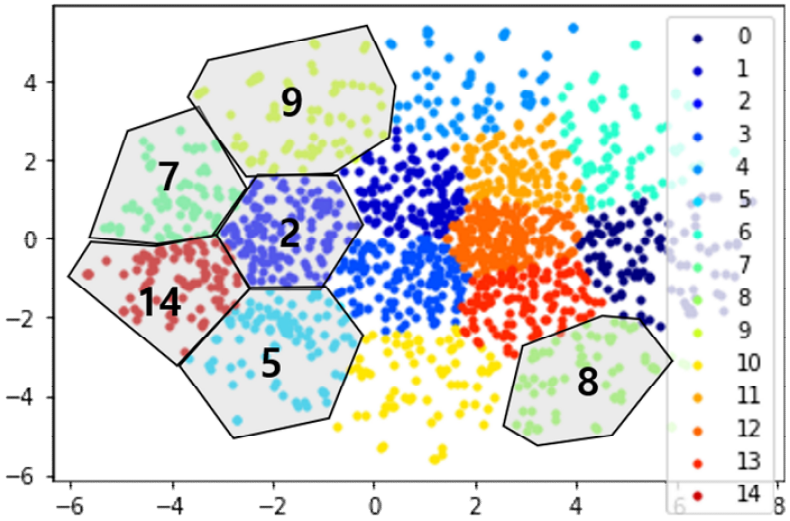
클러스터	LDA 키워드
0	['research' 'sleep' 'reward' 'core' 'models' 'cells' 'disease' 'data' 'immune' 'genetic' 'replication' 'cell' 'specific' 'clinical' 'deficits' 'dna' 'aging' 'program' 'processing' 'aim']
1	['research' 'cells' 'studies' 'study' 'core' 'cancer' 'imaging' 'development' 'lp' 'review' 'center' 'clinical' 'lipid' 'specific' 'human' 'obesity' 'mri' 'investigators' 'mast' 'committee']
2	['gut' 'cells' 'microbiome' 'cimp' 'csf' 'microbial' 'cancer' 'aim' 'microbiota' 'shunt' 'crc' 'bacteria' 'data' 'function' 'disease' 'intestinal' 'il' 'bacterial' 'community' 'th']
3	['data' 'gut' 'cancer' 'cells' 'inflammatory' 'risk' 'mammary' 'disease' 'aim' 'immune' 'ed' 'cell' 'host' 'systems' 'lung' 'studies' 'cd' 'mechanisms' 'inflammation' 'hiv']
4	['research' 'syndrome' 'mirna' 'memory' 'data' 'aging' 'genetic' 'core' 'studies' 'signaling' 'models' 'sjren' 'aim' 'disease' 'awareness' 'cancer' 'traits' 'gene' 'function' 'model']
5	['intestinal' 'cell' 'gut' 'cd' 'immune' 'disease' 'hiv' 'cells' 'inflammation' 'microbiota' 'microbiome' 'role' 'responses' 'sepsis' 'lung' 'function' 'human' 'inflammatory' 'cfs' 'airway']
6	['research' 'data' 'specific' 'center' 'children' 'il' 'metabolic' 'core' 'including' 'health' 'aim' 'studies' 'gene' 'provide' 'commons' 'high' 'omics' 'training' 'word' 'dtmp']
7	['microbiota' 'microbial' 'gut' 'microbiome' 'bacterial' 'human' 'om' 'dr' 'studies' 'communities' 'diet' 'difficile' 'high' 'bacteria' 'oral' 'development' 'data' 'cdi' 'rt' 'ehec']

클러스터	LDA 키워드
8	['disease' 'speech' 'care' 'microbiome' 'metabolic' 'research' 'cytp' 'data' 'fviii' 'identification' 'mass' 'eif' 'use' 'bone' 'ci' 'patients' 'tools' 'cancer' 'development' 'body']
9	['cancer' 'study' 'hpv' 'microbiota' 'role' 'risk' 'tumor' 'research' 'hiv' 'host' 'function' 'patients' 'microbiome' 'development' 'inflammation' 'human' 'taste' 'studies' 'ies' 'aim']
10	['cells' 'rsv' 'cell' 'stroke' 'cftr' 'function' 'mhc' 'role' 'provide' 'cas' 'bacterial' 'crispr' 'copd' 'sleep' 'responses' 'cf' 'immune' 'enms' 'disease' 'iof']
11	['core' 'research' 'studies' 'control' 'care' 'data' 'provide' 'network' 'analysis' 'cancer' 'surgical' 'investigators' 'hrfd' 'program' 'human' 'cognitive' 'projects' 'pre' 'support' 'specific']
12	['hiv' 'clinical' 'research' 'program' 'core' 'studies' 'cancer' 'cell' 'disease' 'data' 'cells' 'aim' 'apoe' 'risk' 'center' 'patients' 'new' 'tumor' 'ddis' 'genomics']
13	['risk' 'research' 'cells' 'disease' 'cell' 'development' 'brain' 'dna' 'project' 'il' 'using' 'human' 'replication' 'data' 'cd' 'non' 'aim' 'function' 'bladder' 'patients']
14	['microbiome' 'microbiota' 'gut' 'immune' 'intestinal' 'disease' 'skin' 'mice' 'host' 'rorg' 'human' 'core' 'activity' 'research' 'ifn' 'microbial' 'rt' 'sphingolipids' 'microbes' 'bacteria']

❍ NIH 2018년 클러스터 그래프를 보면, 8번 클러스터가 다른 클러스터에서 떨어져나와 있는 것을 볼 수 있는데, 실제 과제 리스트 검토를 통해 마이크로바이옴 관련 과제의 비중은 낮아 제외해야 하는 것으로 드러남(표 4-7, 그림 4-4)

- 8번 클러스터(62개 과제)의 과제명에는 microbiome, microbiota 등의 단어가 거의 등장하지 않으며, 전체 텍스트(요약문 포함)에서 5회만 등장함(표 4-7)
- 그럼에도 불구하고 LDA 주요 키워드로 microbiome이 추출된 것은 해당 클러스터의 과제들의 과제들이 매우 산개되어있고 통일성이 낮기 때문인 것으로 추측됨

※ 현재 클러스터링 알고리즘의 개선점이 필요함을 시사



[그림 4-4] NIH 2018년 'microbiome' 1,500개 과제 검색 결과 (클러스터 번호 표시)

〈표 4-7〉 NIH 2018년 'microbiome' 8번 클러스터 상위 코사인유사도 10개 과제

과제명(한국어 번역)	코사인 유사도	연구비 (달러)
Dissection of eIF4E dependent mRNA export (eIF4E 의존 mRNA 사출 현상의 해부)	0.335	210,799
A New Disease Platform Leveraging Complex Drosophila and Mammalian Models (복합 드래소필라와 포유류의 모델을 활용하는 새로운 질병 플랫폼)	0.330	1,988,127
The Immunobiology of Factor VIII (팩터 VIII의 면역생물학)	0.328	373,686
New algorithms and tools for large-scale genomic analyses (대량 유전체 분석을 위한 새로운 알고리즘과 도구)	0.312	490,000
Individual differences in cochlear implant users' audiovisual integration and links to speech proficiency (인공 와우 사용자의 시청각 통합 및 언어능력 연결의 개인차)	0.311	29,188
Investigating cysteine PTMs in living cells (살아있는 세포의 시스테인 PTM 연구)	0.310	297,350
Addressing Sparsity in Metabolomics Data Analysis (대사체 데이터 분석에서 성긴 데이터의 해결)	0.305	414,872

과제명(한국어 번역)	코사인 유사도	연구비 (달러)
Predicted lean body mass, fat mass, and risk of lung, pancreatic, colorectal, breast, and prostate cancers (체질량, 체지방량 예측치와 폐암, 췌장암, 결장 직장암, 유방암 및 전립선 암 위험성)	0.301	79,750
Nutrition, Obesity and Atherosclerosis Training Program (영양, 비만과 아테롬성 동맥경화증 훈련 프로그램)	0.297	292,326
Racial and Ethnic Disparities in Chronic Disease Outcomes and Nurse Practitioner Practice (만성질환 경과와 간호 행위의 인종간 차이)	0.296	689,094

○ 2018년 NIH에서 지원된 마이크로바이옴 관련 과제는 총 446개(연구비 약 2억달러)로 집계됨(표 4-8)

※ 단, 클러스터링 분석에 의한 것이고 개별 과제를 검증한 것은 아니므로 '집계'라는 단어에 주의할 필요가 있음

〈표 4-8〉 NIH 2018년 'microbiome' 1,500개 과제 클러스터(15개)별 과제 수 및 연구비

클러스터	과제 수	연구비(달러)
0	102	49,734,808
1	119	43,480,884
2	123	50,070,390
3	133	44,590,534
4	62	25,827,994
5	89	39,066,190
6	73	32,437,139
7	78	40,057,874
8	62	25,775,160
9	69	28,016,382
10	75	31,516,323
11	111	49,136,804
12	195	73,251,996
13	122	49,215,412
14	87	41,653,347
총합계	**1,500**	**623,831,237**

❍ 2017년 NIH 과제를 대상으로 'microbiome'을 검색한 결과, 총 1,500건 과제의 15개 클러스터 중 6개(2, 3, 4, 7, 11, 13번)의 클러스터가 마이크로바이옴에 대한 키워드를 포함하고 있었으며, 실제 내용이 마이크로바이옴과 관련성이 낮은 3, 4번 클러스터를 제외함(표 4-9, 그림 4-5)

- 3, 4번 클러스터의 과제들을 검토한 결과 마이크로바이옴 관련 과제의 비중이 낮아 제외하여야 한다고 판단함(표 4-10, 11)
- 총 426개 과제가 집계되었으며 연구비는 1.76억달러임(표 4-12)

〈표 4-9〉 NIH 2017년 'microbiome' 1,500개 과제 클러스터(15개)별 주요 키워드

클러스터	LDA 키워드
0	['il' 'research' 'aim' 'studies' 'cell' 'pdt' 'function' 'auditory' 'core' 'use' 'cells' 'clinical' 'metabolic' 'beta' 'nssi' 'pain' 'ethics' 'inflammation' 'specific' 'jak']
1	['training' 'cells' 'cell' 'research' 'hiv' 'disease' 'data' 'cancer' 'risk' 'core' 'use' 'baf' 'specific' 'immune' 'high' 'program' 'studies' 'neurons' 'health' 'aim']
2	['gut' 'microbiome' 'host' 'microbiota' 'gvhd' 'microbial' 'human' 'cdi' 'data' 'reg' 'immune' 'antibiotic' 'colon' 'aim' 'changes' 'bacteria' 'disease' 'studies' 'cancer' 'intestinal']
3	['cells' 'data' 'trib' 'determine' 'surgery' 'fsh' 'cell' 'control' 'hiv' 'microbiome' 'genetic' 'development' 'cancer' 'bone' 'function' 'prostate' 'muscle' 'stem' 'model' 'research']
4	['cells' 'klf' 'role' 'circadian' 'patients' 'research' 'asthma' 'disease' 'inflammation' 'cell' 'inflammatory' 'immune' 'microbiota' 'studies' 'lung' 'gvhd' 'prmt' 'human' 'mucosal' 'injury']
5	['cells' 'aim' 'research' 'dna' 'hiv' 'data' 'analysis' 'trunk' 'protein' 'health' 'melanosome' 'cell' 'effects' 'network' 'coverage' 'training' 'face' 'systems' 'pathways' 'brain']
6	['core' 'research' 'center' 'data' 'functional' 'health' 'clinical' 'tissue' 'stress' 'speech' 'administrative' 'resolvins' 'provide' 'development' 'cortisol' 'network' 'community' 'repair' 'nqo' 'investigators']
7	['responses' 'microbiota' 'gut' 'hiv' 'immune' 'cells' 'mucosal' 'intestinal' 'cell' 'inflammation' 'response' 'vaginal' 'disease' 'metabolic' 'intestine' 'studies' 'ph' 'axis' 'host' 'inflammatory']
8	['data' 'genetic' 'beta' 'models' 'model' 'protein' 'exposure' 'research' 'asthma' 'methods' 'disease' 'metabolic'

클러스터	LDA 키워드
	'modeling' 'human' 'structure' 'adhd' 'analysis' 'health' 'core' 'new']
9	['research' 'cancer' 'hiv' 'patients' 'cell' 'tumor' 'cells' 'use' 'data' 'immune' 'exercise' 'aml' 'development' 'based' 'model' 'webcore' 'effects' 'intervention' 'nf' 'ptpn']
10	['cancer' 'clinical' 'core' 'research' 'cd' 'imaging' 'studies' 'data' 'provide' 'genes' 'changes' 'health' 'consortium' 'metabolic' 'specific' 'services' 'support' 'human' 'response' 'folate']
11	['research' 'microbiome' 'intestinal' 'microbial' 'gut' 'obesity' 'cells' 'gi' 'data' 'ibd' 'inflammatory' 'development' 'microbiota' 'disease' 'immune' 'associated' 'study' 'function' 'specific' 'risk']
12	['balance' 'fgf' 'cell' 'research' 'tissue' 'meniscectomy' 'patients' 'otolith' 'cells' 'project' 'speech' 'study' 'vestibular' 'pd' 'aim' 'latently' 'core' 'hiv' 'bone' 'sleep']
13	['immune' 'development' 'gut' 'microbial' 'human' 'disease' 'host' 'microbiome' 'cell' 'data' 'research' 'patients' 'risk' 'gene' 'cancer' 'aim' 'function' 'cfs' 'clinical' 'oral']
14	['pathway' 'career' 'ptsd' 'aim' 'gene' 'data' 'immune' 'stimulation' 'enm' 'sci' 'oncology' 'neural' 'exposure' 'research' 'community' 'host' 'memory' 'propionate' 'brains' 'female']

〈표 4-10〉 NIH 2017년 'microbiome' 3번 클러스터 상위 코사인유사도 10개 과제

과제명(한국어 번역)	코사인 유사도	연구비 (달러)
Effect of sex hormones on HIV infection of cervical and rectal mucosal tissue (자궁 경부 및 직장 점막 조직의 HIV 감염에 대한 성 호르몬의 영향)	0.258	713,593
The Impact of Breast Milk on the Developing Infant Microbiome (모유가 발달하는 영아 미생물 군집에 미치는 영향)	0.251	44,044
Dietary fat ratios influence adolescent depression (식이 지방 비율은 청소년 우울증에 영향을 미친다)	0.236	462,247
Modulation of aged hematopoietic stem cell niches (노화 된 조혈 줄기 세포 틈새의 조절)	0.232	191,875
Methods for the Genetic Epidemiology of Complex Traits (복잡한 형질의 유전 역학을 위한 방법)	0.229	377,975

과제명(한국어 번역)	코사인 유사도	연구비 (달러)
Project 3: Effects of Androgen and Diet on Adipose Function (프로젝트 3 : 안드로겐과식이가 지방 기능에 미치는 영향)	0.227	167,938
Role of long non-coding RNAs in p53 signaling (p53 신호 전달에서 긴 논코딩 RNA의 역할)	0.224	835,910
Cellular, Molecular, and Functional Characterization of Quiescent/Active Intestin (대기/활성 장의 세포, 분자 및 기능적 특성)	0.222	82,995
Communication Outcomes After Head & Neck Cancer (두경부암 이후 의사 소통 결과)	0.220	300,680
Role of Hedgehog Signaling in Chronic Gastritis and Metaplasia (만성 위염 및 Metaplasia에서 고슴도치 신호의 역할)	0.216	427,585

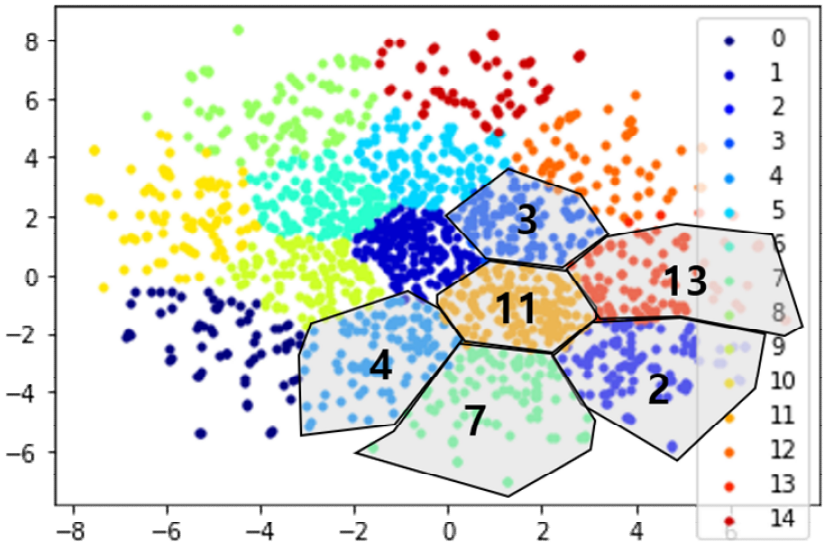
〈표 4-11〉 NIH 2017년 'microbiome' 4번 클러스터 상위 코사인유사도 10개 과제

과제명(한국어 번역)	코사인 유사도	연구비 (달러)
Airway Microbiome in Cystic Fibrosis Pulmonary Exacerbations (낭포성 폐 섬유증의 기도 마이크로바이옴)	0.306	170,640
Intestinal microbiota effects on damage tolerance in aging (노화의 손상 내성에 대한 장내 균총의 효과)	0.263	234,000
Using human stem cell-derived thymic epithelium to remodel T1D immune tolerance (인간 줄기세포 유래 흉선 상피를 사용한 T1D 면역 내성 리모델링)	0.244	578,390
Lung Host Defense in Microgravity (Microgravity 에서의 폐 숙주 방어)	0.242	785,943
Hormones in allergic disease (알레르기 질환의 호르몬)	0.232	404,370
Mechanisms of Chronic Inflammation in Periodontitis (치주염에서 만성 염증의 기전)	0.231	947,033
Modeling the Impact of Targeted Therapy Based on Breast Cancer Subtypes (유방암 아형 기반의 표적 치료의 영향 모델링)	0.230	401,350
Effects of aging on the T follicular helper response to influenza vaccine (노화가 인플루엔자 백신에 대한 난포 도우미 반응에 미치는 영향)	0.226	194,318

과제명(한국어 번역)	코사인 유사도	연구비 (달러)
The UCLA Center for Medical Countermeasures Against Radiation (UCLA 방사능 의학적 대처 연구 센터)	0.225	3,234,284
Role of Circadian Clocks in Aging using Drosophila (Drosophila를 이용하여 노화에서의 생체시계 역할에 대한 연구)	0.222	93,370

〈표 4-12〉 NIH 2017년 'microbiome' 1,500개 과제 클러스터(15개)별 과제 수 및 연구비

클러스터	과제 수	연구비(달러)
0	64	22,777,230
1	165	81,522,717
2	101	45,625,657
3	105	40,681,167
4	95	50,715,275
5	112	50,944,403
6	138	104,686,694
7	89	35,746,493
8	86	50,334,864
9	110	45,824,687
10	85	45,484,475
11	143	52,481,668
12	60	21,025,075
13	93	42,358,423
14	54	17,087,554
총합계	**1,500**	**707,296,382**



[그림 4-5] NIH 2017년 'microbiome' 1,500개 과제 검색 결과 (클러스터 번호 표시)

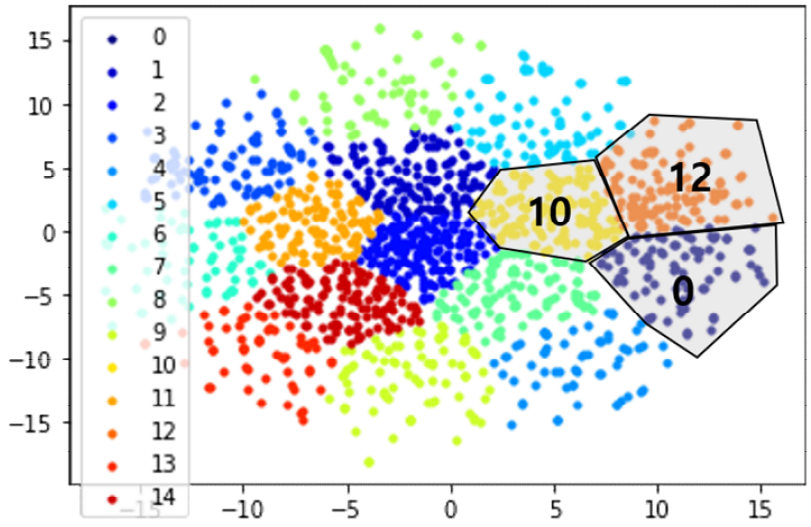
❍ 2016년 NIH 과제를 대상으로 'microbiome'을 검색한 결과, 키워드를 기반으로 판단했을 때 2개(10, 12번)만의 마이크로바이옴 관련 클러스터가 존재하지만, 인접한 클러스터 중 0번 클러스터도 마이크로바이옴에 해당하는 것으로 판명됨(표 4-6, 그림 4-13)

- 0번 클러스터의 주요 키워드 중 'microbiome'은 존재하지 않지만, 'human', 'gut', 'microbial'은 장내 미생물(human gut microbiome)에 관련된 키워드로 보임
- 실제 0번 클러스터의 과제 목록을 보면 마이크로바이옴에 해당하는 내용임을 알 수 있음

〈표 4-13〉 NIH 2016년 'microbiome' 1,500개 과제 클러스터(15개)별 주요 키워드

클러스터	LDA 키워드
0	['hiv' 'infection' 'human' 'gut' 'disease' 'sa' 'hsv' 'bacterial' 'host' 'vaginal' 'women' 'transmission' 'phage' 'cells' 'microbial' 'development' 'resistance' 'nasal' 'aim' 'inflammation']
1	['training' 'research' 'core' 'data' 'health' 'center' 'use' 'hiv' 'exposure' 'projects' 'care' 'administrative' 'program' 'provide' 'pgx' 'science' 'gata' 'clinical' 'community' 'support']
2	['cells' 'research' 'core' 'data' 'aging' 'use' 'cell' 'bone' 'response' 'cancer' 'infections' 'disease' 'factors' 'based' 'risk' 'changes' 'proteins' 'aim' 'cannabis' 'treatment']
3	['treatment' 'stigma' 'children' 'development' 'anxiety' 'vocal' 'use' 'memory' 'hrp' 'learning' 'early' 'study' 'skills' 'associated' 'ptsd' 'data' 'brain' 'visual' 'behavioral' 'studies']
4	['hiv' 'proteins' 'research' 'personality' 'dna' 'intestinal' 'use' 'cells' 'affinity' 'shivs' 'aging' 'fiv' 'clinical' 'melatonin' 'il' 'risk' 'ucsf' 'viral' 'study' 'models']
5	['younger' 'data' 'cv' 'produce' 'core' 'obesity' 'research' 'cells' 'il' 'obese' 'em' 'aim' 'lung' 'preterm' 'scientists' 'group' 'use' 'health' 'safety' 'food']
6	['data' 'core' 'clinical' 'ezh' 'cell' 'food' 'cells' 'gvhd' 'patients' 'research' 'effects' 'models' 'study' 'obstruction' 'allergy' 'also' 'dem' 'cgd' 'whether' 'use']
7	['development' 'research' 'use' 'core' 'intervention' 'specific' 'obesity' 'hypoglycemia' 'adhesins' 'cell' 'glycosylation' 'also' 'cells' 'infants' 'study' 'effects' 'blood' 'brain' 'nutrition' 'risk']
8	['research' 'biomedical' 'use' 'center' 'platelet' 'nu' 'lti' 'students' 'aging' 'mentor' 'cells' 'core' 'yoga' 'program' 'urinary' 'scientific' 'cda' 'studies' 'curli' 'cell']
9	['data' 'research' 'editing' 'cells' 'wnt' 'aid' 'cancer' 'igf' 'msi' 'ape' 'cell' 'adolescents' 'systems' 'signaling' 'program' 'disease' 'students' 'dili' 'brain' 'runx']
10	['cells' 'microbiome' 'associated' 'bacterial' 'disease' 'research' 'gut' 'role' 'mice' 'cancer' 'development' 'studies' 'data' 'microbiota' 'ahr' 'microbial' 'core' 'pgn' 'mechanisms' 'human']
11	['cells' 'activation' 'signaling' 'il' 'stress' 'disease' 'use' 'risk' 'disorders' 'core' 'treatment' 'cd' 'response' 'et' 'mechanisms' 'aim' 'development' 'study' 'propionate' 'studies']

클러스터	LDA 키워드
12	['microbiota' 'gut' 'microbiome' 'disease' 'intestinal' 'microbial' 'human' 'gi' 'cell' 'asthma' 'gvhd' 'cells' 'cd' 'bacteria' 'aim' 'patients' 'identify' 'community' 'immune' 'study']
13	['cls' 'pathway' 'health' 'data' 'cells' 'project' 'genes' 'evaluation' 'cortisol' 'risk' 'research' 'develop' 'genome' 'gene' 'models' 'aim' 'studies' 'cancer' 'core' 'dynamics']
14	['disease' 'trib' 'cell' 'protein' 'gene' 'data' 'cancer' 'pathways' 'expression' 'genes' 'risk' 'nf' 'endothelial' 'replication' 'specific' 'regulation' 'rna' 'studies' 'novel' 'development']



[그림 4-6] NIH 2016년 'microbiome' 1,500개 과제 검색 결과 (클러스터 번호 표시)

○ 2016년 NIH 마이크로바이옴 검색 결과의 클러스터링 분석 결과, 총 331개 과제(1.36억 달러)의 마이크로바이옴 관련 과제가 집계되었음(표 4-14)

〈표 4-14〉 NIH 2016년 'microbiome' 1,500개 과제 클러스터(15개)별 과제 수 및 연구비

클러스터	과제 수	연구비(달러)
0	88	39,460,141
1	134	79,761,034
2	174	74,680,076
3	93	41,273,787
4	61	40,273,549
5	82	46,924,698
6	67	29,948,626
7	109	37,028,885
8	73	25,734,063
9	81	43,336,334
10	125	44,756,931
11	111	52,002,372
12	118	51,372,968
13	61	27,351,861
14	123	60,551,551
총합계	**1,500**	**694,456,876**

