기관 2018-036

기계학습 기반 바이오의료분야 과학기술정보데이터 분석 · 활용 모형 고도화

김 한 해 외

KISTEP

기관 2018-036

기계학습 기반 바이오의료분야 과학기술정보데이터 분석 · 활용 모형 고도화

김 한 해 외

KISTEP

제 출 문

한국과학기술기획평가원 원장 귀하

본 보고서를 “기계학습 기반 바이오의료분야 과학기술정보데이터 분석활용 모형 고도화”의 최종보고서로 제출합니다.

2019. 1.

연구기관명 : 한국과학기술기획평가원
연구책임자 : 김 한 해 부연구위원
연 구 원 : 김 은 정 연구위원
연 구 원 : 김 주 원 연구위원
연 구 원 : 박 지 현 연구위원
연 구 원 : 유 거 송 부연구위원
연 구 원 : 김 행 미 부연구위원
연 구 원 : 홍 세 호 부연구위원
연 구 원 : 황 은 혜 연구원
연 구 원 : 이 현 익 연구원

요 약 문

1. 제목 : 기계학습 기반 과학기술정보데이터 분석·활용 모형 고도화

2. 연구목적 및 필요성

가. 연구목적

ㅇ 바이오의료분야 과학기술정보 분석·활용 모형 고도화 및 신규 분석·활용 기능 추가

- 딥러닝 등 최신 기계학습 방법론 기반으로 개발된 분석·활용 모형의 사용 편의성을 제고
- 예산배분조정 업무에 활용 가능한 관계성(유사정도) 분석 및 과제분류 기능 고도화, 신규 분석·활용 기능 추가 발굴 및 적용

ㅇ 17년도 정부 신약개발 R&D 과제 DB 구축

- 17년도 신약개발연구과제DB 구축을 위해 의약과제정보 분류모형을 시범적 적용하고 지속 활용 방안 마련

나. 연구필요성

ㅇ 과학기술정보통신부는 최근 과학기술지식정보서비스(NTIS)에서 정보 접근성, 이용자 편의성 측면에서 서비스를 전면 개편하고 제공하는 과학기술정보의 공개 범위를 대폭 확대함에 따라 해당 정보의 분석·활용 장이 마련됨

ㅇ 동 연구진은 KISTEP 업무 수행과정에서 투자포트폴리오 구축, 유사한 과제 검색 등의 업무 프로세스 개선을 위해 과학기술지식정보(빅데이터)에 인공지능 연구의 기반이 되는 기계학습을 접목한 분석·활용 모형의 고도화를 추진하고자 함

- 기 수행된 과제에서 2008~2016년간 축적된 바이오의료분야 과학기술정보(약 14만 건)를 활용하여 모형을 구축함

※ 연구과제 내용을 임베딩하여 내재된 공간상에서 벡터값을 할당하는 최신 기계학습 방법론(word2vec, doc2vec)을 적용

- 해당 모형의 고도화를 통해 실무적 활용 가능성을 제고하고 사용환경 및 제반여건을 확보하고자 함

3. 연구내용

가. 바이오의료분야 과학기술정보 분석·활용 모형 고도화 및 신규 분석·활용 기능 추가

[분석·활용 모형 인터페이스 개선]

ㅇ 현 커맨드 입력방식의 분석·활용 모형 인터페이스를 마우스 등의 입력도구 사용방식으로 개선하여 사용자 편의성 측면에서 범용성을 확보

[연구개발 사업 및 과제 간 관계성 분석 기능 고도화]

ㅇ 현 연구과제 관계성(유사정도) 분석기능을 세부사업수준으로 확대하고 바이오분야 R&D 투자포트폴리오 정보 분류기능 추가

[연구과제 분류기능 고도화]

ㅇ 바이오분야 R&D 투자포트폴리오 등 훈련정보 기반 범용적인 분류모형 구축을 통해 타 기술분야 확대적용을 위한 기반 마련

- 분류 기능 고도화 과정에서 17년도 신약개발연구과제DB를 구축하고, 모형의 신규과제 분류 결과와 전문가 분류 결과의 비교분석을 함께 수행

나. 딥러닝 기반 토픽 클러스터링 분석 방법론 연구

ㅇ Pubmed 논문 초록정보*를 바탕으로 국내 바이오의료분야 연구과제 추진 현황, 글로벌 연구내용 간 종합적 비교분석 수행

* 바이오분야 모든 논문의 초록정보를 무료 제공

다. 17년도 정부 신약개발 R&D과제 DB구축

ㅇ 생