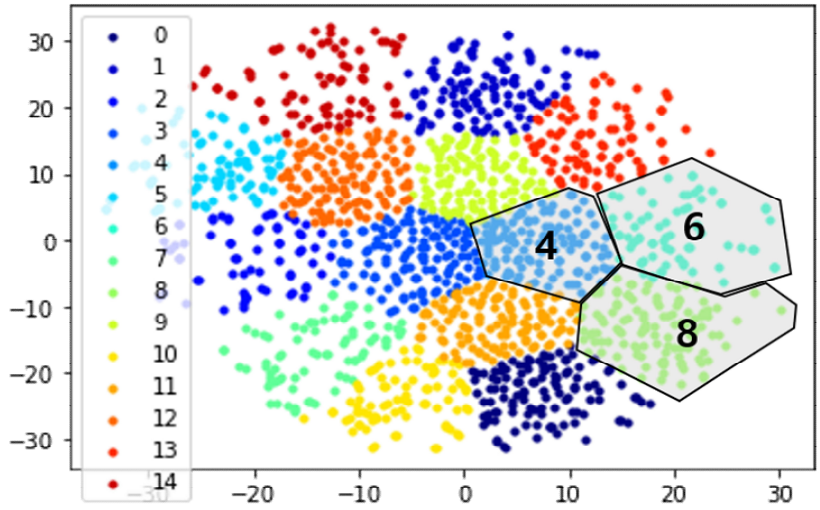
❍ 2015년 NIH 과제를 대상으로 검색한 결과, 3개 클러스터(4, 6, 8번)가 마이크로바이옴에 직접적인 관련이 있는 것으로 드러남(표 4-15, 그림 4-7)

- 해당 과제는 총 305개이며, 1.11억 달러로 집계됨(표 4-16)

〈표 4-15〉 NIH 2015년 'microbiome' 1,500개 과제 클러스터(15개)별 주요 키워드

클러스터	LDA 키워드
0	['cells' 'immune' 'responses' 'cell' 'il' 'role' 'inflammatory' 'cd' 'response' 'sit' 'specific' 'infection' 'cgd' 'lung' 'th' 'arthritis' 'disease' 'influenza' 'mice' 'aim']
1	['implementation' 'children' 'health' 'care' 'research' 'community' 'study' 'use' 'treatment' 'sci' 'based' 'collaboration' 'program' 'data' 'quality' 'social' 'outcomes' 'food' 'hiv' 'project']

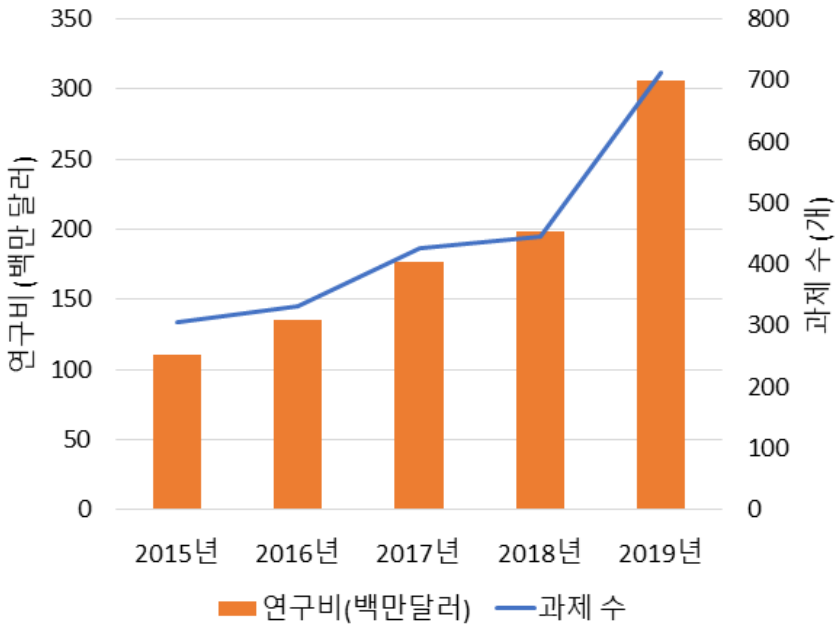
클러스터	LDA 키워드
2	['core' 'exposure' 'data' 'workforce' 'liver' 'research' 'health' 'risk' 'models' 'science' 'fatty' 'pathway' 'development' 'clinical' 'study' 'function' 'genomes' 'psychosocial' 'pain' 'develop']
3	['sf' 'cancer' 'research' 'cells' 'signaling' 'data' 'bone' 'dna' 'disease' 'response' 'variants' 'core' 'mechanisms' 'family' 'program' 'support' 'tumor' 'breast' 'growth' 'function']
4	['bone' 'cells' 'research' 'women' 'immune' 'factors' 'care' 'hiv' 'study' 'cancer' 'health' 'associated' 'induced' 'cd' 'stress' 'disease' 'microbiota' 'role' 'functional' 'program']
5	['research' 'core' 'response' 'clinical' 'human' 'review' 'models' 'health' 'administrative' 'center' 'genome' 'data' 'provide' 'development' 'monitoring' 'segregation' 'ensure' 'training' 'program' 'project']
6	['insulin' 'fat' 'study' 'cells' 'mice' 'gut' 'ba' 'gh' 'development' 'effects' 'src' 'microbiota' 'ee' 'signaling' 'obesity' 'aim' 'research' 'cancer' 'changes' 'inflammation']
7	['asthma' 'core' 'patients' 'research' 'human' 'cancer' 'study' 'clinical' 'genetic' 'prostate' 'immune' 'analysis' 'studies' 'identify' 'provide' 'environmental' 'aki' 'risk' 'autism' 'response']
8	['metabolites' 'microbial' 'environmental' 'hiv' 'gut' 'microbiome' 'host' 'human' 'species' 'intestinal' 'immune' 'community' 'research' 'bacteria' 'microbiota' 'interactions' 'bacterial' 'health' 'aim' 'provide']
9	['developmental' 'research' 'develop' 'data' 'bone' 'recovery' 'study' 'using' 'core' 'based' 'language' 'memory' 'social' 'listeners' 'development' 'clinical' 'changes' 'also' 'food' 'eating']
10	['asthma' 'risk' 'study' 'immune' 'disease' 'research' 'patients' 'clinical' 'microbial' 'tlr' 'hsv' 'using' 'infections' 'nhl' 'il' 'hla' 'human' 'lung' 'data' 'nec']
11	['cells' 'il' 'cell' 'par' 'inflammatory' 'disease' 'arg' 'tumor' 'tfh' 'aim' 'inflammation' 'response' 'immune' 'colitis' 'role' 'human' 'acth' 'oral' 'variant' 'determine']
12	['data' 'research' 'core' 'center' 'test' 'imaging' 'provide' 'information' 'administrative' 'study' 'children' 'investigators' 'results' 'consortium' 'develop' 'mapp' 'hiv' 'development' 'oaic' 'analysis']
13	['sleep' 'research' 'children' 'study' 'circadian' 'alcohol' 'health' 'factors' 'social' 'use' 'life' 'development' 'increased' 'care' 'risk' 'stress' 'diabetes' 'air' 'data' 'early']
14	['research' 'health' 'dna' 'gata' 'skeletal' 'core' 'trex' 'sac' 'care' 'scattering' 'function' 'neurons' 'cells' 'control' 'faculty' 'provide' 'data' 'community' 'specific' 'ubiquitin']



[그림 4-7] NIH 2015년 'microbiome' 1,500개 과제 검색 결과
(클러스터 번호 표시)

〈표 4-16〉 NIH 2015년 'microbiome' 1,500개 과제 클러스터(15개)별 과제 수 및 연구비

클러스터	과제 수	연구비(달러)
0	114	54,874,582
1	108	42,542,296
2	69	46,071,448
3	124	49,637,271
4	118	43,108,858
5	85	34,028,209
6	73	32,835,976
7	80	31,296,193
8	114	34,785,794
9	117	47,434,883
10	82	41,586,926
11	118	50,747,190
12	122	68,176,595
13	94	33,149,218
14	82	42,597,902
총합계	**1,500**	**652,873,341**



[그림 4-8] NIH 마이크로바이옴(microbiome) 분야 2015~2019년 지원과제 추이

❍ NIH 마이크로바이옴 분야 최근 5년간 과제(2015~2019)는 전체적으로 증가하여 2015년 대비 2019년 3배 정도의 연구비 규모로 투자가 확대되었음 (그림 4-8)

- 과제 수 역시 연구비 규모와 거의 동일한 패턴으로 변화하여 2015년 대비 2.3배 증가하였음

❍ 동 연구에서는 미국 NIH의 마이크로바이옴 분야 투자액을 기존에 보고된 수치보다 높게 추정하였으나, 검색 노이즈를 감안하였을 때 충분히 참고할만한 수치인 것으로 판단됨

- 기존에 보고된 바에 따르면, 미국은 국가 마이크로바이옴 이니셔티브 사업(National Microbiome Initiative)을 통하여 2016~2017년 휴먼 마이크로바이옴 분야에 1.21억 달러를 투자함[5)]

5) 황은혜, 김은정, 남영도(2018) KISTEP 기술동향브리프 : 휴먼 마이크로바이옴

- 본 연구에서 추정된 2016~2017년 마이크로바이옴 분야 NIH 투자액은 약 4억 달러로 해당 수치보다 높음
- 미국 NIH 펀딩 중 NMI 사업 외 각종 기초연구과제, 병원 중개연구과제 등 다양한 형태의 마이크로바이옴 관련 연구가 수행될 것으로 추측되며, 동 연구진이 마이크로바이옴 관련으로 분류한 클러스터 내에 마이크로바이옴에 해당하지 않는 과제들이 노이즈로 섞여있기 때문에 수치의 차이가 존재하는 것으로 생각됨

❍ NIH RePORT에서 제공하는 분야별 투자액(categorical spending) 통계[6] 또는 단순 키워드 검색 결과에 비하면 적게 추정하고 있어, 본 방법론은 어느정도 보수적인 추정치를 제시하는 것으로 볼 수 있음

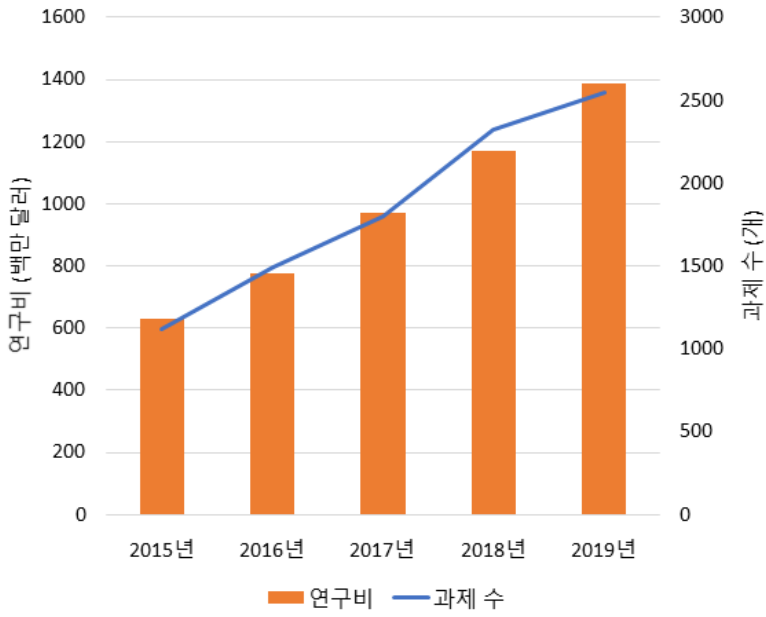
- NIH 분야별 투자액(categorical spending)은 NIH에서 자체적인 분류체계로 산출한 통계치로, 마이크로바이옴 분류는 2019년 신설되어 2018년 이전의 통계치는 없으며, 하나의 과제가 여러개의 분류에 중복 계상되는 통계로 어느정도 과대 계상된다고 볼 수 있음
- NIH에서 명시적으로 제공하는 마이크로바이옴 분야 2019년 투자액은 7.66억 달러로(표 4-17), 본 연구에서 추정한 3억 달러는 이보다 훨씬 적음

〈표 4-17〉 마이크로바이옴 분야 NIH 분야별 투자액(RePORTER)

(단위 : 백만 달러)

2019년	2020년 추정치	2021년 추정치
766	812	745

6) https://report.nih.gov/categorical_spending.aspx



[그림 4-9] NIH 과제 키워드 검색 결과

- NIH RePORTER에서 'microbiome' or 'microbiota' 검색식으로 2015~2019년 과제를 검색한 결과, 과제 수 및 연구비 규모는 동 연구에서 추정한 값의 4~5배 정도로 큼(그림 4-9)
- 하지만 이것도 과제명 또는 요약문에 해당 단어가 한 번이라도 등장하면 집계하는 방식으로 매우 과대 계상되는 방식임

❍ 결론적으로, 지능형 분석시스템에 탑재되어있는 doc2vec 기반의 문서 클러스터링 알고리즘은 NIH 연구과제에도 적용이 가능하며, 이를 통해 연도별 과제 수 및 연구비를 추산할 수 있음

- 단, 유전체(genomics) 같이 매우 포괄적인 분야에 대해 검색할 경우 검색결과가 너무 많아지기 때문에 클러스터링을 하더라도 클러스터별 주제를 특정하기 어려워짐
- 그리고 연구과제들의 융합성 및 텍스트마이닝 알고리즘의 기술적 한계로 인해 문서 클러스터링의 정확도 및 정밀도는 어느정도 떨어질 수 밖에 없음

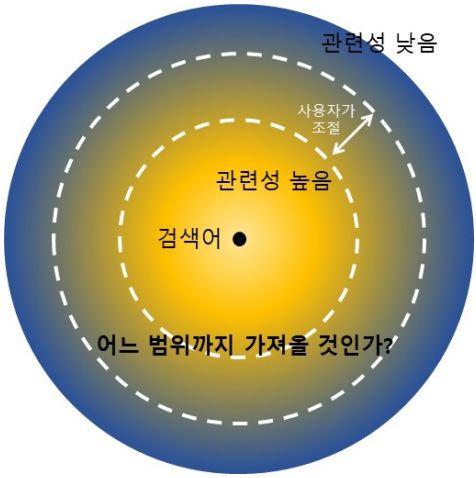
- 이로 인해 어떠한 주제로 검색을 한다고 하여도 유사도가 낮아짐에 따라 해당 주제에 해당하는 과제와 그렇지 않은 과제들이 불명확하게 섞여있게 되며, 어느 정도(threshold)를 기준으로 검색 결과를 끊고, 클러스터링 할 것인지는 본 연구와 같은 접근의 난점이라고 할 수 있음
  ※ 이러한 논의점은 2018, 2019년 연구결과에서도 다룬 바 있음

❍ 이러한 난점은 그림 4-10처럼 표현할 수 있는데, 검색어와 과제 간의 관련성은 예/아니오로 명확히 나뉘지 않고 상대적인 차이가 있을 뿐이므로 그림과 같이 그라데이션 같은 분포를 가지게 됨

- 과제 검색 시 어느 범위까지 가져올 것인지가 주관적 판단의 영역에 놓이게 되며, 클러스터링도 얼마나 세분화 하여 나눌 것인지가 주관적 판단의 영역에 놓이게 됨

❍ 이는 알고리즘의 성능의 문제라기보다는 본질적인 문제이며, 서로 다른 분야의 융합 연구가 더욱 활성화되고 있는 추세이므로 텍스트마이닝을 통한 자동적 과제 분류는 더욱 어려워질 것임

❍ 따라서, 이러한 과제들의 분포에 대하여 얼마나 잘 시각화 하는지가 관건이라고 생각되며, 제4절 소결 및 한계점에서 개선방안에 대해 후술하겠음

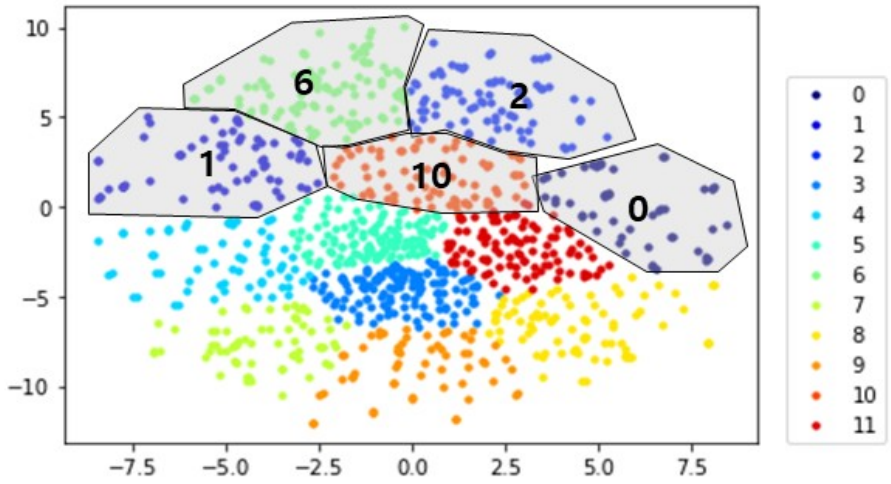


[그림 4-10] 텍스트임베딩 기반 과제검색 및 클러스터링의 난점

나. 미국 NSF(국립 과학재단) 과제 분석

□ 미국 NSF는 과학기술 전 분야의 기초연구 단계를 지원하는 전문관리기관으로서, 보건의료 이외 다양한 분야의 과제가 포함되어 있어 농림수산식품(종자 개발) 분야의 분석을 시험적으로 수행하였음

○ 2015~2019년 과제에 대하여 'breeding'의 검색어로 검색하여, 유사도 상위 1,000개 과제를 12개 클러스터로 분류하였고 그 결과는 다음과 같음(그림 4-11, 표 4-18)



[그림 4-11] NSF 2015~2019년 'breeding' 1,000개 과제 검색 결과

〈표 4-18〉 NSF 2015~2019년 'breeding' 1,000개 과제 클러스터(12개)별 주요 키워드

클러스터	LDA 키워드
0	['research' 'br' 'project' 'plant' 'quantum' 'sleep' 'spin' 'high' 'development' 'conference' 'new' 'movement' 'control' 'soil' 'students' 'animal' 'biology' 'iron' 'award' 'plants']
1	['br' 'species' 'project' 'nitrogen' 'change' 'research' 'genes' 'data' 'biodiversity' 'genetic' 'populations' 'gene' 'students' 'variation' 'ocean' 'model' 'anthropogenic' 'reproductive' 'digitization' 'deposition']

클러스터	LDA 키워드
2	['br' 'project' 'plant' 'genome' 'research' 'sequence' 'plants' 'wheat' 'indigo' 'gene' 'maize' 'stomatal' 'develop' 'crop' 'using' 'genes' 'new' 'genetic' 'species' 'stress']
3	['research' 'br' 'graduate' 'stem' 'project' 'grfp' 'program' 'geometry' 'students' 'duplicate' 'support' 'virtual' 'data' 'education' 'conference' 'gene' 'nsf' 'new' 'fellowship' 'award']
4	['br' 'forest' 'research' 'project' 'ice' 'management' 'data' 'students' 'aluminum' 'climate' 'using' 'water' 'clustering' 'sea' 'across' 'new' 'wildlife' 'time' 'puberty' 'variable']
5	['br' 'graduate' 'research' 'workshop' 'conference' 'grfp' 'ice' 'students' 'project' 'new' 'carbon' 'shelf' 'program' 'stem' 'education' 'algebras' 'support' 'black' 'protein' 'resilience']
6	['species' 'plant' 'br' 'project' 'research' 'plants' 'evolution' 'pollen' 'genetic' 'data' 'traits' 'evolutionary' 'host' 'seed' 'schiedea' 'pollination' 'understanding' 'stress' 'different' 'self']
7	['br' 'research' 'physics' 'conference' 'gluon' 'award' 'students' 'matter' 'workshop' 'nuclear' 'social' 'geometry' 'black' 'project' 'support' 'plasma' 'space' 'nucleus' 'analysis' 'new']
8	['students' 'research' 'stem' 'project' 'br' 'program' 'support' 'engineering' 'energy' 'science' 'high' 'student' 'teacher' 'mathematics' 'retention' 'year' 'grain' 'development' 'university' 'teachers']
9	['br' 'particle' 'research' 'manufacturing' 'workshop' 'conference' 'computer' 'project' 'lhc' 'engineering' 'students' 'data' 'science' 'materials' 'solar' 'plant' 'programming' 'using' 'physics' 'new']
10	['br' 'project' 'research' 'plant' 'species' 'grasses' 'carbon' 'students' 'division' 'cell' 'nitrogen' 'changes' 'plants' 'using' 'stem' 'award' 'development' 'fellowship' 'growth' 'program']
11	['br' 'project' 'cell' 'research' 'algebraic' 'imaging' 'workshop' 'sorting' 'fellowship' 'sliding' 'robotics' 'amino' 'soil' 'geometry' 'glacier' 'snowpac' 'data' 'theory' 'students' 'acid']

❍ 주요 키워드를 기준으로 판단했을 때, 12개 클러스터 중 육종과 관련되었을 가능성이 높은 클러스터는 0, 1, 2, 6, 10번 클러스터이며, 총 371개의 과제에 해당함

- 그림 4-11에서 확인할 수 있듯 해당 과제들은 전체 분포 중 위쪽 반원에 모여있음

❍ NSF 과제 데이터를 분석할 때 유의할 점은, 주제를 특정하지 않은 일반적인 장학금(펠로우십)에 해당하는 과제들이 상당량 포함되어 있는데, 이 과제들은 구체적인 연구 내용을 알 수 없으며 인건비 성격이므로 제외할 필요가 있음

- 예를 들면 「Graduate Research Fellowship Program(GRFP)」이라는 제목을 가지고 있는 과제가 있는데, 초록에 해당 장학금 프로그램의 일반적인 소개 내용만 동일하게 기술되어 있어(아래 박스에 표시) R&D 투자동향 분석 등 동 연구의 활용 목적에는 필요 없는 과제들임

※ Graduate Research Fellowship Program(GRFP) 과제 초록(abstract)

The National Science Foundation(NSF) Graduate Research Fellowship Program(GRFP) is a highly competitive, federal fellowship program. GRFP helps ensure the vitality and diversity of the scientific and engineering workforce of the United States. The program recognizes and supports outstanding graduate students who are pursuing research-based master's and doctoral degrees in science, technology, engineering, and mathematics (STEM) and in STEM education. The GRFP provides three years of financial support for the graduate education of individuals who have demonstrated their potential for significant research achievements in STEM and STEM education. This award supports the NSF Graduate Fellows pursuing graduate education at this GRFP institution. This award reflects NSF's statutory mission and has been deemed worthy of support through evaluation using the Foundation's intellectual merit and broader impacts review criteria.

(국립과학재단(NSF)의 대학원 연구 장학금 프로그램(GRFP)은 매우 경쟁률이 높은 연방 장학금 프로그램이다. GRFP는 미국의 과학기술 인력의 활력과 다양성을 제고하는 데 기여한다. 이 프로그램은 과학, 기술, 공학, 수학(STEM) 및 STEM 교육 분야에서 연구중심 석사 및 박사 과정에 진학하는 뛰어난 대학원생을 발굴하여 지원한다. GRFP는 STEM 및 STEM 교육에서 상당한 연구 업적을 위한 잠재력을 입증한 개인들의 대학원 교육을 위해 3년간의 재정적인 지원을 한다. 이 과제는 GRFP 기관에서 대학원 과정을 수강하는 NSF 대학원생 펠로우를 지원한다. 이 과제는 NSF의 설립 미션을 반영하며, 재단의 지적 수월성과 파급효과에 대한 리뷰 기준을 통한 평가를 거쳐 선정된 과제이다.

○ 상기 분석 결과는 해당 과제들을 제거하지 않은 분석 결과로, 해당 과제는 총 157개이며, 대부분이 3, 5, 11번 클러스터에 분포해있음(그림 4-11의 중앙 부분)

- ‘research’, ‘engineering’ 등의 일반적인 단어들로 구성되어 있어 doc2vec 학습 결과 어떤 검색어로 상위 유사도 과제를 추출하더라도 일정 확률로 나타나게 됨
- 이 과제들의 검색어(breeding)와의 유사도(cosine similarity) 평균은 0.494±0.022로, 1,000개 전체 평균(0.486±0.022)보다 약간 높은 수준임

○ 각 클러스터를 검색어와의 코사인 유사도의 평균 순으로 정렬한 결과, 실제 관련성이 높은 0, 1, 2, 6, 10번 클러스터가 상위에 있었음(표 4-19)

- 각 클러스터에 포함된 과제들의 코사인유사도 값들을 평균낸 결과, 2, 6, 3, 1, 0, 10번 클러스터 순이었는데, 표준편차를 고려하면 통계적 유의미성이 높지는 않다는 점을 고려할 필요
- 단, 장학금 과제가 많은 3번 클러스터도 상위에 있었다는 점은 향후 데이터 전처리 과정에서 장학금 과제들을 제외시켜야 한다는 것을 시사함

〈표 4-19〉 NSF 2015~2019년 대상 ‘breeding’ 1,000개 과제 검색결과의 클러스터별 코사인유사도 평균 및 표준편차

클러스터	과제 수	유사도 평균	유사도 표준편차
2	76	0.496542	0.030052
6	106	0.4914	0.027783
3	137	0.490193	0.022949
1	66	0.489725	0.025886
0	50	0.487924	0.024994
10	73	0.485551	0.021439
11	99	0.482684	0.019059
5	127	0.482608	0.018377
9	59	0.481879	0.017374
7	65	0.478842	0.014837
4	63	0.478562	0.015259
8	79	0.477309	0.014923
총합계	1,000	0.485635	0.022497

❍ 위에서 언급한 장학금 과제들("Fellowship"이라는 단어가 들어간 과제)을 제외하고 클러스터별 유사도의 평균을 계산할 경우, 실제 육종과 관련된 것으로 보이는 0, 1, 2, 6, 10번 클러스터가 상위 5개에 위치함

- 0, 1, 2, 6, 10번 클러스터에도 소수(10개 미만)의 장학금 과제가 포함되어 있었으므로 과제 수가 다소 줄어든 것을 확인할 수 있음

〈표 4-20〉 NSF 2015~2019년 대상 'breeding' 1,000개 과제 검색결과(일반 장학금 과제 제외)의 클러스터별 코사인유사도 평균 및 표준편차

클러스터	과제 수	유사도 평균	유사도 표준편차
2	71	0.496349	0.03004
6	101	0.491226	0.028296
1	64	0.489883	0.026263
0	49	0.486588	0.023413
10	65	0.486411	0.021521
3	101	0.482101	0.019304
9	58	0.481999	0.017499
5	69	0.480403	0.019253
11	66	0.47868	0.017087
4	63	0.478562	0.015259
7	61	0.478147	0.014724
8	75	0.476732	0.014399
총합계	843	0.484136	0.022321

❍ 각 클러스터의 세부적인 주제를 탐색하기 위해 과제들을 검토한 결과, 0, 1, 2, 6, 10번 클러스터에도 육종과 관련되지 않은 과제들이 상당 수 관찰되어 주제 판정이 어려웠음

- 2번 클러스터를 제외하고는 주제를 파악하기 어렵거나, 육종보다는 자연생태계 또는 진화에 대한 연구가 주를 이루는 클러스터가 대다수였음
- 육종은 자연적으로 일어나는 진화를 인위적으로 조작하여 인간이 원하는 형질을 얻어내는 행위기 때문에, 생태(ecology)나 진화(evolution)와 관련성이 높다고 볼 수 있음

- NSF가 지원하는 연구가 산업적 목적보다는 기초연구 자체를 표방하고 있기 때문에, 특정 작물의 육종에 대해서 연구하기보다는 기초적인 진화나 생태에 대해서 연구하는 과제가 다수인 것은 자연스러워 보임
- 육종과 관련성이 높은 2번 클러스터의 경우, 작물 또는 테크닉 면에서 여러 가지 과제들이 혼합되어 있는 경향을 보여 더 세부적인 분석이 필요함

〈표 4-21〉 NSF 2015~2019년 'breeding' 1,000개 과제 클러스터별 주제

클러스터	주제
0	주제 통일성 없음
1	생태, 진화
2	육종 관련 기초연구(더 세부적으로는 파악 불가)
6	생태, 진화
10	주제 통일성 없음

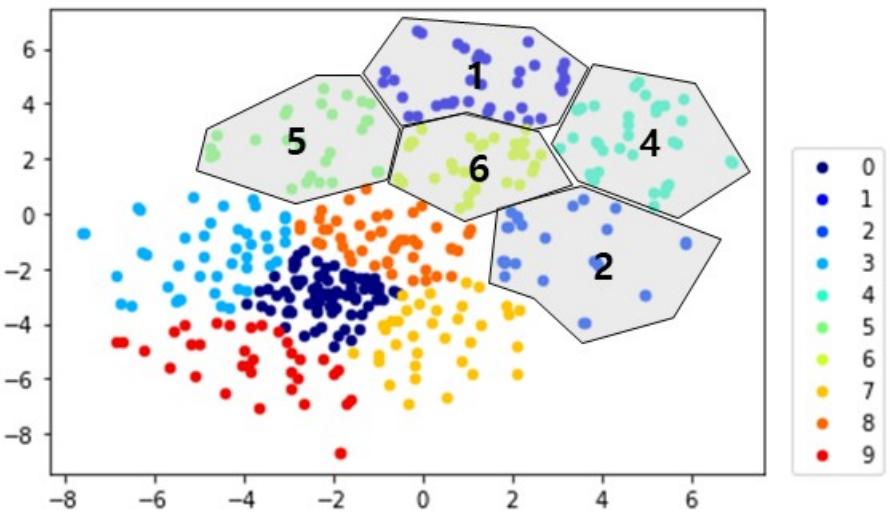
❍ 따라서, 검색결과 수를 400개로 줄이고 10개 클러스터로 분리하여 분석을 수행한 결과, 작물의 육종에 직접적으로 관련되어 있을 가능성이 높은 클러스터가 관찰되었음

- 해당 클러스터는 1, 2, 4, 5, 6번이며, 우측 상단 절반을 차지하고 있음을 알 수 있음(표 4-22, 그림 4-12)

〈표 4-22〉 NSF 2015~2019년 'breeding' 400개 과제 클러스터(10개)별 주요 키워드

클러스터	LDA 키워드
0	['graduate' 'stem' 'br' 'new' 'research' 'programming' 'resistance' 'education' 'grfp' 'students' 'data' 'mentoring' 'host' 'project' 'fuel' 'fellowship' 'theory' 'program' 'languages' 'parasites']
1	['br' 'drought' 'research' 'species' 'project' 'seed' 'plant' 'yam' 'traits' 'mucilage' 'size' 'stress' 'recombination' 'new' 'tolerance' 'plants' 'population' 'mosses' 'genome' 'genetic']
2	['plant' 'br' 'project' 'research' 'species' 'reproductive' 'cell' 'engineering' 'domestication' 'new' 'salt' 'genetic' 'changes' 'crop' 'responses' 'animal' 'conference' 'tolerance' 'farming' 'light']

클러스터	LDA 키워드
3	['students' 'br' 'research' 'project' 'temperature' 'stem' 'water' 'analysis' 'imagery' 'upwelling' 'genetic' 'georgia' 'irrigation' 'conference' 'disease' 'understanding' 'fellowship' 'biomass' 'forest' 'imaging']
4	['species' 'research' 'br' 'project' 'temperature' 'plant' 'genetic' 'plants' 'birds' 'traits' 'evolution' 'microbes' 'diversity' 'host' 'self' 'insects' 'evolutionary' 'using' 'including' 'new']
5	['br' 'genetic' 'species' 'project' 'gene' 'research' 'evolution' 'phloem' 'new' 'reproduction' 'plant' 'notothenioid' 'learning' 'data' 'rate' 'parasites' 'anatomy' 'including' 'memory' 'expression']
6	['species' 'plant' 'tomato' 'br' 'research' 'project' 'plants' 'data' 'genome' 'rice' 'photosynthesis' 'stress' 'maize' 'tcn' 'sequence' 'use' 'used' 'genes' 'heat' 'genetic']
7	['br' 'nitrogen' 'research' 'project' 'carbon' 'deposition' 'graduate' 'drought' 'data' 'peromyscus' 'travel' 'gut' 'program' 'nsf' 'biology' 'science' 'support' 'sleep' 'soil' 'dust']
8	['project' 'br' 'research' 'new' 'soil' 'alleles' 'graduate' 'division' 'data' 'protein' 'plant' 'fellowship' 'abilities' 'amino' 'award' 'soybean' 'host' 'program' 'sciences' 'genes']
9	['research' 'br' 'software' 'project' 'stem' 'conference' 'cosmic' 'genes' 'change' 'power' 'support' 'wildlife' 'models' 'students' 'workshop' 'materials' 'theory' 'researchers' 'feedback' 'imprinted']



[그림 4-12] NSF 2015~2019년 'breeding' 400개 과제 검색 결과

❍ 각 클러스터의 과제를 검토하면서 주제를 파악한 결과, 주요 키워드에서 나타난 것과 마찬가지로 1, 2, 4, 5, 6번 클러스터와 육종의 관련성이 높았는데, 4, 5번 클러스터의 경우 작물 육종보다는 진화, 생태 관련 기초연구에 해당함(표 4-23)

- 이는 4, 5번 클러스터의 주요 키워드에 관련 키워드가 더 적은 결과와도 맥락을 같이 함

〈표 4-23〉 NSF 2015~2019년 'breeding' 400개 과제 클러스터별 주제

클러스터	주제
0	Graduate Research Fellowship Program 이 절대 다수
1	작물 육종, 생태 연구 혼합
2	작물 육종, 생태 연구 혼합
3	주제 통일성 없음
4	진화, 생태
5	진화, 생태
6	작물 육종, 생태 연구 혼합
7	주제 통일성 없음
8	주제 통일성 없음
9	주제 통일성 없음

❍ 하지만 클러스터를 상당히 세분화 시켰음에도 육종에 직접적으로 관련되어 있는 클러스터를 정확히 구분하기는 어려운 것으로 판명됨

- 1, 2, 6번 클러스터는 작물 육종 관련 연구과제들도 포함하고 있었지만 생태나 자연적 진화에 대한 연구도 혼합되어 있었음
- 1, 2, 6번 클러스터의 과제 수는 각각 36, 26, 34개로 상당히 적으며, 그림 39 상에서 해당 클러스터의 과제들이 분산되어 있는 편임을 감안할 때 육종 과제들은 다소 흩어져 클러스터링이 잘 이루어지지 않는 것으로 판단됨

❍ 이러한 사유는 NSF 기관의 특성상 농산업의 직접적 적용을 위한 목적형 연구보다는 기초연구를 지원하는 비중이 크기 때문에 육종에 해당하는 과제 자체가 적기 때문으로 사료됨

- NSF의 Award Search 사이트에서 advanced search 기능을 사용하여 2015.1.1.~2019.12.31. 사이에 수행된 모든 과제를 검색한 결과 총 438개의 과제가 검색됨
  ※ 연구시작일이 2015.1.1. 이후이면서 연구종료일이 2019.12.31. 이전인 과제를 검색
- 438개 과제의 대과제(Program)명을 볼 때, 동식물 육종에 관련된 과제는 이 중 소수로 파악되며, 대과제 안에서도 육종의 목적보다는 자연현상 연구에 가까워보이는 과제들이 혼합되어 있는 것으로 보임(표 4-24)
- 과제 수 상위 10개 대과제 중 육종에 관련이 있는 것으로 보이는 대과제는 3개 정도임(표 4-24에 표시함)

〈표 4-24〉 NSF 2015~2019년 'breeding' 관련 상위 10개 대과제

대과제(Program) 명	과제 수
Plant Genome Research Project(식물 게놈 연구 프로젝트)	57
Animal Behavior(동물 행동)	27
NPGI PostDoc Rsrch Fellowship(NPGI 포닥 장학금)	22
Physiolgcl Mechnsms&Biomechnsm(생리학적 기작 & 바이오메커니즘)	11
ANT Organisms & Ecosystems(ANT 생물체 & 생태계)	11
EVOLUTIONARY GENETICS(진화유전학)	10
Integrtv Ecological Physiology(통합적 생태 생리학)	9
Plant Genome Research Resource(식물게놈 연구자원)	8
PHYLOGENETIC SYSTEMATICS(시스템 계통유전학)	8

❍ NSF Award Search 기능을 통해 키워드 검색으로 수집한 과제와 본 연구의 알고리즘으로 검색한 결과를 비교한 결과, 동물이나 곤충 육종에 대한 과제는 누락되는 것이 발견됨

- 대과제 중 "Animal Behavior"에 해당하는 과제들은 동물이나 곤충의 육종에 연관성이 있는데, 해당 과제들은 동 알고리즘의 검색 결과에 거의 나타나지 않았음

- 인공신경망의 특성상 이것의 이유를 정확히 판명하기는 어려우나, 향후 개선이 필요한 사항임

❍ NSF의 경우 '육종'과 같은 구체적인 단어로 검색하기는 적합치 않은 것이 드러남

- 관련성이 높은 과제가 실제로 매우 적은 경우, 어느정도 노이즈가 존재하는 동 검색 알고리즘의 특성상 충분히 동작하지 않음
- 동 알고리즘은 소수의 과제를 정확히 찾아내기보다는 거시적인 그룹핑에 더욱 적합함

□ 기초연구를 폭넓게 지원하는 NSF 기관 특성을 반영하여, 포괄적인 검색어를 통하여 보다 거시적인 분석을 시험하기 위해 "biology(생물학)"에 대한 검색을 수행하였으며, 2차원 축소 이전에 클러스터링을 수행하였음

❍ 5,000개의 광범위한 검색 결과를 2차원으로 축소할 경우, 충분히 클러스터링이 수행되지 않을 수 있는 점을 고려하여, 정보가 손실되기 전인 200차원 상태에서 k-means 클러스터링을 수행함

- 그룹을 먼저 나눈 후, t-SNE로 차원축소를 실행(perplexity = 30, 반복수 = 1,000)함으로써 차원축소 과정에서 클러스터간 관계성이 잘 보존되는지, 클러스터링 자체는 잘 일어나는지 분석함
- 각 클러스터의 주제가 명확히 파악되는지 평가함

❍ 2015~2019년 과제를 대상으로 5,000개의 과제를 검색하고 20개의 클러스터로 분류한 결과, 주요 키워드는 표 4-25과 같았음

※ 클러스터들의 내용이 광범위할 것으로 예상되어 주요 키워드를 50개씩 표시하도록 함

- 각 클러스터의 주요 키워드를 기준으로 판단하였을 때, 생물학이 아닌 우주(지구과학)라거나 물리학에 관련된 과제들도 상당수 검색되었지만 클러스터링 되어 제외 가능하게 분리되었음이 관찰됨
- '생물학'으로 검색하였지만, 검색 결과 수학, 물리학과 같은 다른 기초과학에 관련된 과제 및 수학 · 과학 등을 가르치는 STEM 교육에 관련된 연구도 많이 검색된 것을 알 수 있음

〈표 4-25〉 NSF 2015~2019년 'biology' 5,000개 과제 클러스터(20개)별 주요 키워드

클러스터 (주제)	LDA 키워드
0 (식물 유전체 관련)	['research' 'br' 'cell' 'amino' 'project' 'protein' 'students' 'proteins' 'plant' 'rna' 'tools' 'acid' 'biology' 'signaling' 'iron' 'circular' 'gene' 'transcription' 'function' 'new' 'genes' 'cells' 'dna' 'molecular' 'high' 'division' 'genetic' 'understanding' 'development' 'expression' 'networks' 'ethylene' 'methods' 'rnap' 'csld' 'plants' 'regulation' 'human' 'asymmetric' 'also' 'species' 'data' 'provide' 'mathematical' 'hydrogen' 'maize' 'polarity' 'different' 'using' 'important']
1 (물리화학 관련)	['systems' 'br' 'research' 'models' 'new' 'pools' 'project' 'phase' 'dynamics' 'cold' 'quantum' 'adsorption' 'dipole' 'nonlocal' 'data' 'understanding' 'using' 'clouds' 'cortical' 'theory' 'fluid' 'computational' 'many' 'internal' 'monsoon' 'particles' 'equations' 'state' 'provide' 'energy' 'modeling' 'analysis' 'surface' 'structures' 'properties' 'statistical' 'distribution' 'two' 'interface' 'applications' 'mathematical' 'black' 'stochastic' 'physics' 'foam' 'flow' 'interactions' 'suspensions' 'water' 'south']
2 (생화학 관련)	['research' 'br' 'oxygen' 'project' 'chemistry' 'students' 'hydrogels' 'synthesis' 'compounds' 'chemical' 'reactions' 'molecules' 'new' 'catalysts' 'program' 'high' 'polymerization' 'protein' 'activity' 'gold' 'work' 'systems' 'organic' 'mppp' 'cellulose' 'carbon' 'molecular' 'dna' 'production' 'bacteria' 'understanding' 'microbial' 'materials' 'dioxide' 'enzymes' 'development' 'using' 'properties' 'rna' 'catalysis' 'also' 'enzyme' 'graduate' 'acids' 'dr' 'rare' 'professor' 'active' 'university' 'plant']
3 (생물학 분야 학회 지원)	['biology' 'br' 'conference' 'workshop' 'research' 'plant' 'workshops' 'meeting' 'biological' 'cell' 'new' 'behavior' 'science' 'signaling' 'scientists' 'monographs' 'field' 'synthetic' 'engineering' 'gas' 'data' 'researchers' 'well' 'grs' 'symposium' 'ethylene' 'change' 'fellows' 'rural' 'development' 'learning' 'government' 'epigenetics' 'students' 'life' 'ice' 'collections' 'protocol' 'systems' 'understanding' 'discussion' 'chronobiology' 'grc' 'presentations' 'taxonomic' 'modeling' 'experts' 'hydroids' 'mitonuclear' 'metabolic']
4 (ICT-BT 융합)	['project' 'br' 'research' 'cancer' 'data' 'software' 'systems' 'new' 'networks' 'system' 'information' 'autonomous' 'infrastructure' 'smart' 'acid' 'nsf' 'high' 'control' 'science' 'linear' 'mismatch' 'university' 'iot' 'breast' 'brain' 'security' 'using' 'vehicles' 'impacts' 'performance' 'support' 'use' 'potential' 'wildlife' 'virtual' 'hardware' 'multi' 'design' 'lift' 'algorithms' 'learning' 'energy' 'decision' 'broader' 'memory' 'detection' 'ai' 'enable' 'large' 'students']

클러스터 (주제)	LDA 키워드
5 (환경, 생태)	['data' 'project' 'br' 'research' 'carbon' 'soil' 'water' 'climate' 'microbial' 'ocean' 'ice' 'sea' 'asm' 'wetland' 'activity' 'fracturing' 'nsf' 'understanding' 'environmental' 'ballast' 'system' 'denitrification' 'change' 'award' 'impacts' 'processes' 'students' 'biotite' 'vessel' 'dom' 'high' 'air' 'new' 'also' 'methods' 'iodine' 'co' 'isotope' 'plant' 'field' 'species' 'changes' 'nitrogen' 'support' 'biodiversity' 'fossil' 'weathering' 'oceanographic' 'provide' 'using']
6 (우주, 지구과학)	['research' 'br' 'astronomy' 'clusters' 'galaxies' 'quantum' 'data' 'earth' 'particle' 'physics' 'tsunami' 'black' 'students' 'quark' 'solar' 'science' 'astrophysics' 'waves' 'galaxy' 'supersubstorms' 'mineral' 'wave' 'project' 'stars' 'debris' 'gas' 'gravitational' 'award' 'gluon' 'new' 'large' 'dark' 'information' 'learning' 'particles' 'fiord' 'team' 'space' 'using' 'nuclear' 'simulations' 'observations' 'matter' 'study' 'elements' 'structures' 'magnetosphere' 'machine' 'physical' 'lhc']
7 (생물정보학)	['data' 'project' 'research' 'br' 'computational' 'new' 'analysis' 'models' 'human' 'high' 'systems' 'learning' 'algorithms' 'extreme' 'based' 'dimensional' 'model' 'methods' 'power' 'clustering' 'brain' 'framework' 'statistical' 'datasets' 'bayes' 'space' 'develop' 'using' 'modeling' 'robotic' 'objects' 'massive' 'network' 'disease' 'objective' 'events' 'dynamics' 'networks' 'large' 'small' 'mathematical' 'machine' 'applications' 'information' 'cancer' 'behavior' 'genetic' 'software' 'neural' 'uncertainty']
8 (수학, 물리학 분야 학회 지원)	['conference' 'br' 'university' 'workshop' 'research' 'students' 'topology' 'meeting' 'researchers' 'systems' 'theory' 'science' 'symposium' 'scientists' 'support' 'engineering' 'award' 'stem' 'geometry' 'dynamics' 'graduate' 'data' 'georgia' 'international' 'travel' 'energy' 'th' 'physics' 'analysis' 'quantum' 'participants' 'women' 'new' 'student' 'storage' 'nsf' 'mathematics' 'processes' 'computer' 'representation' 'topics' 'career' 'stdc' 'mentoring' 'field' 'rna' 'provide' 'held' 'mathematical' 'together']
9 (다양한 분야 장학금)	['research' 'conference' 'br' 'graduate' 'project' 'data' 'theory' 'fellowship' 'imaging' 'workshop' 'gravitational' 'grfp' 'operator' 'high' 'physical' 'chemistry' 'researchers' 'ice' 'software' 'program' 'students' 'nsf' 'geometry' 'cell' 'study' 'support' 'award' 'sciences' 'materials' 'algebras' 'postdoctoral' 'mathematical' 'stem' 'dynamics' 'new' 'hoxa' 'technology' 'university' 'manufacturing' 'redox' 'education' 'light' 'waves' 'host' 'methods' 'cells' 'system' 'information' 'temperature' 'scientific']

클러스터 (주제)	LDA 키워드
10 (STEM 교육)	['stem' 'students' project' 'research' 'faculty' 'teachers' 'engineering' 'undergraduate' 'support' 'br' 'program' 'master' 'need' 'education' 'high' 'conference' 'experiences' 'university' 'schools' 'student' 'science' 'school' 'mathematics' 'community' 'mentoring' 'success' 'teacher' 'scholars' 'teaching' 'college' 'ncuwm' 'development' 'doctoral' 'nreup' 'retention' 'advance' 'noyce' 'biology' 'data' 'women' 'mathematical' 'nsf' 'underrepresented' 'based' 'programs' 'tutorials' 'future' 'chemistry' 'graduate' 'geoscience']
11 (진화, 생태)	['project' 'br' 'species' 'research' 'traits' 'gene' 'tcn' 'data' 'evolution' 'genetic' 'host' 'human' 'study' 'plant' 'using' 'biology' 'students' 'collections' 'diversity' 'seed' 'evolutionary' 'expression' 'understanding' 'genome' 'fellowship' 'changes' 'also' 'specimens' 'coral' 'mucilage' 'new' 'regulatory' 'award' 'speciation' 'fishes' 'nsf' 'patterns' 'high' 'provide' 'physiology' 'learning' 'contact' 'genes' 'biological' 'california' 'science' 'tropical' 'movement' 'models' 'biodiversity']
12 (교육, 인문사회)	['stem' 'science' 'research' 'br' 'brain' 'discourse' 'health' 'children' 'data' 'language' 'project' 'education' 'engineering' 'matter' 'interpretation' 'surface' 'development' 'students' 'mathematics' 'community' 'information' 'online' 'learning' 'social' 'management' 'career' 'support' 'city' 'change' 'citizen' 'families' 'faculty' 'harassment' 'understanding' 'human' 'cajun' 'collections' 'design' 'climate' 'projects' 'parent' 'coastal' 'meta' 'particle' 'simultaneous' 'life' 'robotics' 'fisheries' 'new' 'polymer']
13 (수학)	['research' 'theory' 'br' 'manifolds' 'project' 'geometry' 'hypergraph' 'conference' 'algebras' 'mathematical' 'fano' 'problems' 'dimensional' 'new' 'study' 'women' 'first' 'spaces' 'analysis' 'group' 'algebraic' 'learning' 'geometric' 'pi' 'systems' 'finite' 'higher' 'stem' 'field' 'soft' 'number' 'models' 'fellowship' 'symplectic' 'random' 'mathematics' 'polynomials' 'topology' 'categories' 'adic' 'harmonic' 'equations' 'math' 'floer' 'hypergraphs' 'algorithms' 'applications' 'quantum' 'percolation' 'groups']
14 (물리 등 주제 혼합)	['research' 'students' 'program' 'br' 'halos' 'project' 'field' 'olfactory' 'data' 'arctic' 'science' 'reu' 'cybersecurity' 'dark' 'crocodilians' 'used' 'division' 'graduate' 'matter' 'conference' 'odor' 'telomeric' 'new' 'umass' 'galaxies' 'site' 'university' 'study' 'education' 'station' 'subduction' 'imaging' 'dna' 'sea' 'cshl' 'big' 'climate' 'training' 'interdisciplinary' 'stars' 'public' 'award' 'scientific' 'engineering' 'spinal' 'nsf' 'support' 'physics' 'based' 'sheet']

클러스터 (주제)	LDA 키워드
15 (세포 기초생물학)	['cell' 'br' 'mechano' 'research' 'cells' 'tissue' 'pressure' 'mammalian' 'project' 'membrane' 'new' 'cardiac' 'brain' 'tree' 'mechanical' 'protein' 'students' 'encapsulation' 'chromatin' 'biomanufacturing' 'development' 'biology' 'stem' 'engineering' 'arbor' 'function' 'advanced' 'biological' 'cytoskeleton' 'synthetic' 'bacteria' 'signaling' 'award' 'understanding' 'using' 'systems' 'division' 'experiments' 'hydrogel' 'cellular' 'living' 'rna' 'magnetic' 'based' 'mechanisms' 'tissues' 'conference' 'proteins' 'polymer' 'model']
16 (분석화학, 나노 융합)	['br' 'research' 'project' 'nanoparticles' 'dna' 'chemical' 'dynamics' 'single' 'infrared' 'cell' 'nom' 'new' 'award' 'students' 'proteins' 'structure' 'hydrogen' 'chemistry' 'quantum' 'imaging' 'absorption' 'protein' 'probe' 'spectrometer' 'molecular' 'rna' 'using' 'molecules' 'stabilities' 'science' 'release' 'species' 'membrane' 'high' 'used' 'electron' 'biological' 'methods' 'university' 'vibrations' 'system' 'energy' 'isoferritins' 'small' 'lipid' 'afm' 'instrument' 'surface' 'undergraduate' 'materials']
17 (다양한 분야 장학금)	['br' 'research' 'graduate' 'conference' 'stem' 'blood' 'fellowship' 'project' 'data' 'grfp' 'program' 'students' 'science' 'award' 'institution' 'mathematics' 'university' 'support' 'theory' 'workshop' 'materials' 'engineering' 'computational' 'mathematical' 'new' 'big' 'pressure' 'algebras' 'researchers' 'differential' 'physics' 'sciences' 'geometry' 'education' 'health' 'nsf' 'provide' 'also' 'biology' 'fellowships' 'chemistry' 'field' 'host' 'quantum' 'workforce' 'techniques' 'climate' 'understanding' 'using' 'postdoctoral']
18 (STEM 교육)	['students' 'stem' 'research' 'project' 'undergraduate' 'biology' 'course' 'science' 'br' 'software' 'data' 'learning' 'cs' 'mathematics' 'teachers' 'ipls' 'school' 'engineering' 'education' 'value' 'training' 'development' 'computer' 'models' 'cyle' 'workshop' 'scientific' 'communication' 'knowledge' 'resources' 'student' 'mathematical' 'courses' 'active' 'argumentation' 'high' 'thinking' 'modeling' 'physics' 'graduate' 'liberal' 'using' 'computational' 'network' 'curricular' 'change' 'scientists' 'program' 'environmental' 'instruction']
19 (소재 관련)	['materials' 'br' 'research' 'new' 'titanium' 'based' 'project' 'organic' 'design' 'fuel' 'properties' 'high' 'quantum' 'soft' 'ion' 'water' 'solid' 'nanoscale' 'nanomaterials' 'energy' 'engineering' 'hysteresis' 'students' 'biological' 'electrodes' 'cellulose' 'field' 'devices' 'resistance' 'alloys' 'robotics' 'cathode' 'solar' 'ferroelectric' 'nitride' 'award' 'thermal' 'science' 'magnetic' 'develop' 'young' 'applications' 'lithium' 'graphene' 'using' 'material' 'interfacial' 'development' 'polymers' 'surface']

〈표 4-26〉 NSF 2015~2019년 'biology' 5,000개 과제 20개 클러스터별 (평균 유사도 순) 과제 수 및 주제

클러스터	과제 수	평균 유사도	주제
9	410	0.598	다양한 분야 장학금
3	114	0.564	생물학 분야 학회 지원
8	447	0.561	수학, 물리학 분야 학회 지원
14	308	0.560	물리 등 주제 혼합
0	178	0.554	식물 유전체 관련
17	389	0.552	다양한 분야 장학금
18	200	0.552	STEM 교육
15	148	0.551	세포 기초생물학
13	286	0.551	수학
12	181	0.550	교육, 인문사회
11	330	0.549	진화, 생태
10	292	0.548	STEM 교육
16	164	0.547	분석, NT융합
5	288	0.545	환경, 생태
1	199	0.545	물리화학 관련
6	135	0.545	우주, 지구과학
7	256	0.544	생물정보학
19	217	0.544	소재 관련
4	301	0.544	ICT-BT 융합
2	157	0.543	생화학 관련
총합계	5,000	0.554	

❍ 각 클러스터의 검색어(biology)와의 평균 유사도를 계산할 시, 예상한 것과 다르게 일반적인 장학금이나 학회 지원에 관한 클러스터가 더 상위에 있고, 생물학의 하위 분야에 해당하는 환경, 생태나 생화학 관련 클러스터의 유사도가 낮았음(표 4-26)

- 이는 실제로 생물학을 연구하는 과제들은 'biology'라는 단어보다는 더 구체적인 단어를 과제명이나 요약문에 기재하기 때문으로 생각됨
- 오히려 STEM 교육과 같은 인문사회 분야에서 '생물학'이라는 포괄적인 단어를 더 많이 사용하기 때문으로 추측됨

❍ 또한 상위 유사도 과제를 보았을 때, 수학에 대한 과제도 수학의 생물학적 적용 등 생물학과 관련이 있는 것으로 보이므로(표 4-27), 클러스터의 주요 키워드만 보고 정확히 주제를 잡아내기에는 어려운 점이 존재함

- 점차 학문간의 융합이 가속화되고 있기 때문에 이러한 어려움은 심화되서 보다 고차원적인 분석 방법이 필요할 것으로 생각됨
- 분석 결과의 신뢰성 및 정확성 제고를 위해 분야의 지식 체계(domain knowledge)를 결합할 필요

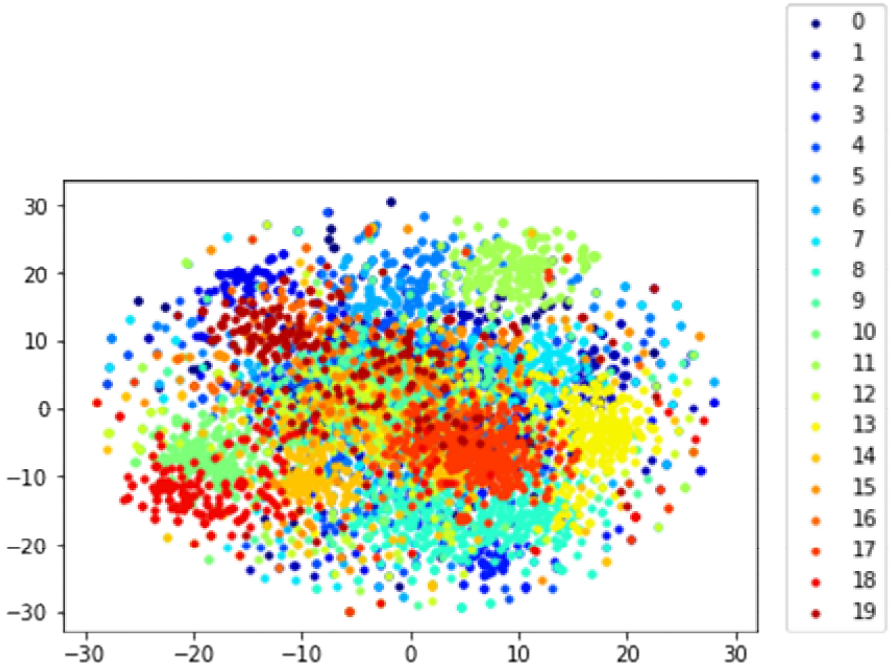
〈표 4-27〉 NSF 2015~2019년 'biology' 5,000개 과제 중 유사도 상위 10개 과제

유사도	과제명	소속 클러스터
0.723	Geometry, Topology and Complexity of Manifolds, and Applications to Biology (기하학, 위상학, 매니폴드의 복잡도 및 생물학적 적용)	9
0.682	Meeting: Evolutionary Impacts of Seasonality; a Symposium for the Annual Meeting of SICB, New Orleans, LA, Jan 4-8 2017 (학회: 계절이 진화에 끼치는 영향; SICB 연례 학회 심포지움)	3
0.679	Workshop on Future Directions in Network Biology (네트워크 생물학의 향후 방향에 대한 워크샵)	8
0.679	Targeted Infusion Project: Innovative Jarvis Undergraduate Mathematics Program(I-JUMP): Embedding Computational and Mathematical Biology into Life STEM (목적 지향 확산 프로젝트: 자비스 혁신 학부생 수학 프로그램(I-JUMP): 계산/수학 생물학의 생물 STEM 융합)	18
0.678	CONFERENCE: Post-transcriptional Gene Regulation in Plants to be held July 14-15, 2016 at the Austin Convention Center in Austin, TX (학회: 식물 내 전사된 유전체의 조절)	3

유사도	과제명	소속 클러스터
0.669	REU Site: Environmental Biology for Pacific Islanders (REU Site: 태평양 섬 지역을 위한 환경 생물학)	14
0.666	2018 FASEB Summer Research Conference on Yeast Chromosome Biology & Cell Cycle; July 15-20, 2018; Steamboat Springs, Colorado (효모 염색체 생물학, 세포 사이클에 관한 2018 FASEB 여름 학회)	3
0.664	EAGER: STEM NOLA: An Exploration of an Urban Community-Based STEM Learning Model (EAGER: STEM NOLA: 도시 커뮤니티 기반 STEM 학습 모델 탐색)	12
0.658	Graduate Research Fellowship Program(GRFP) 대학원생 연구 장학생 프로그램(GRFP)	9
0.658	NSF Student Travel Grant for the 2018 International Workshop on Bio-Design Automation(IWBDA) 2018 바이오 디자인 자동화 국제 워크샵를 위한 NSF 학생 출장비 지원사업	8

❍ 본 분석에서는 200차원의 데이터를 2차원으로 변환하기 전에 클러스터링을 수행하였는데, 2차원 상태에서 각 클러스터들은 일부 지역적인 위치를 가지고 있으나, 중앙부의 경우 여러 클러스터들이 서로 겹쳐서 뭉쳐져 있음

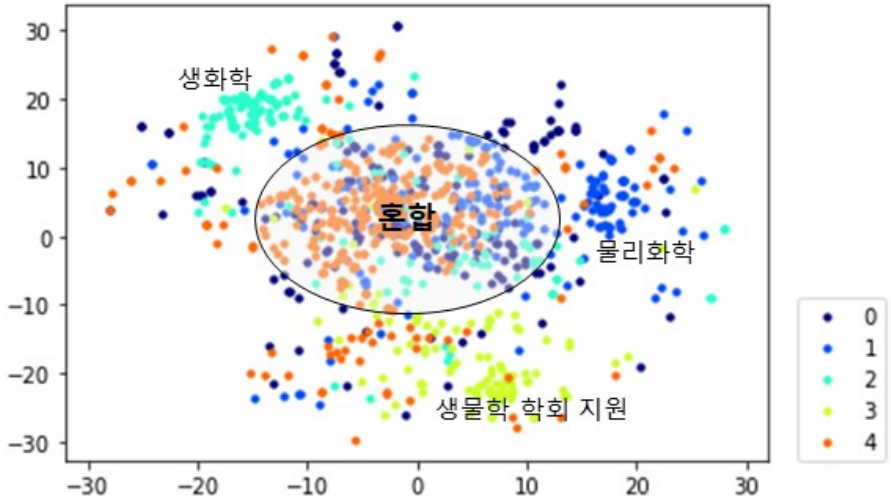
- 각 과제들이 매우 명확하게 분류되지 않고 서로 연속성을 가지는 상태에서 다량의 고차원 데이터 점들을 2차원으로 축소하여 표현하는 것에는 한계가 있음을 보여주는 예시라고 볼 수 있음
- 그림 73에서 볼 수 있듯, 20개의 클러스터가 서로 겹쳐서, 클러스터 간의 영역을 명확하게 관찰하기 어렵기 때문에, 5개 클러스터씩 분리하여 표시함 (그림 4-13~17)
- 지면 상에 표현하기 불편한 단점이 있지만, 데이터의 특성을 명확히 보기 위해서는 2차원이 아닌 3차원으로의 차원축소를 통해 정보 손실을 줄일 필요가 있음



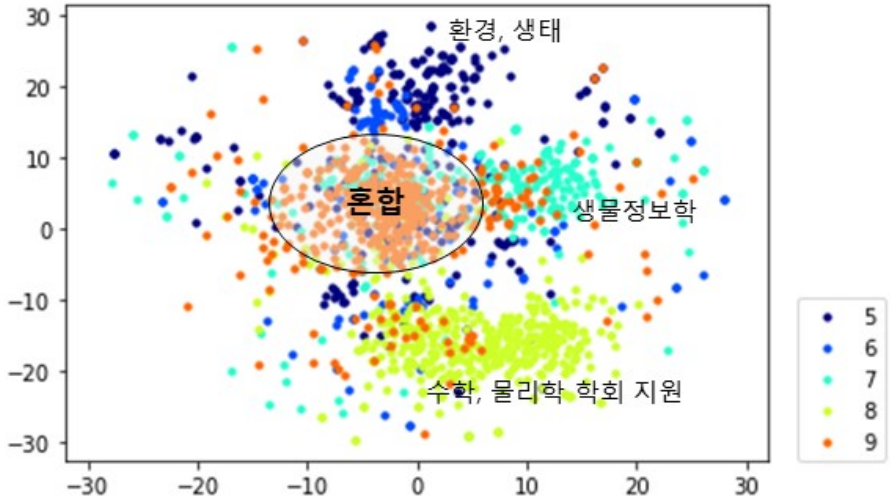
[그림 4-13] NSF 2015~2019년 'biology' 5,000개 과제 검색 결과 (차원 축소 이전 클러스터링 수행)

❍ 클러스터들 중 높은 응집성을 가지는 클러스터들은 주로 주변부에 위치하고 있으며, 중앙부에는 산개되어 과제 간 응집성이 낮은 클러스터들이 위치하고 있음

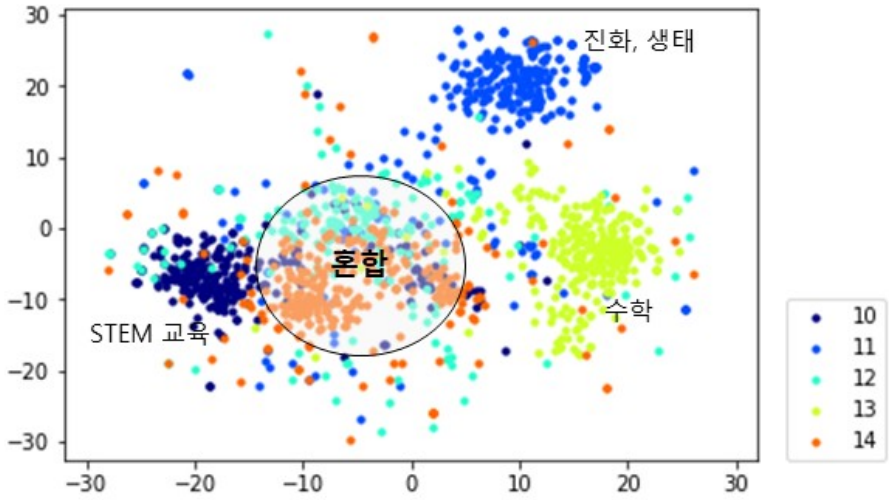
- 그림 4-14의 0, 1, 4번 클러스터는 그래프 전역에 산개되어 서로 겹쳐져 있는 모양을 띠는 반면, 2, 4번 클러스터는 상대적으로 응집성이 높음
- 그림 4-15의 5, 7, 8번 클러스터는 각자의 위치를 가지고 있으나, 6, 9번 클러스터는 중앙에 겹쳐있음
- 그림 4-16의 10~14번 클러스터는 12번을 제외하고는 응집성이 높은 편이며, 12, 14번이 서로 겹쳐있는 특성을 보임
- 그림 4-17의 17, 18번 클러스터는 응집성이 높지만, 15, 16, 19번 클러스터는 좌측 상단부에 혼합되어 있음



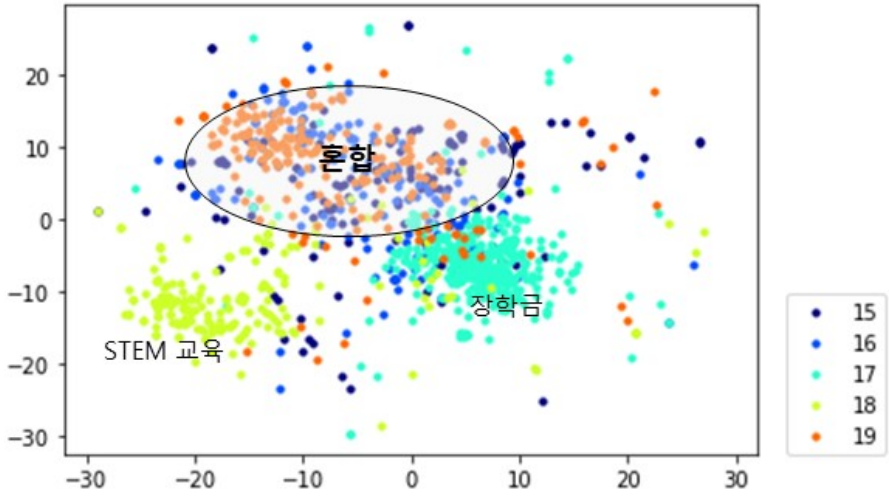
[그림 4-14] NSF 2015~2019년 'biology' 5,000개 과제 검색 결과
(클러스터 0~4번)



[그림 4-15] NSF 2015~2019년 'biology' 5,000개 과제 검색 결과
(클러스터 5~9번)



[그림 4-16] NSF 2015~2019년 'biology' 5,000개 과제 검색 결과 (클러스터 10~14번)



[그림 4-17] NSF 2015~2019년 'biology' 5,000개 과제 검색 결과 (클러스터 15~19번)

❍ 위치에 따른 내용 차이가 존재하여, 위쪽에는 주로 환경, 진화, 생태, 우측 하단은 수학이나 물리학, 좌측 하단은 STEM 교육에 관련된 과제들이 위치함

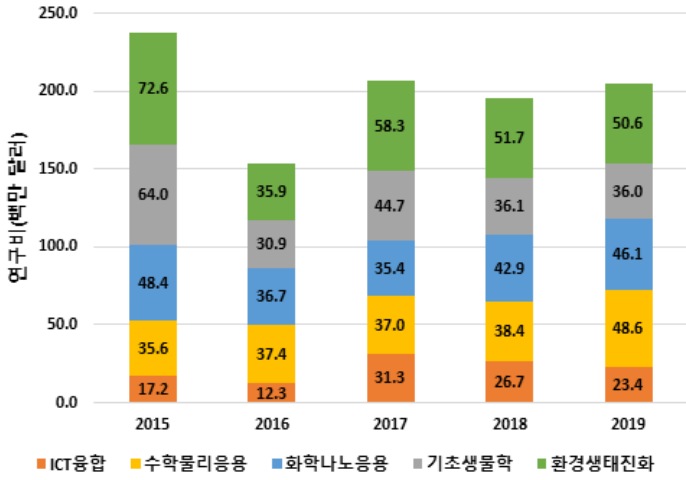
- 5, 11번 클러스터는 환경, 생태, 진화에 대한 내용이며, 그림의 상단~우측 상단에 분포하고 있음
- 1(물리화학), 7(생물정보학), 13(수학)번 클러스터는 수학이나 물리학이 응용되는 분야로 모두 그림의 우측부에 위치함
- 3, 8번 클러스터는 모두 학회 개최 비용을 지원하는 과제들로 그림의 하단부에 위치함
- 10, 18번 클러스터는 STEM 교육에 관련된 내용으로 그림의 좌측부에 모여있음
- 그림의 좌측 상단의 경우 2, 16, 19번 클러스터가 위치한다고 볼 수 있는데, 각각 생화학, 분석화학·NT융합, 소재 관련 클러스터로 화학 · 나노와 연관성이 높은 내용들이 위치하는 것으로 보임
- 나머지(0, 4, 6, 9, 12, 14, 15, 17) 클러스터의 경우 위치적 특성 없이 전체적으로 산개되어 있는 형태임

❍ 상기 분석 결과를 기준으로 NSF의 생물학 분야 투자 현황을 분석한 결과, 전체 투자 규모는 정체 상태이며, 기초생물학 보다는 수학, 물리 또는 ICT, 화학, 나노 등을 응용하는 다학제 연구의 비중이 증가하는 추세로 나타남(표 4-28, 그림 4-18)

- 클러스터 중 생물학과 명백히 관련되지 않거나 일반적 장학금, 학회 개최비, 인문사회에 해당하는 클러스터(3, 6, 8, 9, 10, 12, 14, 17, 18)를 제외한 나머지 클러스터를 유사한 내용끼리 그룹화하고 과제 수와 연구비를 계산함

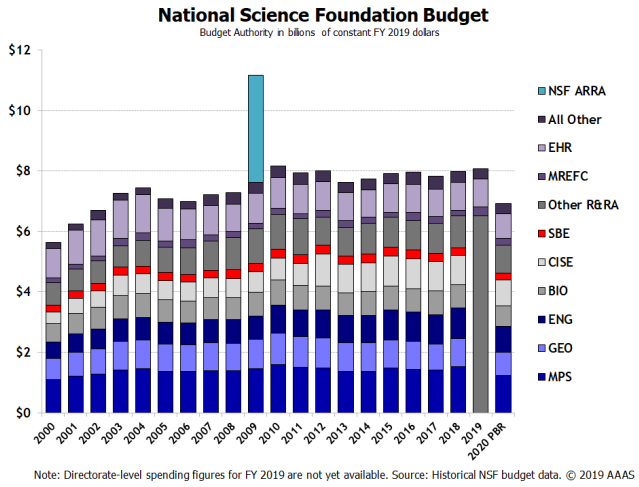
〈표 4-28〉 NSF 2015~2019년 'biology' 관련 과제 통계

분야	클러스터	과제 수(개)					연구비(백만 달러)				
		2015	2016	2017	2018	2019	2015	2016	2017	2018	2019
ICT 융합	4	62	44	47	77	71	17.2	12.3	31.3	26.7	23.4
수학물리응용	1, 7, 13	146	139	115	162	179	35.6	37.4	37.0	38.4	48.6
화학나노응용	2, 16, 19	101	104	92	118	123	48.4	36.7	35.4	42.9	46.1
기초생물학	0, 15	61	68	66	64	67	64.0	30.9	44.7	36.1	36.0
환경생태진화	5, 11	139	111	128	119	121	72.6	35.9	58.3	51.7	50.6



[그림 4-18] NSF 2015~2019년 생물학 관련 분야 투자 추이(추정)

❍ 이 결과는 최근 NSF 중 생물학 분야(BIO)의 예산이 전체적으로 정체 내지 줄어드는 추세인 것(그림 4-19)과 부합하며, 최근 다학제적 연구의 중요성 증대가 실제 과제 선정에 반영된 것으로 사료됨



출처 : Association of American Universities(2019), National Science Foundation(NSF) AAU FY20 Funding Brief

[그림 4-19] 2000~2020년 NSF 예산 추이

다. 영국 UKRI(영국 연구혁신기구) 과제 분석

□ UKRI 과제 대상으로는 암(cancer) 관련 연구 현황을 분석하였음

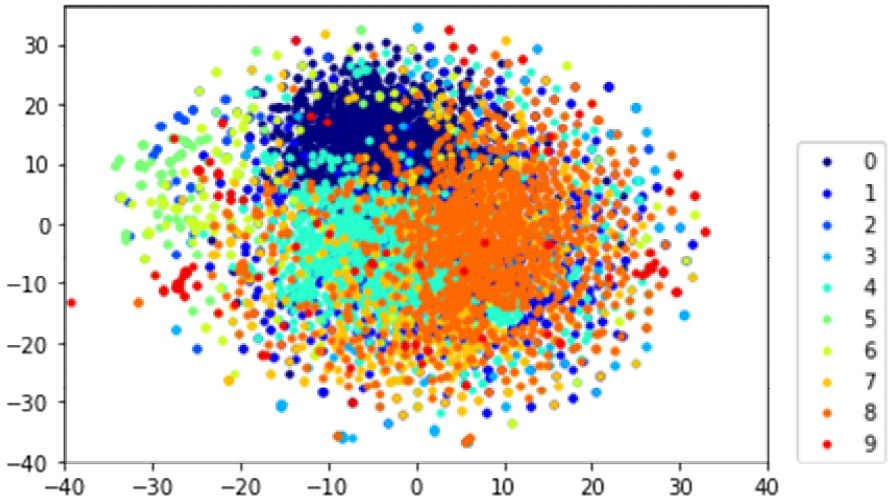
○ 검색어 'cancer'로 2015~2019년 수행 과제를 대상으로 5,000개의 과제를 검색하였고, 10개의 클러스터로 분류하여 거시적인 경향을 관찰함

- 10개 클러스터 중 암과 직접적인 관련이 있는 클러스터는 2, 5, 6, 9번 클러스터로 보임(표 4-29)

〈표 4-29〉 UKRI 2015~2019년 'cancer' 5,000개 과제 클러스터(10개)별 주요 키워드

클러스터 (주제)	LDA 키워드
0	['technology' 'new' 'project' 'training' 'test' 'physics' 'engine' 'university' 'dark' 'atlas' 'growth' 'limited' 'survey' 'development' 'develop' 'cancer' 'research' 'organisation' 'lhcb' 'lz' 'upgrade' 'heat' 'materials' 'system' 'diagnostic' 'code' 'doctoral' 'feasibility' 'security' 'embed']
1	['cost' 'training' 'project' 'research' 'new' 'develop' 'energy' 'development' 'system' 'using' 'data' 'university' 'cells' 'limited' 'pathways' 'enable' 'disease' 'hydrogen' 'organisation' 'health' 'technology' 'food' 'breast' 'immune' 'improve' 'power' 'ffir' 'use' 'uk' 'information']
2	['cancer' 'mutations' 'quot' 'cells' 'melanoma' 'disease' 'ds' 'virus' 'patients' 'hpv' 'hnscc' 'cell' 'cca' 'cancers' 'aim' 'awareness' 'expression' 'ebv' 'new' 'bladder' 'therapies' 'ras' 'stem' 'vietnam' 'malaysia' 'blood' 'development' 'haematuria' 'genes' 'tumour']
3	['physics' 'new' 'uk' 'research' 'memory' 'consortium' 'matter' 'climate' 'space' 'also' 'dark' 'energy' 'neutrino' 'pollution' 'materials' 'particle' 'neutrinos' 'coding' 'lhc' 'high' 'grid' 'lhcb' 'development' 'analysis' 'quot' 'air' 'large' 'single' 'project' 'models']
4	['training' 'circulation' 'new' 'development' 'industrial' 'application' 'atlas' 'dna' 'research' 'organisation' 'develop' 'cells' 'random' 'radiocarbon' 'cell' 'strains' 'lostbox' 'pprp' 'structure' 'cloudnc' 'epithelial' 'geometry' 'uk' 'larger' 'genes' 'immune' 'quot' 'business' 'funded' 'ac']
5	['cancer' 'tumour' 'cell' 'cells' 'specific' 'new' 'proteins' 'tnf' 'patients' 'cancers' 'tumours' 'wnt' 'death' 'models' 'stem' 'liver' 'treatment' 'tissue' 'mutations' 'immune' 'nk' 'genetic' 'metastatic' 'breast' 'development' 'wt' 'clinical' 'metabolic' 'protein' 'mina']

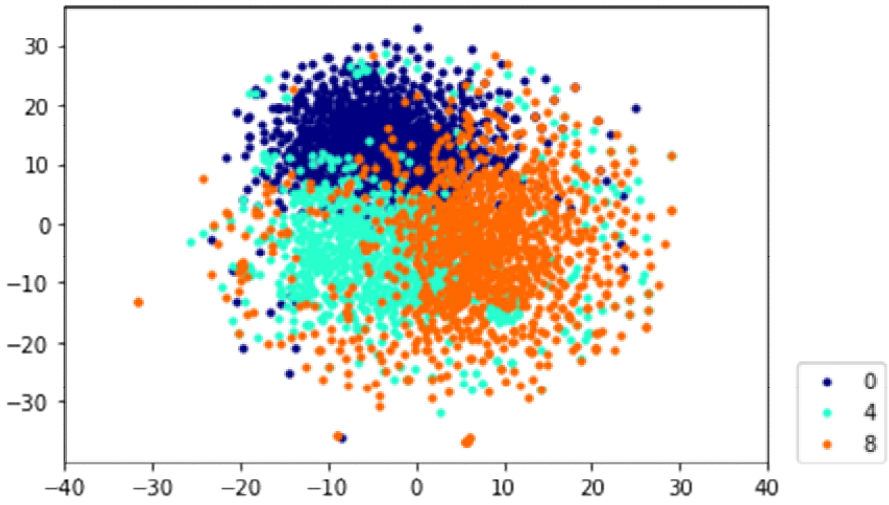
클러스터 (주제)	LDA 키워드
6	['cancer' 'cells' 'cell' 'new' 'data' 'imaging' 'disease' 'treatment' 'structural' 'research' 'patients' 'tissue' 'nanopore' 'develop' 'cancers' 'proteins' 'immune' 'brain' 'radiation' 'pet' 'health' 'technology' 'surgery' 'system' 'structure' 'use' 'project' 'detection' 'breast' 'drugs']
7	['materials' 'project' 'new' 'novel' 'high' 'energy' 'cancer' 'applications' 'training' 'research' 'technology' 'development' 'use' 'test' 'products' 'systems' 'enzymes' 'uk' 'screening' 'market' 'glycan' 'data' 'develop' 'cell' 'prostate' 'cost' 'product' 'using' 'manufacturing' 'based']
8	['training' 'high' 'university' 'research' 'manufacturing' 'project' 'limited' 'new' 'energy' 'modelling' 'micro' 'uk' 'organisation' 'develop' 'information' 'power' 'support' 'consortium' 'cells' 'transfer' 'materials' 'turbine' 'range' 'doctoral' 'studentshipterminology' 'groundwater' 'funded' 'lithium' 'see' 'development']
9	['cancer' 'dna' 'cells' 'cell' 'proteins' 'genome' 'replication' 'repair' 'damage' 'ageing' 'eif' 'genes' 'system' 'protein' 'new' 'ubiquitin' 'complex' 'rna' 'ber' 'germ' 'cdk' 'gene' 'human' 'mitochondrial' 'non' 'stem' 'expression' 'development' 'identify' 'autophagy']



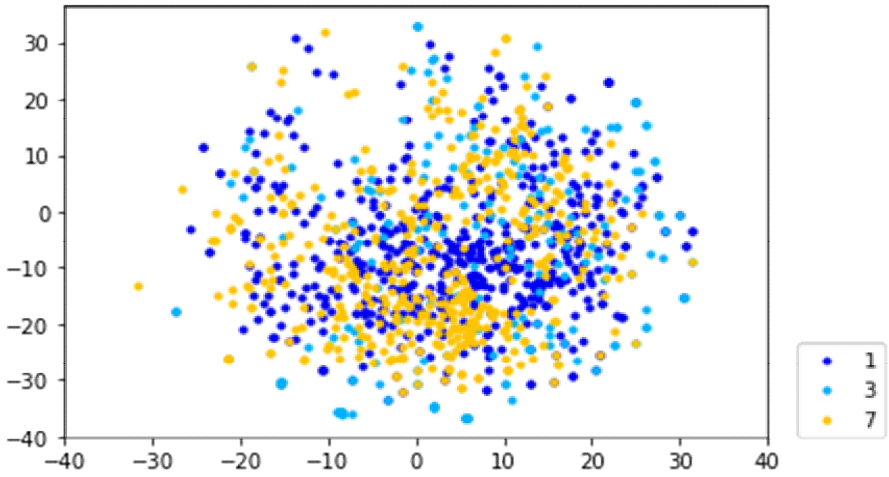
[그림 4-20] UKRI 2015~2019년 'cancer' 관련 과제 5,000개 (10개 클러스터로 분리) 시각화

❍ 분석 결과, 0, 4, 8번 클러스터를 제외하고는 모두 산개되어 있는 것처럼 서로 겹쳐있어 시각화가 용이하도록 3개의 그룹 클러스터들을 나누어 관찰하였음(그림 4-21~23)

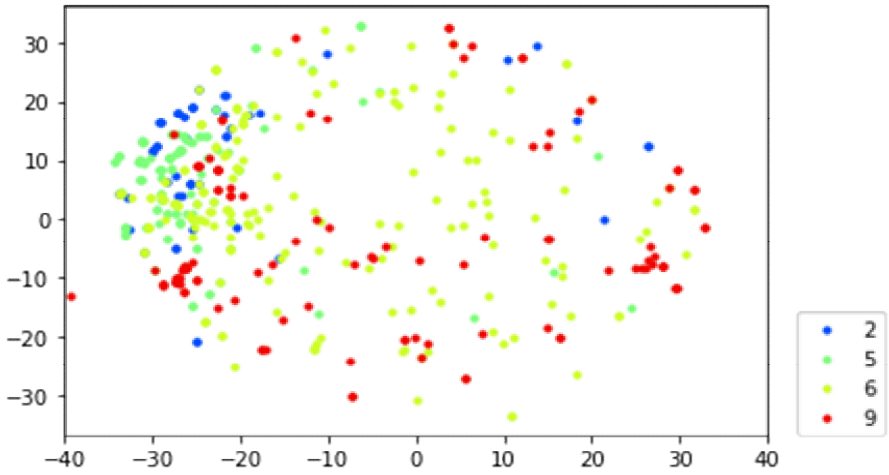
- 0, 4, 8번은 중앙부를 3등분하여 위치하고 있으며, 1, 3, 7번 클러스터는 전체적으로 산개되어 있음
- 검색어와 관련도가 높은 2, 5, 6, 9번 클러스터는 마찬가지로 산개되어 있으나 다소 가장자리에 모여있는 경향을 보임



[그림 4-21] UKRI 2015~2019년 'cancer' 관련 과제 5,000개 (0, 4, 8번 클러스터)



[그림 4-22] UKRI 2015~2019년 'cancer' 관련 과제 5,000개 (1, 3, 7번 클러스터)



[그림 4-23] UKRI 2015~2019년 'cancer' 관련 과제 5,000개 (2, 5, 6, 9번 클러스터)

❍ 검색 결과 중 다수의 과제가 UKRI에서 지원하는 박사후연구원(포닥) 연구비 지원 사업에 해당하여 다음과 같은 단편적인 초록만이 기재되어 있어, 분석 결과에서 제외해야 함이 드러남

Doctoral Training Partnerships: a range of postgraduate training is funded by the Research Councils. For information on current funding routes, see the common terminology at www.rcuk.ac.uk/StudentshipTerminology. Training grants may be to one organisation or to a consortia of research organisations. This portal will show the lead organisation only.
박사급 훈련 파트너십: 연구 위원회(Research Councils)에서는 다양한 포닥 훈련비를 지원함. 현재 지원되는 사업에 대한 정보는 www.rcuk.ac.uk/StudentshipTerminology 의 용어 설명에서 조회할 수 있음. 포닥 연구비는 단일 기관 또는 컨소시움에 지원됨. 이 포털에서는 주관 기관에 대한 정보만 제공함.

- 향후 분석 데이터 자체에서 이러한 과제들은 제거할 필요
- 이 사업에 해당하는 모든 과제가 동일한 요약문을 가지고 있어 학습 시 유사성이 높은 과제로 인식되며(과제명이 다르더라도), 이에 따라 학습 결과를 전반적으로 왜곡할 수 있기 때문임

❍ 이러한 과제(Doctoral Training Partnerships, DTP)는 총 2,846개로 클러스터 0, 4, 8번에 몰려있었으며, 이후 내용은 이러한 과제들은 제외하고 분석한 것임

- 0, 4, 8번 클러스터는 70% 이상의 과제가 DTP에 해당함
- 0, 4번 클러스터는 대다수의 과제가 제외된 것에 반하여, 8번 클러스터는 DTP 과제를 제외하더라도 331개의 과제가 남아있었음
- 이외 클러스터들은 DTP 과제의 비중이 낮음
- 특히, 암과 직접적 내용 연관성이 높은 2, 5, 6, 9번 클러스터는 DTP 과제의 비중이 1.1% 이하로 낮았음

〈표 4-30〉 UKRI 2015~2019년 'cancer' 검색 결과

클러스터	전체 검색결과 과제 수	DTP 과제를 제외한 과제 수	DTP 과제의 비중(%)
0	1,138	165	85.5%
1	524	455	13.2%
2	54	54	0.0%
3	222	218	1.8%
4	978	66	93.3%
5	116	116	0.0%
6	190	188	1.1%
7	466	446	4.3%
8	1,197	331	72.3%
9	115	115	0.0%
총합계	5,000	2,154	56.9%

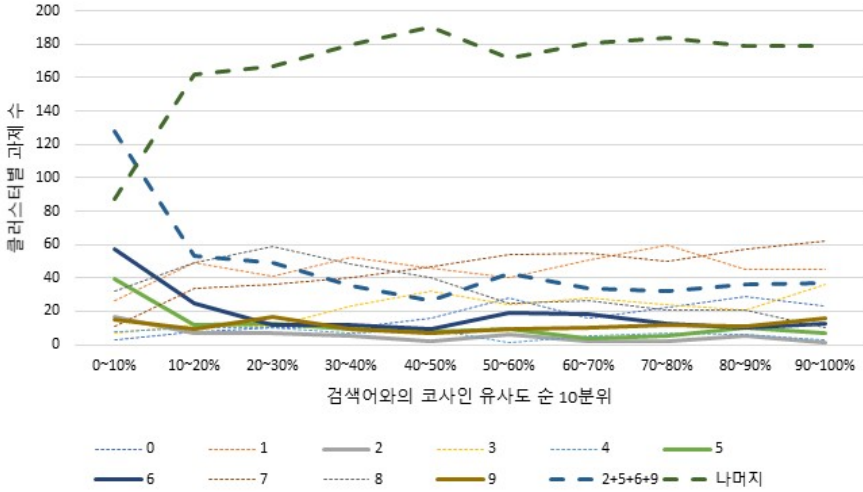
❍ 각 클러스터의 주제를 주요 키워드 및 과제명을 기반으로 추정한 결과, 2, 5, 6, 9번을 제외한 클러스터들은 대부분 주제를 특정할 수 없었음

〈표 4-31〉 UKRI 2015~2019년 'cancer' 5,000개 과제 클러스터(10개)별 주제

클러스터	주제
0	통일된 주제 없음
1	통일된 주제 없음
2	**암진단/치료**
3	물리학 등
4	통일된 주제 없음
5	**암치료제/기초기전연구**
6	**유전체/치료전략**
7	통일된 주제 없음
8	통일된 주제 없음
9	**기초기전연구**

❍ 검색어와의 유사도를 기준으로 검색 결과 과제들을 정렬하여 상위 10%씩 10분위를 나누어 과제 수 분포를 산출한 결과, 암과 관련성이 높은 클러스터와 나머지 클러스터의 분포는 상반되었음

- 상위 10% 과제 중 2, 5, 6, 9번 클러스터의 비중이 높은 반면, 하위 90% 과제는 나머지의 비중이 훨씬 높아짐



[그림 4-24] UKRI 2015~2019년 'cancer' 관련 과제 검색결과 (DTP 과제 제외)의 유사도 10분위별 분포

❍ UKRI는 큰 비중을 차지하는 포닥 장학금(DTP) 과제들의 요약문에서 연구 내용이 드러나지 않는 문제로 인해 이러한 과제들을 모두 제거하지 않으면 학습 자체가 제한되어 분석이 어려운 상태임

- 차후 데이터 전처리에 반영하여 개선이 필요함

제3절 국내외 동향 비교

□ 마이크로바이옴 분야 NIH 연구개발 투자동향 분석 결과와 국내를 비교하기 위하여 旣개발된 지능형 R&D정보데이터 분석시스템을 이용하여 조사분석 과제를 대상으로 마이크로바이옴 과제 현황을 분석함

○ 2015~2019년을 대상으로 '국내/해외 클러스터링 분석' 기능을 이용하여 '마이크로바이옴' 관련 과제를 검색함

○ 2019년 과제를 대상으로 1,000개의 과제를 검색하고 10개의 클러스터로 분류한 결과, 마이크로바이옴과 관련성이 낮은 클러스터가 다량 관찰되어 검색 결과 수를 줄일 필요가 있었음

〈표 4-32〉 조사분석 2019년 '마이크로바이옴' 1,000개 과제 클러스터(10개)별 주요 키워드

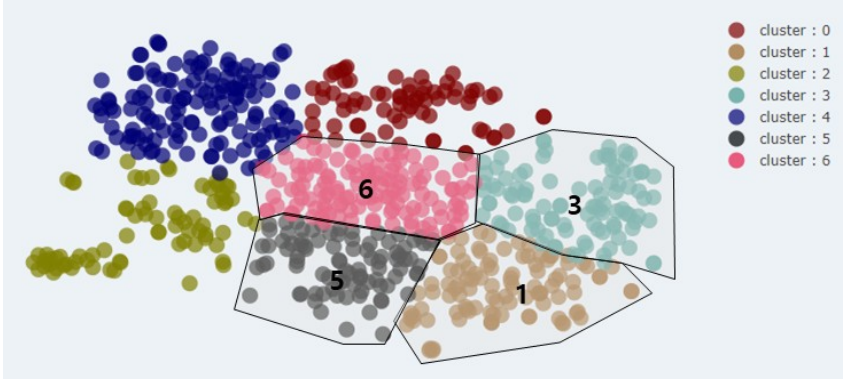
클러스터	주요 키워드
0	치료, 질환, 줄기세포, 세포, 금속, 식품, 펩타이드, 단백질, 나노, 탄소, 기전, 진단, 기술, 운동, 기능, 손상, 모델, 시스템, 소재, 물질
1	소재, 영상, 바이러스, 세포, 치료, 시스템, 조절, 기술, 제어, 발현, 유전자, 조직, 평가, 전이, 단백질, 기전, 생체, 나노, 환자, 대사
2	미생물, 유전체, 유전자, 식품, 면역, 소재, 유용, 김치, 발효, 질환, 기능, 치료, 반응, 바이오, 스마트, 제품, 프로바이오틱스, 개선, 물질, 시험
3	줄기세포, 세포, 나노, 센서, 치료, 유전자, DNA, 반응, 소재, 기술, 질환, 약물, 발현, 환자, 구조, 물질, 이드, 위암, 가노, 진단
4	세포, 소재, 공간, 기술, 유전체, 치료, 유방암, 유전자, 제어, 조절, 알고리즘, 센서, 구조, 로봇, 시스템, 식물, 항체, 데이터, 나노, 에너지
5	엽록체, 미생물, 유전자, **바이옴,** 유전체, 스트레스, 식물, 작물, 기능, 단백질, 세포, 마이크로, 조사, 기술, 물질, 형질전환, 고추, 영상, 광합성, 조절
6	자궁, 내막, 유전체, 조산, 진단, 유전자, 지능, 환자, 말라리아, 물질, 단백질, 세포, 플랫폼, 기능, 기법, 쇼그렌, 동물, 전립선암, AtR, 자극
7	영상, 나노, 치료, 질환, 관절염, 감염, 진단, 시스템, 천연물, 기술, 단백질, 면역, 분석, 탄소, 합성, 구조, 설계, 환자, 피부, 모델
8	유전자, 소자, 면역, 세포, 치료, 단백질, 줄기세포, 플라보노이드, 바이오, 발현, 산화, 조절, 질환, 표적, 구조, 바이러스, 데이터, 시스템, 약물, 환자
9	미생물, 질환, 세포, 진단, **바이옴,** 장내, 바이오, 면역, 공생, 분화, 마이크로, 감염, 염증, 나노, 치료, 호흡기, 유전체, 줄기세포, 분석, 소재

❍ 따라서 검색 결과 수를 700개로 줄이고, 7개의 클러스터로 분류한 결과, 4개의 클러스터(1, 3, 5, 6번)가 마이크로바이옴에 관련된 것으로 드러남

- 6번 클러스터의 경우 키워드 상에서는 마이크로바이옴 관련성이 드러나지 않지만(표 4-33), 포함된 과제들을 보면 마이크로바이옴과 연관성이 높음
- 해당 클러스터들은 중앙부부터 우측 하단 1/4 정도를 차지하고 있으며 (그림 4-25), 클러스터별 주제는 표 4-34와 같음

〈표 4-33〉 조사분석 2019년 '마이크로바이옴' 700개 과제 클러스터(7개)별 주요 키워드

클러스터	주요 키워드
0	나노, 줄기세포, 치료, 진단, 세포, 소재, 바이러스, 면역, 유전체, 구조, 분자, 금속, 생체, 센서, 이온, 식품, 환자, 반응, 로봇, 토양
1	**미생물,** 식품, **유전체,** 유용, 유전자, 질환, **프로바이오틱스,** 기능, 소재, 바이오, 면역, 김치, 전통, 개선, 다중, 치료, 발효, 시험, 물질, 균주
2	유전자, 세포, 공간, 기술, 발현, 유전체, 모델, 개발, 환자, 줄기세포, 조산, 구조, 손상, 약물, 표적, 질환, 플랫폼, 기능, 단백질, 비디오
3	유전자, **바이옴,** 식물, 미생물, 단백질, **엽록체, 작물,** 유전체, 스트레스, **마이크로,** 기능, **약물,** 반응, 세포, 물질, 고추, 질환, 광합성, 기술, 특성
4	세포, 나노입자, 감염, 영상, 면역, 소재, 시스템, 치료, 질환, 조절, 단백질, 나노, 제어, 네트워크, 평가, 발현, 분석, 유전자, 기전, 종양
5	질환, 유전자, **미생물, 장내,** 바이오, **치료,** 세포, **바이옴,** 염증, 데이터, 바이러스, 분화, 면역, 진단, **줄기세포,** 소재, 유전체, 센서, 소자, 유체
6	세포, 환자, 구조, 시스템, 유전자, 질환, 알고리즘, 치료, 유방암, 기술, 제어, 센서, 발현, 환경, 모델, 건강, 간암, 년도, the, 조절



[그림 4-25] 조사분석 2019년 '마이크로바이옴' 700개 과제 7개 클러스터 분류 결과

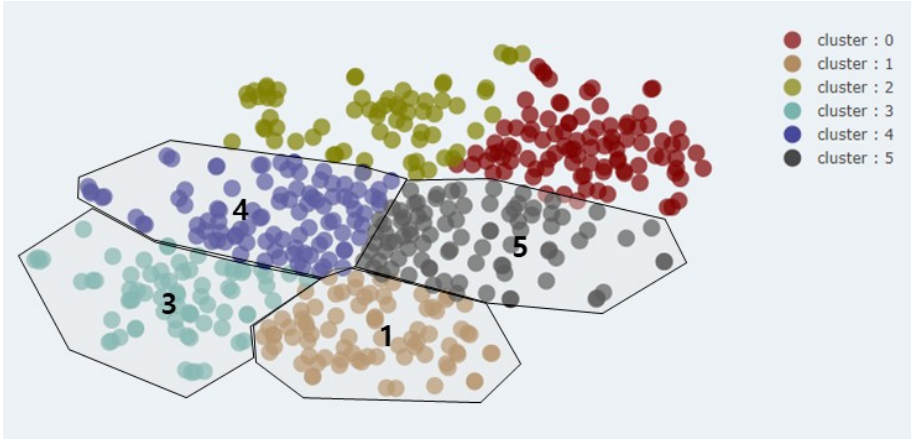
〈표 4-34〉 조사분석 2019년 '마이크로바이옴' 700개 과제 클러스터(7개)별 주제

클러스터	주요 키워드
0	마이크로RNA, 질병기전 등(미해당)
1	가축 마이크로바이옴, 건강증진용 프로바이오틱스 등
2	마이크로RNA, 질병기전 등(미해당)
3	작물 마이크로바이옴(농업 분야)
4	질병 기초기전, 면역치료 등(미해당)
5	마이크로바이옴 이용 치료제 개발, 마이크로바이옴 관련 질병 기전 규명
6	마이크로바이옴 이용 치료제 개발, 마이크로바이옴 관련 질병 기전 규명

❍ 2015~2018년 과제의 경우 보다 적은 결과가 나올 것임에 따라 검색 결과 수를 점차 줄여가면서 분석하였으며, 그 결과는 다음과 같음

〈표 4-35〉 조사분석 2018년 '마이크로바이옴' 500개 과제 클러스터(6개)별 주요 키워드, 주제

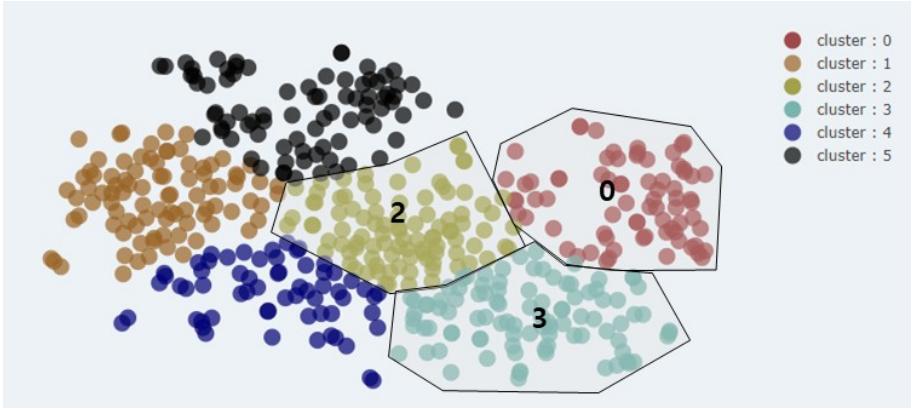
클러스터	주요 키워드	주제
0	단백질, 세포, 나노, 도시, 구조, 물질, 줄기세포, 조직, 치료, 조절, 소재, 메타, 제조, 환경, 프로그램, 모델, 스트레스, 기능, 차원, 평가	질병 기전연구, 면역치료(미해당)
1	바이옴, 마이크로, 미생물, 유전자, 작물, 기능, 식물, 유전체, 스트레스, 식품, 홀로, 프로바이오틱스, 세포, 조절, 효모, 치료, 평가, 당뇨, 바이오, 약물	동식물 마이크로바이옴(농림수산 분야), 김치 등 식품 프로바이오틱스 등
2	유전자, 치료, 나노, 소자, 진단, 불임, 줄기세포, 기술, 운동, DNA, 해석, 정자, 시스템, 공정, CRISPR, IoT, 신경세포, 천연물, 후보, 환자	줄기세포, 질병기전 연구(미해당)
3	프로바이오틱스, 김치, 미생물, 바이오, 장내, 소재, 질환, 기능, 균주, 식품, 배양, 완화, 전통, 공정, 파마, 염증, 주의, 틱스, 유전체, 마이크로	건강증진용 마이크로바이옴 식품(프로바이오틱스) 등
4	파마, 틱스, 바이오, 프로바이오틱스, 미생물, 세포, 물질, 비만, 마이크로, 바이옴, 질환, 장내, 효능, 신경, 모델, 유전자, 면역, 진화, 치료, 염증	마이크로바이옴 이용 치료제 개발, 마이크로바이옴 관련 질병 기전 규명
5	유전자, 세포, 치료, 소자, 염증, 데이터, 물질, 소재, 공정, 억제, 표적, 생산, 년도, 단백질, 발굴, 기능, 줄기세포, 기억, 촉매, 구조	마이크로바이옴 이용 치료제 개발, 마이크로바이옴 관련 질병 기전 규명



[그림 4-26] 조사분석 2018년 '마이크로바이옴' 500개 과제 6개 클러스터 분류 결과

〈표 4-36〉 조사분석 2017년 '마이크로바이옴' 500개 과제 클러스터(6개)별 주요 키워드, 주제

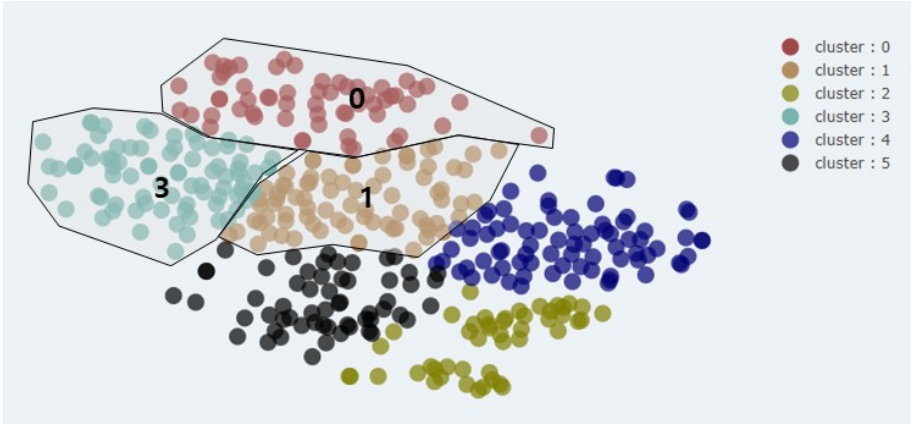
클러스터	주요 키워드	주제
0	단백질, 유전체, 유전자, 미생물, 김치, 작물, 형질전환, 선발, 식물, 유산균, 발현, 평가, 조절, 성분, 기능, 시스템, 식품, 품종, 유용, GMO	동식물 마이크로바이옴(농림수산 분야)
1	세포, 노화, 단백질, 나노, 전이, 물질, 유전자, 치료, 질환, 산화, 전립선암, 조절, 발현, 시스템, 차원, 종양, 촉매, RNA, 구조, 유전체	면역 관련 질환 기전 연구(미해당)
2	세포, 노화, 줄기세포, 약물, 나노, 데이터, 질환, 분석, 환자, 기전, 물질, 시스템, 발굴, 유전자, 치료, 조절, 전이, 전사체, 기반, 안전	마이크로바이옴 이용 치료제 개발, 마이크로바이옴 관련 질병 기전 규명
3	소재, 미생물, 식품, 유전자, 제품, 물질, 천연물, 세포, 개선, 기능, 치료, 조절, 진단, 할랄, 시험, 건강, 유전체, 장내, 항체, 효능	마이크로바이옴 이용 치료제 개발, 마이크로바이옴 관련 질병 기전 규명
4	대장암, 유전자, 세포, 환자, 질환, 시스템, 파킨슨병, 나노, 모델, 진단, 치료, 바이오, 장내, 손상, 중성미자, 대사, 운동, 물질, 톡소플라즈마, 분석	유전자 기반 진단 등(미해당)
5	유전자, 세포, 줄기세포, 나노, 시스템, 치료, 유전체, 데이터, 에너지, 변형, 기술, 망막, 년도, 환자, 조절, 주행, 질환, 물질, 재생, 센서	유전자 기반 진단, 치료 등(미해당)



[그림 4-27] 조사분석 2017년 '마이크로바이옴' 500개 과제 6개 클러스터 분류 결과

〈표 4-37〉 조사분석 2016년 '마이크로바이옴' 400개 과제 클러스터(6개)별 주요 키워드, 주제

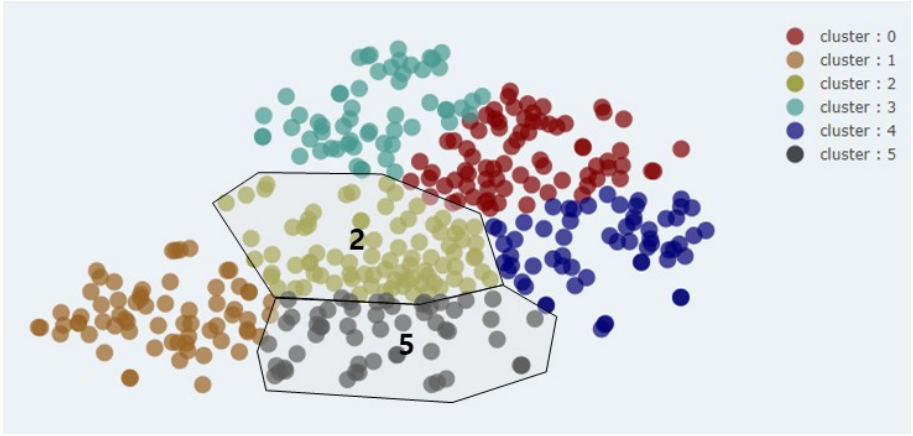
클러스터	주요 키워드	주제
0	GM, 유전체, 유전자, 정보, 미생물, 엽록체, 작물, 분자, 기능, 세포, 조절, 분석, 평가, 형질전환, 육종, 식품, 시스템, 농업, 분화, 유용	작물 마이크로바이옴(농업 분야)
1	유전체, 유전자, 물질, 면역, 치료, 세포, 태양전지, 소프트웨어, 진단, 종양, 데이터, 분석, 모델, 환경, 바이러스 질환, 조산, 영상, 표적, 기능	마이크로바이옴 이용 치료제 개발, 마이크로바이옴 관련 질병 기전 규명
2	바이오, 세포, 프린팅, 줄기세포, 미토콘드리아, 유전체, 재생, 질환, 장애, 소재, 부인, 유전자, 림프관, 제어, 음성, TSM, 다중, Frizzled, 치료, 클라우드	유전자 기반 진단, 치료 등(미해당)
3	전분, 물질, 질환, 원료, 진단, 치료, 세포, 유전체, 유전자, 소재, 평가, 김치, 기능, 시험, 데이터, 작물, 분석, 시스템, 대사체, 독성	동식물 마이크로바이옴(농림수산 분야), 김치 등 식품 프로바이오틱스 등
4	유전자, 단백질, 세포, 질병, 펩타이드, 구조, 치료, 조절, 줄기세포, 양자, 분화, 보존, 추출, 염증, 질환, 전달, 나노, 면역, 항균, 연골	면역 관련 질환 기전, 바이오마커 연구 (미해당)
5	교육, 세포, 약물, 정보, 유전자, 조직, 프로그램, 조절, 기술, 생물, 유전체, 환자, 나노, 자살, 신경, 기전, 러닝, 치료, 항암제, 면역	마이크로RNA, 질병기전 등(미해당)



[그림 4-28] 조사분석 2016년 '마이크로바이옴' 400개 과제 6개 클러스터 분류 결과

〈표 4-38〉 조사분석 2015년 '마이크로바이옴' 400개 과제 클러스터(6개)별 주요 키워드, 주제

클러스터	주요 키워드	주제
0	보안, 과제, 환자, 줄기세포, 크론병, 세포, 질환, 치료, 에너지, 시스템, 면역, 표적, 알고리즘, 단백질, 나노, 항암제, 바이오, 진단, 물질, DNA	면역질환 치료, 마이크로RNA, 유전체연구 등 (미해당)
1	유전체, 유전자, 식중독, 곤충, 식물, 물질, 조절, 조사, 정보, 단백질, 식품, 세포, 작물, 형질전환, 생산, 미생물, 저항, 분자, 추출, 질병	동식물 유전체 연구 기반 육종, 질병 치료(미해당)
2	보안, 질환, 세포, 유전체, 과제, 유전자, 마이크로, 에너지, 혈관, 식품, 물질, 생산, 바이오, 데이터, 소재, 모델, 시스템, 기술, 구조, 한국인	마이크로바이옴 이용 치료제 개발, 마이크로바이옴 관련 질병 기전 규명
3	보안, 분화, 줄기세포, 질환, 과제, 후성, 유전체, 세포, 치료, 데이터, 유전자, 나노입자, 기술, 소재, 기전, 발굴, 항암제, miRNA, 터치, 은하	마이크로RNA, 후성유전체 연구 등 (미해당)
4	세포, 유전자, 치료, 단백질, 진단, 백신, 면역, 바이러스, microRNA, 발현, 천식, 물질, 아토피피부염, 환자, 식물, 점막, 약제, 반응, 지표, 발굴	아토피피부염 등 면역질환 치료 (미해당)
5	보안, 토양, 미생물, 유전체, 유전자, 세포, 영상, 농업, 소재, 정보, 서비스, 지역, 수확, 과제, 물질, 사료, 사업, 은하, 줄기세포, 바이오	동물, 인간 마이크로바이옴 이용 치료제, 질병기전

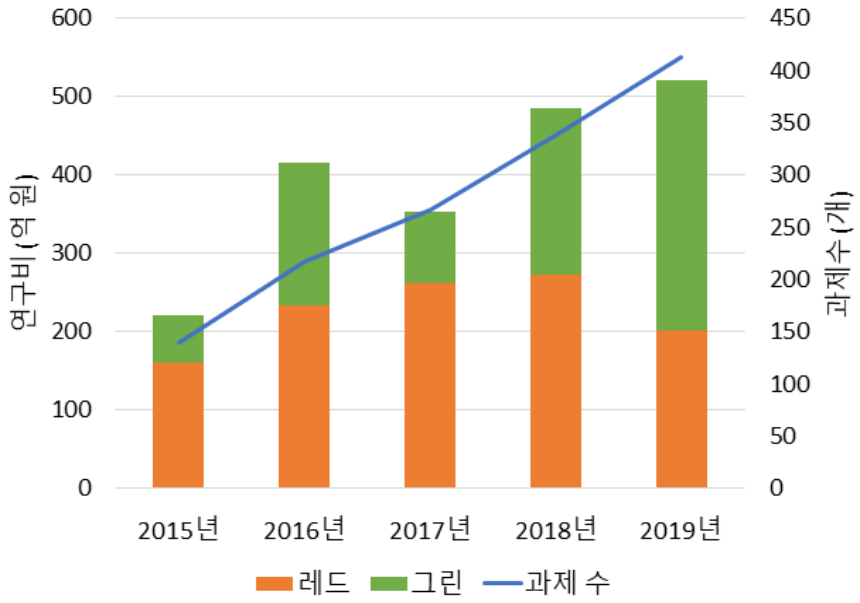


[그림 4-29] 조사분석 2015년 '마이크로바이옴' 400개 과제 6개 클러스터 분류 결과

❍ 연도별 분석 결과, 전체적인 투자 규모는 2015년 220억원에서 2019년 521억원 규모로 증가하였으나, 레드바이오/그린바이오로 분리해서 보면 레드바이오 분야의 투자는 정체 상태에 있음

- 그린바이오 분야의 투자는 동식물 질병 저항성, 수확량 증가를 위해 마이크로바이옴 기반 작물보호제 또는 토양 미생물 균총과의 상호작용 등을 연구하는 과제들이 증가하고, 마이크로바이옴 기반 건강기능 보조식품 등을 개발하는 과제들이 지속적으로 늘어나는 추세임
- 반면에 레드바이오 분야는, 마이크로바이옴을 이용한 치료제 기반 또는 질환-장내미생물 상호작용 연구 등 수년 전 연구 내용이 크게 변화없이 투자되고 있는 것으로 보임
- 동 분석에서 제시된 투자 규모는 기존의 연구에서 발표된 투자금액*과 어느정도 일치하여 정확성이 담보되는 것으로 판단됨

  * '16년 휴먼 마이크로바이옴 R&D 투자 : 242억원(KISTEP, 2018)



[그림 4-30] 조사분석 2015~2019년 '마이크로바이옴' 분야 R&D 투자 추이

□ 본 분석결과를 미국 NIH 마이크로바이옴 분야 투자현황 분석 결과와 비교하면 다음과 같은 시사점을 도출할 수 있음

○ 휴먼 마이크로바이옴 분야의 투자액은 미국 대비 절대적으로 큰 차이가 존재하지만, 바이오헬스 분야에 전반적인 투자 규모의 차이를 고려하면 마이크로바이옴 분야에 국내에서 더 집중적인 투자가 이루어짐을 알 수 있음

- '19년 기준 NIH는 3.1억 달러(약 3,400억원)를 투자한 반면, 국내의 투자 규모는 201억원(NIH의 5.9%)임
- NIH 전체 R&D 예산 382억 달러(약 42.0조원)와 국내 바이오헬스 R&D 예산 규모(1.0조원)를 감안할 경우 마이크로바이옴 분야의 투자 비중은 우리나라가 1.9%로 NIH(0.8%)보다 높음

○ NIH의 휴먼 마이크로바이옴 분야 투자가 '19년 들어 급증한 것과 반대로, 국내는 휴먼 마이크로바이옴 분야 투자가 정체 상태에 있음

❍ 국내는 절대적 투자액은 적으나 투자 비중이 높아 마이크로바이옴 분야의 예산 투자 확대는 어려울 가능성이 존재하므로, 향후 현황 진단을 통한 투자 효율화가 필요하다고 볼 수 있음

제4절 소결 및 한계점

□ 동 연구의 doc2vec 텍스트임베딩 알고리즘에 기반한 해외 정부R&D 과제 분석은 어느정도 신뢰할 수 있는 통계치를 제공하는 것으로 확인되었으나, 세부적인 클러스터(군집)를 확인하고 주제들을 확인하는데는 조사분석 데이터 기반으로 旣개발되어 있는 지능형 R&D정보데이터 분석시스템에 비하여 한계점을 노출하였음

❍ 마이크로바이옴, 생물학, 육종, 암 등 바이오헬스 관련 다양한 검색어에 대하여 국내외 클러스터링 비교분석을 수행한 결과, 원하는 검색어에 대해 기존에 발표된 규모와 유사한 개수 및 연구비에 해당하는 과제를 검색할 수 있었음

❍ 클러스터별 주요 키워드를 검토한 결과 검색 결과 내에서 세부분야 간의 연도별 투자 추이 등을 관찰할 수 있었으며, 국내외 결과 비교를 통해 시사점을 도출할 수 있었음

- 검색어와 실제 관련성이 높은 클러스터들은 관련성이 낮은 클러스터들과 위치상으로 분류되어 시각화하였을 때 뭉쳐있는 경향성을 뚜렷하게 보여주었으며, 새롭게 떠오르는 분야의 경우 과거에 비하여 최근에 과제들이 증가하고 있는 것이 명확히 관찰됨

❍ 따라서 동 연구에서 사용한 방법론은 향후 국내외 연구개발 동향분석에 활용될 수 있는 잠재력이 매우 높으며, 아래 기술할 일부 기술적 한계점들을 개선하면 보다 적극적으로 실무에 적용 가능할 것으로 사료됨

❍ 분석 결과가 영어로 되어있음에 따른 해석 능력의 한계가 있었으며, 과학기술 표준분류와 같은 임베딩 성능을 향상시킬 수 있는 지도학습 태그값 없이 순수히 텍스트만 학습시켰기 때문에 임베딩 및 클러스터링 성능 저하가 나타난 것으로 사료됨

- 기존 지능형 분석시스템에서는 조사분석 데이터의 '과학기술표준분류-중1'(가장 가중치가 높은 분류)을 지도학습 태그 정보로 활용하여 임베딩 성능을 향상시키고 있음
- 과학기술표준분류는 상호 배타성이 있는 분류체계로 텍스트 학습 시 매우 효과적인 성능 향상을 제공하는 것으로 판단됨

○ 원시 데이터의 문제점도 존재했는데, NSF, UKRI의 경우 연구분야가 특정되지 않는 인건비 지원 또는 기관 지원 성격의 과제들이 다량 포함되어 있어 텍스트 학습 시 오차를 일으키는 원인으로 작용했을 것임

□ 다음 이슈들은 본 연구 및 분석시스템에서 취하고 있는 과제분류 접근법의 근본적인 이슈들임

○ 200~300차원의 doc2vec 임베딩 벡터를 그대로 시각화할 수 없어 2차원으로 차원 축소하는 t-SNE 알고리즘을 사용하고 있는데, t-SNE 알고리즘은 perplexity와 같은 파라미터에 매우 민감하게 반응할 수 있는 알고리즘으로, 보통 강건하다고 알려져 있지만[7] 이용자가 필요시 파라미터를 조정할 필요성이 있어 시스템 사용의 난이도가 높아지는 문제가 있음
- 이러한 파라미터들은 일정한 규칙이 있는 것이 아니라 여러번 시행착오를 거쳐 원하는 결과가 나오도록 조정해야 하는데, 이것이 비효율적임

○ 또한 2차원으로 차원이 축소되어 정보가 상당히 손실된 상태에서 클러스터링을 하는 것은 클러스터링 품질을 떨어트리는 문제로 작용하고 있음

○ 비지도학습 텍스트마이닝은 텍스트 그 자체에 크게 의존하기 때문에 서로 다른 부처의 과제들을 섞어서 분석할 경우 요약문 등의 작성 형식이 달라 동등하게 비교가 불가능해 부처별로 따로 분석해야 하는 한계점 존재
- 여러 부처 과제를 한번에 분석할 경우 부처별로 유사도가 전반적으로 차이가 나게 되어 하나의 부처에 검색 결과가 편중되게 됨

7) How to Use t-SNE Effectively(https://distill.pub/2016/misread-tsne/)

- 이러한 문제는 같은 부처라도 서로 다른 연도에 대해서 발생할 수 있는데, 연도가 달라지면서 작성 형식 또는 스타일이 달라질 수 있기 때문임

❍ 또한 문서 클러스터링의 가장 근본적인 문제는 정답지가 없는 비지도학습이기 때문에 결과의 적절성 · 적합성을 판단하기가 어려움

- 즉, 학습 및 클러스터링이 잘 되었는지 정량적으로 판단할 수 있는 도구가 없고 분석자 또는 사용자가 정성적으로 판단해야 하는 문제가 있음

❍ 여러 기술분야의 융합, 기술의 복잡성으로 인하여 단편적으로 과제가 분류되는 것이 아니라 연속성을 가지게 되므로 클러스터 간의 구분이 불명확하게 됨

- doc2vec 등의 알고리즘으로 과제들을 임베딩 할 경우, 마치 과제들을 공간상에 흩뿌려 지도를 그리는 것과 유사하다고 볼 수 있는데, 현재 기술적 문제로 인해 해상도가 낮아 큰 분야는 키워드를 보고 파악이 가능하지만 세부적인 분야가 어떠한 비율로 군집을 이루고 있는지는 정확히 파악하기가 어려운 상태라고 볼 수 있음

□ 현재 이러한 이슈들을 해결하기 위해서는 분석자의 분야 지식(domain knowledge)이 높거나, 지도학습이 가능한 태그정보가 과제에 부여되어 있거나, 또는 과학기술 분야에 고도로 최적화된 언어모델이 요구됨

❍ 언어모델 고도화의 경우, 최근 일반적인 언어에 대해서 학습되어 있는 모델을 원하는 분야에 특화하여 미세조정(fine tuning)하는 전이학습(transfer learning)을 말하는데, 해당 분야의 전문 단어들이 포함된 말뭉치를 다량 학습시켜 언어모델 자체의 성능을 높이는 것임

❍ 이에 대해서는 결론에서 보다 상세히 후술하겠음

□ 동 연구의 분석 결과를 바이오 분야 전문가 면담을 통하여 공유하고 의견을 수렴한 결과, 다음과 같은 개선점이 필요하다는 제안이 있었음

❍ 영문 분석의 경우 주요 키워드 출력 시 일반적으로 자주 등장하는 단어(study, research 등)의 불용어 처리가 현재 미흡하여 클러스터 간의 차이를 보기 어려운 상태임

❍ 분석 결과의 정확성을 평가하기 위해서는 세부사업, 연구개발단계, 과학기술표준분류 등의 다른 과제 정보를 연동하여 피벗테이블 분석을 수행하면 도움이 될 수 있음

❍ 현재 마이크로바이옴 분석 결과를 검토한 결과, 기초기전 연구와 치료제 개발, 치료전략 개발 등의 서로 상이한 연구개발단계(분야)가 잘 분류되지 않고 여러 클러스터에 혼합되어 있는 상태임

- 단순히 단어의 배열만을 보고 학습하고 있기 때문에, 같은 마이크로바이옴 분야이나 연구개발단계 등이 다른 경우 잘 구별되지 않고 있어, 조사분석 데이터의 다른 항목을 학습에 이용하거나, 여타 전문 지식을 활용하는 등 학습의 고도화가 필요해보임

❍ 현재 검색 결과의 수를 사용자가 직접 입력하도록 되어 있는데, 이는 큰 불편점으로 향후 자동적으로 검색 결과의 수를 설정해주는 방법을 고안할 필요

제5장 신약개발 정부R&D 투자포트폴리오 분석

제1절 신약개발 투자포트폴리오 분류기준

제2절 2019년도 신약개발 R&D 투자포트폴리오 분석

제1절 신약개발 투자포트폴리오 분류기준[8)]

□ 생명보건의료분야 예산심의대상 정부연구개발사업 중 신약개발분야에 해당하는 사업('19년 기준 45개)을 중심으로, 국가과학기술정보서비스(NTIS)에서 제공하는 국가연구개발사업 조사분석데이터('19)의 과제정보를 활용

○ 신약개발분야 전문가를 통해 신약개발 목적의 과제를 선별, 신약개발단계, 의약품 종류, 타겟 질환 등의 분류기준*에 따라 과제 분류(표 5-1)

* 「신약개발 R&D 투자 효율화 방안(2012)」에서 제안된 분류기준으로, 생명의료전문위 등 신약분야 관련 전문가 의견을 반영하여 수립

○ 기초 및 기전연구(수행대상: 개별 연구자) 과제는 최종 목표 설정 전 수행되는 과제로 간주, 제외하고 분석을 수행함

〈표 5-1〉 신약개발분야 정부 R&D 투자포트폴리오 분류기준

구분	대분류	중분류	소분류
신약개발 단계	기초기전연구	기초기전연구	기초기전연구
	타겟발굴및검증	타겟발굴및검증	타겟발굴및검증
	후보물질도출 및 최적화	후보물질도출 및 최적화	후보물질도출 및 최적화
	비임상	비임상	비임상
	임상	임상1상	임상1상
		임상2상	임상2상
		임상3상	임상3상
	인프라	신약플랫폼기술	타겟발굴 플랫폼
			후보물질 발굴 플랫폼
			전임상 플랫폼
			질환동물 플랫폼
			임상 플랫폼
		인력양성	인력양성
		제도·정책	제도·정책
		인·허가	인·허가
	기타	기타	기타
의약품 종류	**신약**	**합성신약**	합성신약
		바이오신약	단백질 치료제
			펩타이드 치료제

8) 신약분야 전문가를 통해 구축한 19년도 정부 신약개발 R&D과제 DB 결과 분석

구분	대분류	중분류	소분류
			유전자 치료제
			세포 치료제
			백신
			항체기반신약
			기타
		한약(생약제제)	한약(생약제제)
	개량신약	개량신약(합성)	개량신약
		바이오베터	단백질 치료제
			펩타이드 치료제
			유전자 치료제
			세포 치료제
			백신
			항체기반신약
			기타
	바이오시밀러	바이오시밀러	바이오시밀러
	공통기반기술 및 기타	공통기반기술	공통기반기술
		기타	기타
질환	혈관질환, 호흡기질환, 종양질환(혈액암포함), 근골격계질환, 면역계질환, 감염증, 정신질환, 퇴행성뇌질환, 내분비질환, 비만, 그 외 희귀질환, 기타		

제2절 2019년도 신약개발 R&D 투자포트폴리오 분석

가. 신약개발 분야 정부 R&D 투자 현황

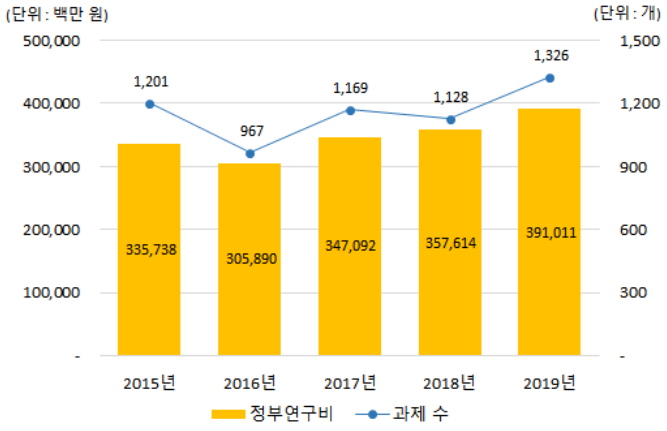
□ 신약개발분야 정부 R&D 투자 현황

○ 2019년 신약개발분야 정부 R&D 투자 규모는 3,910억원으로, 연구비 기준 5년간('15~'19) 연평균 약 3.9% 증가

- 신약개발과제 수는 '15년 1,201건에서 '19년 1,326건으로 연평균 약 2.5% 증가

〈표 5-2〉 신약개발분야 정부 R&D 투자 규모(2015~2019)

구분	2015년	2016년	2017년	2018년	2019년	연평균 증가율
과제 수 (개)	1,201	967	1,169	1,128	1,326	2.5%
정부연구비 (백만원)	335,738	305,890	347,092	357,614	391,011	3.9%



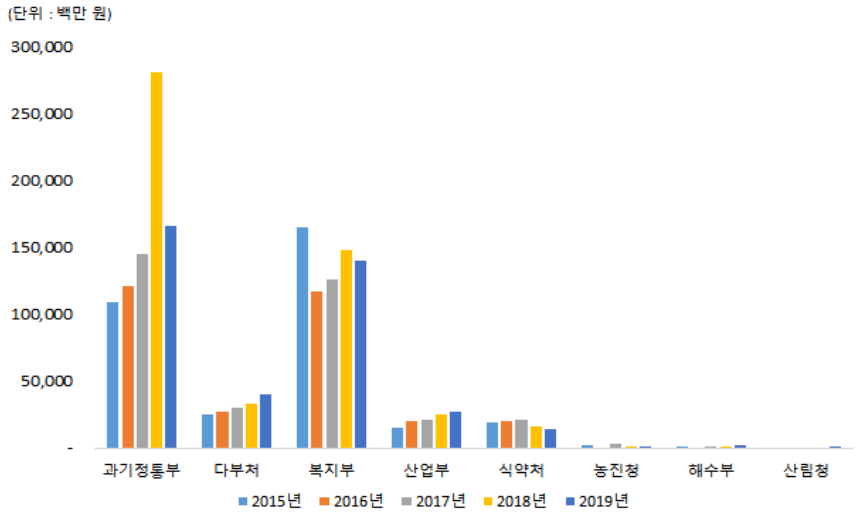
[그림 5-1] 신약개발분야 정부 R&D 투자 현황(2015~2019)

□ 부처별 투자 현황

❍ 2019년 신약개발 분야에 정부연구비 총액 3,910억원 중 과기정통부의 투자가 1,656억원으로 가장 큰 비중(42.3%)을 차지(연평균 11.0% 증가)

❍ 다음으로 복지부(1,395억원, 35.7%), 다부처(401억원, 10.2%) 순으로 신약개발과제 지원

- 3개 부처가 전체 예산의 88.3%를 차지



[그림 5-2] 신약개발분야 정부 R&D 부처별 투자 현황(2015~2019)

〈표 5-3〉 신약개발분야 정부 R&D 부처별 투자 현황(2015~2019)

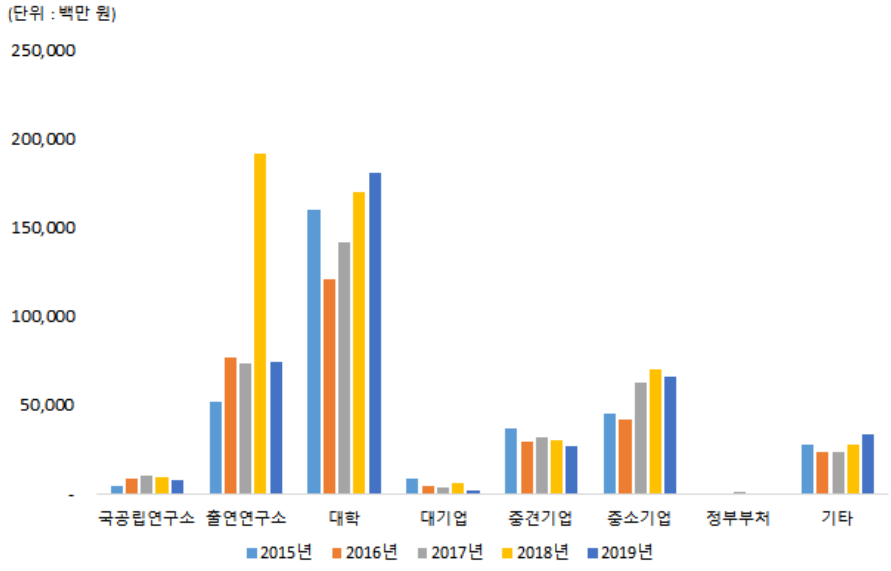
구분	2015년		2016년		2017년		2018년		2019년		연평균 증가율 (%)
	연구비 (백만원)	비중(%)	연구비 (백만원)	비중(%)	연구비 (백만원)	비중(%)	연구비 (백만원)	비중(%)	연구비 (백만원)	비중 (%)	
과기정통부	108,970	32.5	121,485	39.7	144,700	41.7	142,134	39.7	165,579	42.3	11.0
다부처	25,308	7.5	27,300	8.9	30,600	8.8	33,000	9.2	40,051	10.2	12.2
보건복지부	164,748	49.1	117,026	38.3	125,799	36.2	138,806	38.8	139,534	35.7	△4.1
산업부	14,835	4.4	20,062	6.6	21,235	6.1	25,428	7.1	26,668	6.8	15.8
식약처	19,345	5.8	20,017	6.5	20,896	6.0	15,981	4.5	14,572	3.7%	△6.8
농진청	2,150	0.6	-	-	2,782	0.8	930	0.3	1,010	0.3	△17.2
해수부	382	0.1	-	-	1,080	0.3	1,335	0.4	2,477	0.6	59.6
산림청	-	-	-	-	-	-	-	-	1,120	0.3	-
합계	335,738	100.0	305,891	100.0	347,092	100.0	357,614	100.0	391,011	100.0	70.4

□ 연구수행주체별 투자 현황

❍ 2019년 기준 대학에서 1,806억원(46.2%) 규모의 가장 높은 투자 비중을 보임

- 다음으로 출연연구소(742억원, 19.0%), 중소기업(660억원, 19.0%), 기타(337억원, 8.6%) 순으로 투자

❍ 대기업 및 중견기업의 신약개발분야 정부 R&D 투자는 축소된 반면(연평균 △33.7%, △7.5%), 국공립연구소는(연구비 기준) '15년 48억원에서 '19년 78억원으로(연평균 약 13.2%)로 가장 많이 증가



[그림 5-3] 신약개발분야 정부 R&D 연구수행주체별 투자 현황(2015~2019)

〈표 5-4〉 신약개발분야 정부 R&D 연구수행주체별 투자 현황(2015~2019)

구분	2015년		2016년		2017년		2018년		2019년		연평균 증가율 (%)
	연구비 (백만원)	비중 (%)	연구비 (백만원)	비중 (%)	연구비 (백만원)	비중 (%)	연구비 (백만원)	비중 (%)	연구비 (백만원)	비중 (%)	
국공립연구소	4,780	1.4	8,559	2.8	10,502	3.0	8,482	2.4	7,481	2.0	13.2
출연연구소	52,054	15.5	76,675	25.1	73,784	21.3	67,234	18.8	74,190	19.0	9.3
대학	160,062	47.7	121,253	39.6	141,791	40.9	148,678	41.6	180,588	46.2	3.1
대기업	8,892	2.6	4,345	1.4	3,400	1.0	5,831	1.6	1,722	0.4	△33.7
중견기업	36,750	10.9	29,424	9.6	31,757	9.1	29,929	8.4	26,887	6.9	△7.5
중소기업	45,034	13.4	42,155	13.8	62,440	18.0	69,803	19.5	66,036	16.9	10.0
정부부처	-	-	-	-	200	0.1	-	-	-	-	-
기타	28,167	8.4	23,480	7.7	23,218	6.7	27,656	7.7	33,746	8.6	4.6
합계	335,738	100.0	305,891	100.0	347,092	100.0	357,614	100.0	391,011	100.0	△1.0

□ 주요 대상사업별 투자 현황

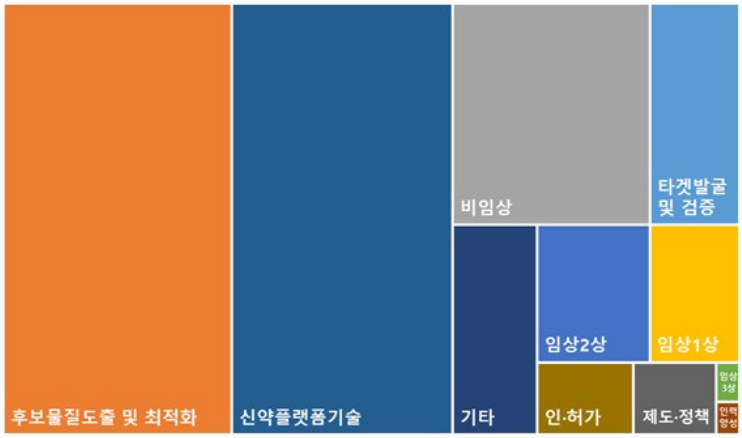
○ 신약개발분야를 지원하는 주요 사업으로는 과기정통부의 바이오·의료기술개발이 1,082억원(27.7%)을 지원하고 있었으며, 동 사업 내 신약개발분야 투자 비중은 40.5% 수준

○ 다음으로 복지부의 첨단의료기술개발(334억원, 8.6%), 범부처전주기신약개발(286억원, 7.3%), 산업부의 바이오산업핵심기술개발(221억원, 5.7%) 순으로 신약개발분야 주요 사업으로 나타남

〈표 5-5〉 신약개발분야 정부 R&D 주요 사업(2019)

사업명	총 사업비 (백만원) (A)	의약분야 투자액(백만원)		비중 (%) (B/A)
		(B)	비중 (%)	
100세사회대응고령친화제품연구개발	2,795	300	0.1	10.7
가속기기반신약개발지원	5,950	5,820	1.5	97.8
감염병관리기술개발연구	20,395	5,055	1.3	24.8
감염병위기대응기술개발	22,633	13,168	3.4	58.2
공익적질병극복연구지원사업	6,036	2,187	0.6	36.2
국가보건의료연구인프라구축	13,635	1,108	0.3	8.1
국가치매극복기술개발	11,541	2,937	0.8	25.5
국가항암신약개발사업	14,224	14,224	3.6	100.0
글로벌프론티어지원	75,301	5,927	1.5	7.9
뇌과학원천기술개발	51,591	3,886	1.0	7.5
만성병관리기술개발연구	11,222	200	0.1	1.8
바이오.의료기술개발	267,128	108,174	27.7	40.5
범부처전주기신약개발	28,647	28,647	7.3	100.0
바이오산업핵심기술개발	59,315	22,103	5.7	37.3
산림생명자원소재발굴연구	5,015	820	0.2	16.4
산림생물종연구	10,430	300	0.1	2.9
선도형특성화연구사업	5,170	3,042	0.8	58.8
스마트임상시험플랫폼기반구축사업	2,780	2,669	0.7	96.0

사업명	총 사업비 (백만원) (A)	의약분야 투자액(백만원)		비중 (%) (B/A)
		(B)	비중 (%)	
안전기술선진화	3,140	2,100	0.5	66.9
안전성평가기술개발연구	13,582	2,942	0.8	21.7
안전성평가연구소연구운영비지원	32,106	3,460	0.9	10.8
암연구소및국가암관리사업본부운영	58,000	7,019	1.8	12.1
양·한방융합기반기술개발	1,825	433	0.1	23.7
연구개발사업관리	2,561	183	0.0	7.2
연구자주도질병극복연구	14,496	5,439	1.4	37.5
연구중심병원육성	34,050	17,046	4.4	50.1
의약품등안전관리	21,056	9,347	2.4	44.4
인공지능신약개발플랫폼구축사업	7,500	7,500	1.9	100.0
임상연구인프라조성	23,198	14,657	3.7	63.2
질환극복기술개발	29,769	9,999	2.6	33.6
차세대바이오그린21	53,686	1,010	0.3	1.9
창의산업미래성장동력	10,306	3,565	0.9	34.6
첨단의료기술개발	49,246	33,447	8.6	67.9
첨단의료복합단지기반기술구축	3,329	1,723	0.4	51.8
첨단의료복합단지미래의료산업원스톱지원	5,604	3,340	0.9	59.6
포스트게놈산산업육성을위한다부처유전체사업	10,100	3,566	0.9	35.3
한국생명공학연구원연구운영비지원	89,065	14,387	3.7	16.2
한국한의학연구원연구운영비지원	48,447	5,655	1.4	11.7
한국화학연구원연구운영비지원	76,767	7,703	2.0	10.0
한의기반융합기술개발	3,565	1,965	0.5	55.1
한의약선도기술개발	9,945	2,917	0.7	29.3
해양바이오전략소재개발및상용화지원	4,980	238	0.4	4.8
해양수산생명공학기술개발	24,892	2,239	0.6	9.0
혁신신약파이프라인발굴	8,000	8,000	2.0	100.0
혁신형의사과학자공동연구사업	3,751	564	0.1	15.0
총합계	1,256,774	391,011	100.0	31.1



[그림 5-4] 신약개발분야 정부 R&D 신약개발단계별 투자 현황(2019)

나. 신약개발단계별 정부 R&D 투자 현황

□ 신약개발단계별 정부 R&D 투자 현황

○ 중분류 기준 타겟발굴 및 검증단계는 '15년 155억원에서 '19년 243억원으로 연평균 11.9%, 후보물질도출 및 최적화 단계는 연평균 30.0%('15) 426억원 → ('19) 1,216억원)으로 증가한 반면 임상3상 단계는 연평균 △37.7%('15) 74억원 → ('19) 11억원)로 감소

○ '19년 대분류 기준 인프라(1,341억원, 34.3%), 후보물질 도출 및 최적화(1,216억원, 31.1%), 타겟발굴 및 검증(243억원, 6.2%) 순으로 투자가 이루어지고 있음

- 정부 투자는 신약개발단계 후반부로 갈수록 감소하나 인프라에 대한 투자는 확대

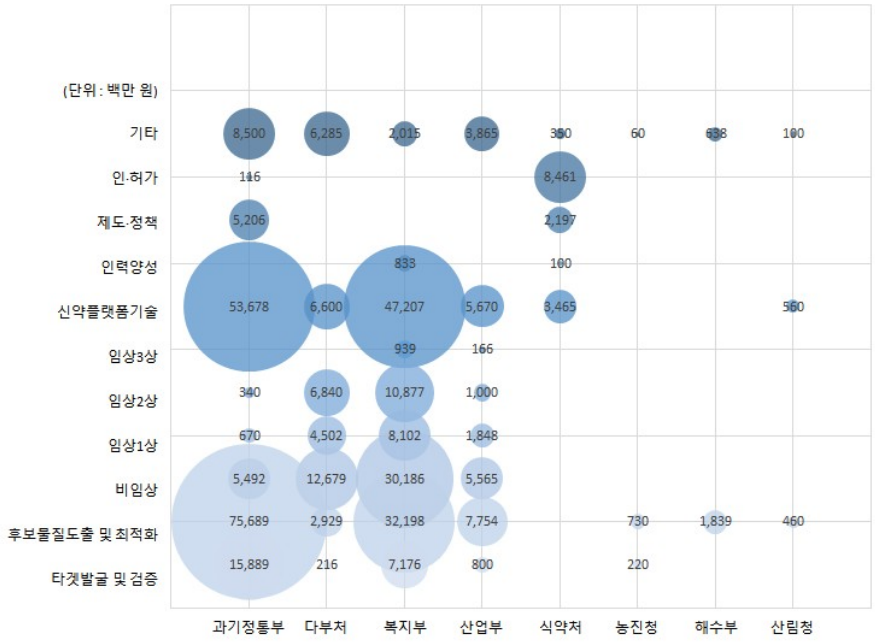
〈표 5-6〉 신약개발분야 정부 R&D 신약개발단계별 투자 현황(2015~2019)

구분			2015년 ·		2016년 ·		2017년		2018년		2019년		연평균 증가율 (%)
			연구비 (백만원)	비중 (%)	연구비 (백만원)	비중 (%)	연구비 (백만원)	비중 (%)	연구비 (백만원)	비중 (%)	연구비 (백만원)	비중 (%)	
타겟발굴 및 검증	타겟발굴 및 검증		15,519	4.6	37,268	12.2	38,419	11.1	31,078	8.7	24,301	6.2	11.9
후보물질도출 및 최적화	후보물질도출 및 최적화		42,563	12.7	67,366	22.0	113,837	32.8	119,445	33.4	121,599	31.1	30.0
비임상	비임상		63,828	19.0	29,678	9.7	34,708	10.0	51,151	14.3	53,922	13.8	△4.1
임상	임상1상		27,122	8.1	16,417	5.4	14,820	4.3	23,529	6.6	15,122	3.9	△13.6
	임상2상		16,474	4.9	19,150	6.3	28,423	8.2	24,781	6.9	19,058	4.9	3.7
	임상3상		7,356	2.2	7,223	2.4	5,535	1.6	822	0.2	1,105	0.3	△37.7
인프라	신약플랫폼기술	타겟발굴 플랫폼	11942	3.6	7,845	2.6	8,349	2.4	7,201	2.0	6,880	1.8	△12.9
		후보물질 발굴 플랫폼	28271	8.4	26,952	8.8	34,199	9.9	27,628	7.7	55,664	14.2	18.5
		비임상 플랫폼	42978	12.8	24,767	8.1	8,221	2.4	15,698	4.4	24,463	6.3	△13.1
		질환동물 플랫폼	6485	1.9	17,870	5.8	16,718	4.8	15,652	4.4	15,484	4.0	24.3
		임상 플랫폼	18735	5.6	9,632	3.1	6,895	2.0	10,665	3.0	14,689	3.8	△5.9
	인력양성		1,150	0.3	-	-	60	0.0	668	0.2	933	0.2	△5.1
	제도·정책		6,308	1.9	4,426	1.4	7,557	2.2	5,276	1.5	7,402	1.9	4.1
	인·허가		14,407	4.3	21,107	6.9	13,904	4.0	10,264	2.9	8,577	2.2	△12.2
기타	기타		32,600	9.7	16,190	5.3	15,447	4.5	13,757	3.8	21,813	5.6	△9.6
합계			335,738	100.0	305,891	100.0	347,092	100.0	357,614	100.0	391,011	100.0	3.9

□ 부처별 단계별 투자 현황

❍ '19년 신약개발분야에 가장 많이 지원한 과기정통부는 임상 이전 단계에 투자를 집중하였으며, 복지부 역시 후보물질 도출 및 최적화 단계(322억원)에 투자를 주력하였음

- 과기정통부는 후보물질 도출 및 최적화(757억원), 신약플랫폼기술(537억원), 타겟발굴 및 검증(159억원) 순으로 투자
- 복지부는 신약플랫폼기술(472억원), 후보물질도출 및 최적화 단계(322억원), 비임상(302억원) 순으로 투자
- 다부처 및 산업부의 경우 임상단계에의 투자 비중이 타 부처에 비해 높게 나타남



[그림 5-5] 신약개발분야 정부 R&D 부처별 단계별 투자 현황(2019)

〈표 5-7〉 신약개발분야 정부 R&D 부처별 단계별 투자 현황(2019)

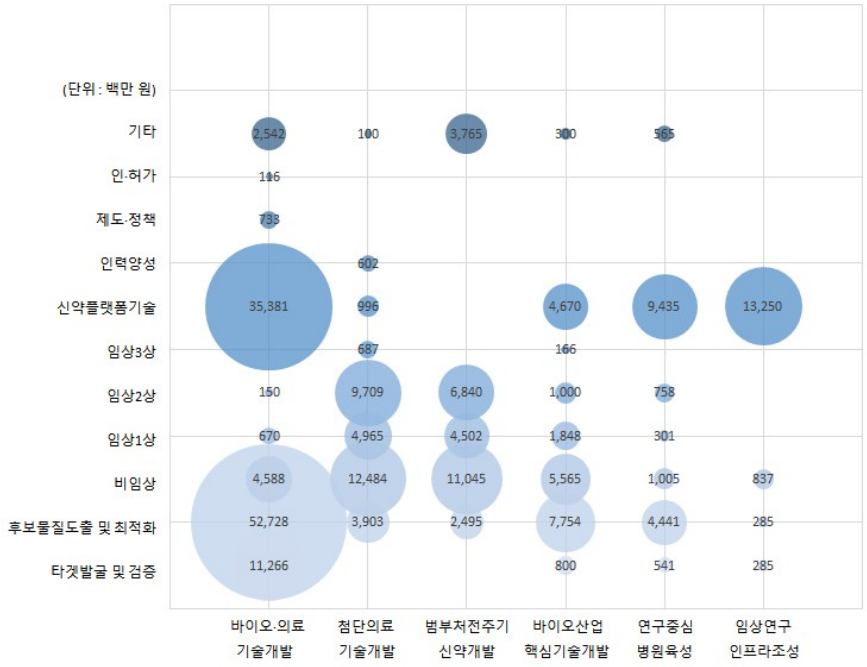
(단위: 백만원)

구분		과기정통부	다부처	복지부	산업부	식약처	농진청	해수부	산림청	합계
타겟발굴 및 검증		15,889	216	7,176	800	-	220	-	-	24,301
후보물질도출 및최적화		75,689	2,929	32,198	7,754	-	-	-	-	121,599
비임상		5,492	12,679	30,186	5,565	-	-	-	-	53,922
임상1상		670	4,502	8,102	1,848	-	-	-	-	15,122
임상2상		340	6,840	10,877	1,000	-	-	-	-	19,058
임상3상		-	-	939	166	-	-	-	-	1,105
신약플랫폼기술	타겟발굴 플랫폼	3,118	2,050	1,712	-	-	-	-	-	6,880
	후보물질 발굴 플랫폼	29,156	2,975	20,382	2,300	290	-	-	56	55,664
	비임상 플랫폼	5,397	450	17,116	850	650	-	-	-	24,463
	질환동물 플랫폼	13,292	-	2,052	-	140	-	-	-	15,484
	임상 플랫폼	2,715	1,125	5,944	2,520	2,385	-	-	-	14,689
인력양성		-	-	833	-	100	-	-	-	933
제도·정책		5,206	-	-	-	2,197	-	-	-	7,402
인·허가		116	-	-	-	8,461	-	-	-	8,577
기타		8,500	6,285	2,015	3,865	350	60	638	100	21,813
합계		165,579	40,051	139,534	26,668	14,572	1,010	2,477	1,120	397,011

□ 주요사업별 단계별 투자 현황

❍ 신약개발분야 주요 사업 중 가장 비중이 큰 과기정통부의 바이오·의료기술개발사업은 후보물질도출 및 최적화(527억원)에 가장 많이 투자하였으며, 후보물질 발굴 플랫폼(204억원), 타겟발굴 및 검증(113억원) 순으로 지원

- 복지부의 첨단의료기술개발은 비임상 및 임상단계를 중점적으로 투자하였으며, 비임상(125억원), 임상2상(97억원), 임상1상(50억원) 순
- 범부처전주기신약개발은 비임상(235억원)에 가장 많이 투자하고 있으며, 다음으로 임상2상(165억원), 임상1상(95억원), 후보물질도출 및 최적화(64억원) 순으로 지원



[그림 5-6] 신약개발분야 정부 R&D 주요사업별 단계별 투자 현황(2019)

〈표 5-8〉 신약개발분야 정부 R&D 주요사업별 단계별 투자 현황(2019)

(단위: 백만원)

구분		바이오_의료 기술개발	첨단의료기술개발	범부처 전주기신약개발	바이오 산업핵심기술개발	연구중심병원육성	임상연구 인프라조성
타겟발굴 및 검증		11,266	-	-	800	541	285
후보물질도출 및 최적화		52,728	3,903	2,495	7,754	4,441	285
비임상		4,588	12,484	11,045	5,565	1,005	837
임상1상		670	4,965	4,502	1,848	301	-
임상2상		150	9,709	6,840	1,000	758	-
임상3상		-	687	-	166	-	-
신약 플랫폼 기술	타겟발굴 플랫폼	1,864	229	-	-	1,133	-
	후보물질 발굴 플랫폼	20,431	105	-	2,300	5,036	100
	비임상 플랫폼	1,181	275	-	85	1,941	12,317
	질환동물 플랫폼	10,597	-	-	-	633	-
	임상 플랫폼	1,308	387	-	1,520	693	833
인력양성		-	602	-	-	-	-
제도·정책		733	-	-	-	-	-
인·허가		116	-	-	-	-	-
기타		2,542	100	3,765	300	565	-
합계		108,174	33,447	28,647	22,103	17,046	14,657

다. 의약품 종류별 정부 R&D 투자 현황

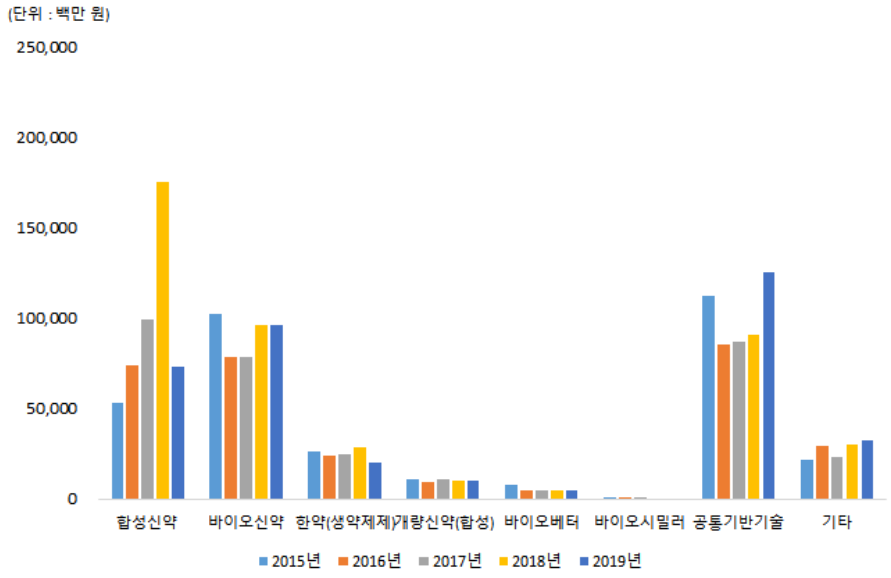
□ 의약품 종류별 투자 현황

○ 합성신약에 대한 투자는 연평균 약 8.1%, 바이오신약에 대한 투자는 연평균 약 5.4% 수준으로 증가

- 바이오신약 중 세포치료제에 대한 투자는 '15년 130억원에서 '19년 510억원으로 가장 많이 확대됨(연평균 약 40.7%)

○ '19년 기준 투자 규모의 절반 이상이 신약개발(56.1%)*에 집중되어 있으며, 바이오신약(1,261억원, 32.3%), 공통기반기술(1,255억원, 32.1%), 합성신약(729억원, 18.6%) 순으로 투자

* 개량신약 포함 시 59.6%



[그림 5-7] 신약개발분야 정부 R&D 의약품 종류별 투자 현황(2015~2019)

〈표 5-9〉 신약개발분야 정부 R&D 의약품 종류별 투자 현황(2019)

(단위: 백만원)

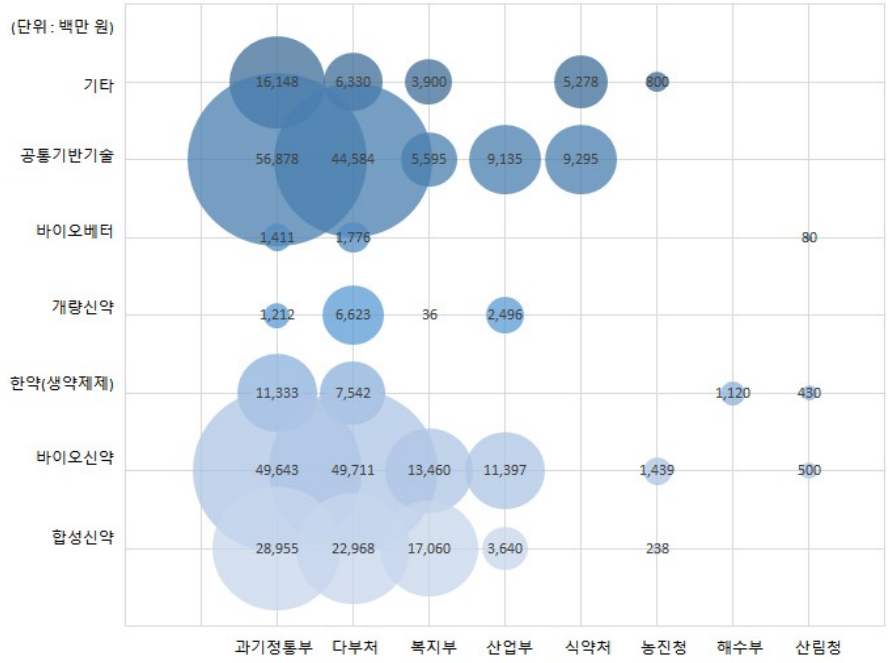
구분		2015년		2016년		2017년		2018년		2019년		연평균증가율 (%)
		연구비	비중(%)	연구비	비중(%)	연구비	비중(%)	연구비	비중(%)	연구비	비중(%)	
합성신약		53,426	15.9	73,623	24.1	99,204	28.6	96,459	27.0	72,861	18.6	8.1
바이오 신약	단백질 치료제	18,180	5.4	14,384	4.7	18,888	5.4	14,073	3.9	11,585	3.0	△10.7
	펩타이드치료제	-	-	-	-	-	-	6,993	2.0	7,662	2.0	2.3
	유전자치료제	11,426	3.4	7,632	2.5	10,451	3.0	8,111	2.3	8,078	2.1	△8.3
	세포치료제	45,793	13.6	17,313	5.7	19,714	5.7	32,454	9.1	36,668	9.4	△5.4
	백신	13,049	3.9	25,878	8.5	24,171	7.0	23,088	6.5	17,519	4.5	7.6
	항체기반신약	13,791	4.1	13,568	4.4	22,930	6.6	26,715	7.5	32,348	8.3	23.8
	기타	-	-	-	-	-	-	10,735	3.0	12,291	3.1	△1.9
한약(생약제제)		26,486	7.9	24,201	7.9	24,942	7.2	22,049	6.2	20,425	5.2	△6.3
개량신약(합성)		11,005	3.3	8,955	2.9	10,756	3.1	10,065	2.8	10,366	2.7	△1.5
바이오 베터	단백질 치료제	2,050	0.6	1,573	0.5	3,008	0.9	1,671	0.5	1,613	0.4	△5.8
	유전자 치료제	250	0.1	-	-	-	-	325	0.1	244	0.1	△0.6
	세포 치료제	130	0.0	-	-	-	-	480	0.1	510	0.1	40.7
	백신	3,883	1.2	1,523	0.5	575	0.2	623	0.2	-	0.0	△100.0
	항체기반신약	1,265	0.4	1,714	0.6	890	0.3	270	0.1	-	0.0	△100.0
	기타	-	-	-	-	-	-	600	0.2	900	0.2	10.7
바이오시밀러		1,080	0.3	1,200	0.4	1,200	0.3	-	-	-	0.0	△100.0
공통기반기술		112,607	33.5	85,199	27.9	86,924	25.0	82,670	23.1	125,486	32.1	△2.7
기타		21,317	6.3	29,126	9.5	23,441	6.8	20,233	5.7	32,455	8.3	11.1
총 합계		335,738	100.0	305,891	100.0	347,092	100.0	357,614	100.0	391,011	100.0	3.9

□ 부처별 의약품종류별 투자 현황

❍ 과기정통부는 공통기반기술(569억원), 바이오신약(496억원), 합성신약(290억원) 순으로 투자

❍ 복지부는 바이오신약(497억원)의 투자가 높고 공통기반기술(446억원), 합성신약(230억원) 순으로 지원

❍ 다부처는 합성신약(171억원), 산업부는 바이오신약(114억원), 식약처는 공통기반기술(93억원)에 가장 많이 투자하는 것으로 나타남



[그림 5-8] 신약개발분야 정부 R&D 부처별 의약품종류별 투자 현황(2019)

〈표 5-10〉 신약개발분야 정부 R&D 부처별 의약품종류별 투자 현황(2019)

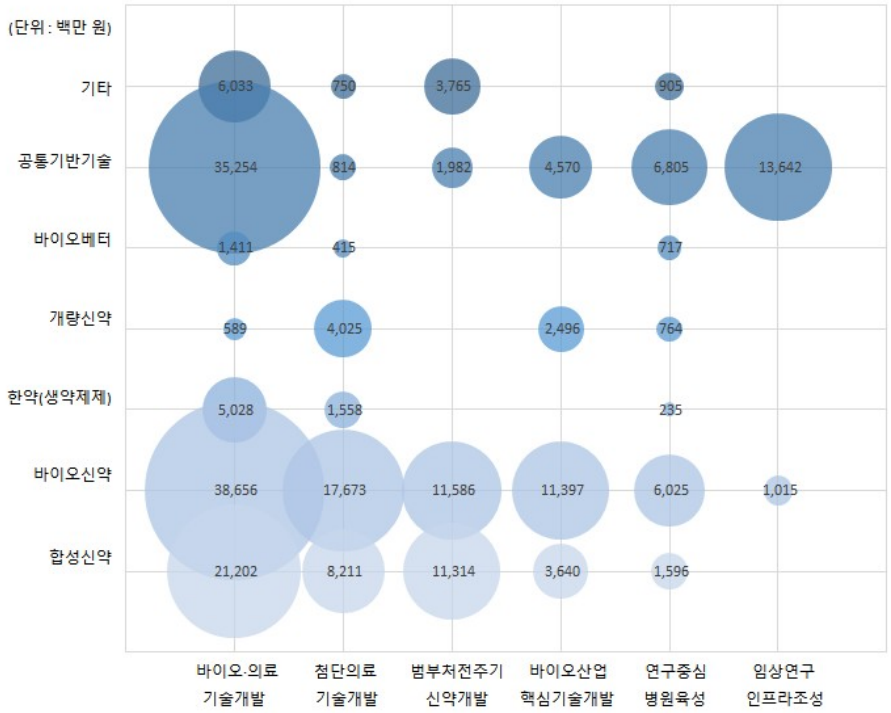
(단위: 백만원)

구분		과기정통부	다부처	복지부	산업부	식약처	해수부	산림청	농진청	합계
합성신약		28,955	22,968	17,060	3,640	-	238	-	-	72,861
바이오신약	단백질 치료제	3.752	2,513	831	2,900	-	1,439	-	150	11,585
	펩타이드 치료제	3.190	2,088	814	1,450	-	-	-	120	7,662
	유전자 치료제	4.444	1,926	1,707	-	-	-	-	-	8,078
	세포 치료제	14.089	17,766	2,140	2,673	-	-	-	-	36,668
	백신	1,597	15,003	-	749	-	-	-	170	17,519
	항체기반 신약	15,315	7,173	7,725	2,075	-	-	-	60	32,348
	기타	7,257	3,242	243	1,550	-	-	-	-	12,291
한약 (생약제제)		11,333	7,542	-	-	-	-	1,120	430	20,425
개량신약 (합성)		1,212	6,623	36	2,496	-	-	-	-	10,366
바이오베터	단백질 치료제	736	797	-	-	-	-	-	80	1,613
	유전자 치료제	-	244	-	-	-	-	-	-	244
	세포 치료제	200	310	-	-	-	-	-	-	510
	백신	-	-	-	-	-	-	-	-	-
	항체기반 신약	-	-	-	-	-	-	-	-	-
	기타	475	425	-	-	-	-	-	-	900
공통기반기술		56,878	44,584	5,595	9,135	9,295	-	-	-	125,486
기타		16,148	6,330	3,900	-	5,278	800	-	-	32,455
총 합계		165,579	139,534	40,051	26,668	14,572	2,477	1,120	1,010	391,011

□ 주요사업별 의약품종류별 투자 현황

○ 전반적으로 전년도와 같이 합성신약, 바이오신약, 천연물신약 등의 신약 개발분야 투자 비중이 큰 양상을 보임

- 과기정통부의 바이오·의료기술개발은 바이오신약(387억원)이 가장 높은 투자 비중을 보였으며, 복지부의 첨단의료기술개발은 바이오신약(177억원)의 투자가 높았음



[그림 5-9] 신약개발분야 정부 R&D 주요사업별 의약품종류별 투자 현황(2019)

〈표 5-11〉 신약개발분야 정부 R&D 주요사업별 의약품종류별 투자 현황(2019)

(단위 : 백만원)

구분		바이오·의료 기술개발	첨단의료 기술개발	범부처전주기신약개발	바이오산업 핵심 기술개발	연구중심 병원육성	임상연구인프라조성
합성신약		21,202	8,211	11,314	3,640	1,596	-
바이오신약	단백질 치료제	3,677	775	831	2,900	788	-
	펩타이드 치료제	3,190	357	814	1,450	-	-
	유전자 치료제	3,844	-	1,707	-	-	990
	세포 치료제	12,629	9,114	2,094	2,673	3,712	25
	백신	1,597	1,100	-	749	939	-
	항체기반 신약	7,013	3,853	6,139	2,075	587	-
	기타	6,707	2,474	-	1,550	-	-
한약 (생약제제)		5,028	1,558	-	-	235	-
개량신약 (합성)		589	4,025	-	2,496	764	-
바이오베터	단백질 치료제	736	-	-	-	717	-
	유전자 치료제	-	-	-	-	-	-
	세포 치료제	200	140	-	-	-	-
	백신	-	-	-	-	-	-
	항체기반 신약	-	-	-	-	-	-
	기타	475	275	-	-	-	-
공통기반기술		35,254	814	1,982	4,570	6,805	13,642
기타		6,033	750	3,765	-	905	-
총 합계		108,174	33,447	28,647	22,103	17,046	14,657

제6장 결 론

제1절 연구결과 요약

제2절 향후 발전 방안

제1절 연구결과 요약

□ 동 연구과제를 통하여 지능형 R&D정보데이터 분석시스템을 K2Base와 연계하여 온라인화하였으며, 분석 알고리즘을 해외 정부R&D 데이터에 적용하여 국내외 R&D 현황을 비교분석하는 방법론을 검증하였음

❍ 분석시스템의 온라인화 작업은 K2Base 용역개발 과제와 연계하여 추진하였으며, 추후 KISTEP 직원들이 사용할 수 있도록 K2Base 메뉴에 포함될 예정임

- 기관 내 서버 확충 및 자원 할당이 현재 진행 중인 상태로 아직 정식 서비스 시작은 되지 않았으며, '21년 상반기 중으로 예정
- 분석시스템 온라인화 작업 과정에서 미세한 버그들 역시 수정되었으며, 사용 편의성 제고를 위한 그래픽 인터페이스의 개선 작업이 병행되었음

❍ 국내외 R&D 현황 비교를 위하여 미국 NIH의 2015~2019년 데이터를 웹크롤링 기법으로 수집하였으며, 분석시스템에 탑재된 알고리즘을 적용하여 클러스터링 분석을 수행한 결과, 어느정도 신뢰할만한 연구비 투자규모가 산출되는 것이 확인되었으며 국내외 비교분석을 통해 정책적 시사점을 도출할 수 있다고 보여짐

- '마이크로바이옴(microbiome)' 분야에 대해 NIH와 국내 조사분석 과제의 비교분석을 수행하였으며, 검색 결과 기존에 보고된 투자규모와 유사한 수준의 연구비 규모가 산출되었음
- 국내외 현황 차이로는, 국내 투자규모는 NIH에 비하여 매우 적은 수준이지만, 바이오헬스 분야 전체 투자규모를 고려하면 마이크로바이옴 분야에 대한 투자집중도는 높아 국내 관련 예산 확대 가능성은 높지 않다고 보여짐
- 다만, 국내 마이크로바이옴 분야 투자는 인간 마이크로바이옴뿐만 아니라 농업 분야의 투자가 같이 집계되었는데, NIH의 연구내용에 대응되는 인간 마이크로바이옴 분야 투자는 최근 5년간 정체 상태에 있었음

❍ 미국 NSF와 영국 UKRI의 2015~2019년 데이터를 각 기관의 인터넷 사이트에서 내려받아 클러스터링 분석을 적용해본 결과, 현재 자연어처리 알고리즘 수준은 매우 세부적인 주제의 탐지보다는 거시적인 그룹핑에 더 적합한 것으로 드러났으며, 분야 전문지식(domain knowledge)의 적용 및 3차원 클러스터링 분석 등의 방법론 고도화를 통해 분석 품질을 제고해야 할 필요성이 제기됨

- NSF 데이터를 대상으로 육종(breeding) 분야에 대해서 검색한 결과, 육종 분야에 해당하는 과제들을 검출할 수 있었으나, 동물/식물 육종 또는 육종 테크닉 등 세부적인 분야·기술에 따른 분리는 일어나지 않아 미시적 수준에서의 클러스터링은 아직 미흡한 것으로 나타남
- NSF 데이터를 대상으로 보다 포괄적인 생물학(biology) 분야에 대해서 검색한 결과, 과제들이 복잡한 상호 연관관계를 가지고 있음에 따라 2차원 차원축소 상태에서는 클러스터들이 서로 겹치는 현상이 많이 발생했으나, 연도별 경향에서 최근 다학제적 연구의 비중이 늘어나고 있는 현상은 포착이 가능하였음
- UKRI 데이터를 대상으로 암(cancer)에 대해서 분석을 시도한 결과, 유사도 상위 10%에 실제 내용적으로 암에 관련된 과제들이 모여있는 현상이 뚜렷하였으나, NSF 데이터를 대상으로 생물학에 대해서 분석한 것과 유사하게 클러스터 간의 경계가 명확하지 않았음
- NSF, UKRI 데이터에서 공통적으로 관찰된 현상은, 각 기관의 특성상 큰 비중을 차지하고 있는 분야 불특정 일반 장학금(우리나라의 BK21 등과 유사) 과제들이 학습에 노이즈로 작용하여 분석 결과에 악영향을 미친다는 점임

※ 해당 과제들은 요약문(abstract)에 세부 연구내용이 아닌 프로그램 자체의 소개 내용만이 기재되어 있어 서로 구분이 잘 되지 않으며, 검색 결과에서 큰 부분을 차지하나 의미는 별로 없음

※ NSF와 UKRI 과제 분석의 품질을 제고하기 위해서는 해당 장학금 과제들은 일괄 제거하고 재학습할 필요가 있음

제2절 향후 발전 방안

□ 현재의 한계점을 지금의 기술로 해결하기에는 복잡하며, 결과의 적절성을 사람이 전문지식에 의존하여 판단해야 한다는 점에 큰 어려움이 있으나, 분야 전문가의 전문지식, 과제 데이터베이스의 다른 정보(연구개발단계, 수행주체 등)의 활용, 차원축소 · 시각화 방법론의 개선 등을 통해 분석 품질을 제고할 수 있을 것이라고 보여짐

○ 이는 인공지능 전반의 문제점으로 지적되고 있는 사항으로, 인공신경망을 이용한 단순 텍스트 분석만으로는 결과의 도출 과정을 이해할 수 없어 원하지 않는 결과가 나왔을 때 해결책을 도출하기 어려움

○ 동 시스템이 보다 폭넓게 실무에 활용될 수 있기 위해서는, 보다 정밀한 클러스터링을 통해 세부적인 분야를 나누어줄 수 있어야 하고, 거시적인 분석 시 클러스터를 보다 효과적으로 시각화할 수 있어야 함

- 클러스터링의 정밀도 향상을 위해서는 데이터 학습 시 과제 원시데이터에서 얻을 수 있는 다른 정보들을 활용하고, 전문가들의 전문 지식을 학습에 반영할 수 있는 방안을 강구할 필요

  ※ 예시로 전년도 연구에서 과학기술표준분류 또는 PubMed의 MeSH heading을 학습 지표 중 하나로 활용했던 것이 있음

- 클러스터 시각화의 개선을 위해서는 현재 2차원으로 축소하는 시각화를 3차원으로 높인다거나 하여 정보의 손실을 줄이는 방법 등이 있음

○ 또한, 현재 사용자가 직접 입력해주어야 하는 파라미터들(검색 결과 수, 클러스터 수 등)을 자동으로 설정해주는 방법론에 대한 개발이 필요함

- 분석시스템의 사용 확산을 위해서는 간단한 절차로도 활용도가 높은 결과값을 제공할 수 있어야 함

□ 최신 언어지능 기술을 이용하여 '지능형 연구개발 정보 데이터 분석 시스템'을 개선시킬 수 있는 세부적 방안은 다음과 같음

❍ 지능형 연구개발 정보 데이터 분석 시스템의 입력은 사용자의 관심 기술분야에 대한 키워드 또는 문장 형태라고 가정함

❍ 출력은 다음의 단계를 따른다고 가정함

- (1단계) 입력에 대해 관련성 높은 과제(혹은 문서)를 제시
- (2단계) 클러스터링 기술을 이용하여 1단계 결과물을 군집화
  ※ 군집별로 기존 시스템의 입력으로 이용하여 과제 수, 연구비 합계 등 통계 정보를 수집
- (3단계) 군집화 결과를 이용하여 각 군집별 주요 keyword 등을 추출함

❍ 지능형 연구개발 정보 데이터 분석 시스템의 분석 대상 분야가 넓기 때문에 먼저 고려해야 할 점은 여러 분야로 쉽게 이식이 가능하도록 도메인 적응(Domain Adaptation) 능력을 고려해야 함

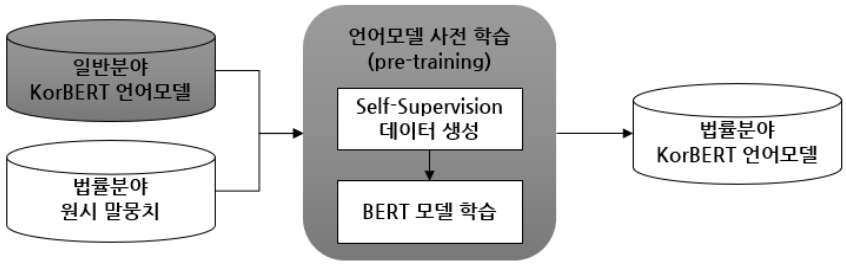
- 클러스터링 기술의 경우 비지도 학습(unsupervised learning)의 특성 때문에 학습 데이터가 필요하지는 않지만 딥러닝 알고리즘을 적용하고 양질의 군집 결과를 얻기 위해서는 딥러닝 언어 모델 활용이 필요함
- BERT와 같은 딥러닝 언어 모델을 활용할 경우 한국어 일반적인 지식을 보유한 상태로 각 특정 분야의 연구개발 정보를 상대적으로 용이하게 분석할 수 있다는 장점이 있음
- 그러나, 한국어를 대상으로 KorBERT, HanBERT, KoBERT 등 일반적인 텍스트 문서를 대상으로 구축한 한국어 딥러닝 언어 모델을 공개되어 있지만 연구개발 정보 데이터를 분석할 수 있는 수준의 한국어 딥러닝 언어 모델은 현재 존재하지 않음

❍ 엑소브레인[9] 과제에서는 상대적으로 많은 양의 텍스트 문서를 수집하기 어려운 법률분야 딥러닝 언어 모델을 구축하기 위해서 소량의 법률 분야 원시 말뭉치(약 186MB)를 활용하여 법률 분야에 특화된 언어 모델을 구축함

- 일반 분야에서 구축되어 성능이 입증된 KorBERT를 기반으로 법률분야 언어 모델을 추가로 구축하였으며(그림 6-1) 이와 같은 방법은 상대적으로

9) http://exobrain.kr/

소량의 텍스트 문서만 존재하는 응용 분야별 특화된 언어 모델 적용 가능성을 제시함



[그림 6-1] 일반 분야 KorBERT 기반 법률분야 특화 KorBERT 구축 개념도

- 비슷한 사례로 유사 특허 검색을 위해 특허 분야에 특화된 언어 모델을 개발하여 유사 특허 분류 태스크에 적용한 결과 딥러닝 언어 모델을 쓰지 않은 고전적인 자연어처리 방법에 비해 6.25%, 일반분야 KorBERT 모델을 사용한 방법에 비해 10.9% 성능이 향상됨
- 지능형 연구개발 정보 데이터 분석 시스템에서 다루는 연구 분야 별로 수집할 수 있는 텍스트 문서 양 등을 고려하여 사전 학습 언어 모델을 구축하고 이를 기반으로 군집화 및 이에 기반한 키워드 추출 태스크를 수행할 경우 현재 방법보다는 개선된 결과를 제시할 수 있을 것으로 예상됨
- 어떠한 접근 방법을 쓰더라도 가장 중요한 것은 한국어 문장 및 문서를 분석하는 가장 기본적인 단위인 형태소 분석에 있어서 적용하려는 분야에 대한 용어 사전 등을 보강하여 정확한 한국어 분석에 기반하여 모든 자연어 처리가 수행되어야 함.
- 구글 등에서 BPE를 기반으로 WPM, SPM 등 토크나이징 방법을 제시하였으나, 이는 다국어 분석을 전제로 하기 때문에 한국어에 특화된 분석 방법이 필요하며, 특화된 분석을 이용할 경우에 좀더 나은 성능을 기대할 수 있음

  ※ BPE 등의 방법은 언어 기본 단위를 단어가 아닌 문자(character, 영어의 경우 알파벳)으로 간주하여 처리하는 방법인데, 한 단어에 쓰인 모든 알파벳이 같은 중요도를 갖는 언어에는 범용적으로 쓰일 수 있지만 교착어인 한국어는 경우 '한국어는'이라는 단어를 한/국/어/는 으로 분리하여 동일하게 처리하면 형식 형태소인 '는'이

실질형태소의 일부인 '한', '국' 등과 같이 학습되어 오류를 일으키는 여지를 남김. 한국어는 보통 명사, 조사와 같은 태그 정보를 추가하여 학습함.

※ 언어에 특화된 토크나이징 방법은 단지 딥러닝 언어모델의 성능 뿐 아니라 이를 이용한 특정 태스크의 성능에도 영향을 미침

❍ 개발 초기 단계인 설명가능한 AI 개념을 도입하는 것도 지능형 연구개발 정보 데이터 분석 시스템의 활용성을 높이는데 기여할 것으로 예상됨

- 현재는 XAI 연구 초기 단계이기 때문에 바로 적용하는 것은 힘들겠지만, 궁극적으로는 정보 분석 시스템의 결과를 사용자가 신뢰할 수 있도록 하기 위해서는 시스템의 결과를 사람이 이해할 수 있도록 제시를 해야 함
- 각 분야별 연구 개발 정보 데이터를 분석자가 모두 이해할 수 없기 때문에 암스테르담 대학의 연구 결과처럼 판단 근거를 하이라이팅 해 주는 정도로도 분석자에게는 도움이 될 수 있음
- 분석시스템 사용자가 분석 시스템의 결과를 확신할 수 없을 경우 XAI 알고리즘을 도입하여 분석 시스템의 결과의 옳고 그름 여부에서부터 분석 시스템은 왜 그렇게 판단했는지를 보고 분석자가 insight를 획득할 수도 있는 장점이 있음

❍ 기술적인 측면에서는 클러스터링 기술과는 추출 요약, 패러프레이즈(paraphrase) 기술을 사용하여 다른 접근 방법도 고려해볼만 함

- 기존 클러스터링 방법은 대상 문서를 딥러닝 언어모델을 이용하여 임베딩하여 벡터화된 문서들을 유사도 계산하여 비슷한 문서를 같은 군집으로 분류
- 기존 방법 이외에도 추출 요약 기술과 패러프레이즈(paraphrase) 기술을 사용하여 유사도를 계산하는 방법도 시도하여 두 결과를 비교하여 더 나은 방법을 선택하거나, 분석 시스템에 앙상블(ensemble)[10] 기법을 사용하여 좀더 나은 결과를 기대할 수 있음
- 문서의 유사도를 비교할 때 모든 문서의 문장을 대상으로 하지 않고 추출 요약 기술을 이용하여 문서를 잘 요약한 문장 만을 대상으로 하여 유사도를 계산하며, 기존 유사도 계산 이외에도 패러프레이즈 기술을 사용한 유사도 계산 방법을 이용하여 다른 관점으로 유사도를 계산할 수 있음.

10) 일반적으로 앙상블 기법을 사용할 경우 2개 이상의 분석 엔진이 작동하기 때문에 작동 시간에 대한 고려가 필요함

- 추출 요약으로 추출된 문장들은 주제 키워드를 추출하거나 시스템 사용자가 문서를 파악할 때도 유용하게 사용될 수 있음.
- 추출 요약이나 패러프레이즈 기술의 경우 지도 학습(supervised learning)으로 학습 데이터 구축 등이 필요하지만 딥러닝 언어 모델을 이용한 전이학습 기법을 사용할 경우 처음부터 시스템을 구축하는 것보다는 용이하게 개발이 가능

□ 그 외 전통적인 키워드 검색 방식의 장점(검색 기준의 명확성)을 취하기 위해 과제 검색 시 꼭 포함되어야 하는 단어 및 제외할 단어를 입력할 수 있게 개선할 필요가 있음

참고문헌

김한해 외. (2017). 기계학습 기반 바이오의료분야 과학기술정보데이터 분석 · 활용 모형 개발. 한국과학기술기획평가원

김한해 외. (2018). 기계학습 기반 바이오의료분야 과학기술정보데이터 분석 · 활용 모형 고도화. 한국과학기술기획평가원

유거송 외. (2019). 바이오 · 의료분야 지능형 연구개발정보데이터 분석시스템의 예산배분 · 조정 활용기법 연구

관계부처 합동. (2019). 바이오헬스 산업 혁신전략

박진욱 외. (2019). ADC: Advanced document clustering using contextualized representations

오렐리앙 제롱. (2017). 핸즈온 머신러닝

Emilio Soria Olivas 외. (2009). Handbook Of Research On Machine Learning Applications and Trends: Algorithms, Methods and Techniques

Jacob Devlin 외. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Kevin Clark 외. (2020). ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators

Mandar Joshi 외. (2019). SpanBERT: Improving Pre-training by Representing and Predicting Spans

Mike Lewis 외. (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension

Tom B. Brown 외. (2020). Language Models are Few-Shot Learners

Ye Zhang 외. (2016). Rationale-Augmented Convolutional Neural Networks for Text Classification

Yutong Li 외. (2020). A Text Document Clustering Method Based on Weighted BERT Model

Zhenzhong Lan 외. (2019). ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations

공공 인공지능 오픈 API · DATA 서비스 포털, http://aiopen.etri.re.kr/

Gensim Doc2vec documentation, https://radimrehurek.com/gensim/models/doc2vec.html

How to Use t-SNE Effectively, https://distill.pub/2016/misread-tsne/

Laurens van der Maaten. t-SNE, https://lvdmaaten.github.io/tsne/

NIH Research Portfolio Online Reporting Tools (RePORT), https://report.nih.gov/

NSF Award Search, https://www.nsf.gov/awardsearch/

Scikit-Learn t-SNE documentation, https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html

UKRI Gateway, https://gtr.ukri.org/

Forbes, “What’s New In Gartner’s Hype Cycle For AI” 2019. 9. 25.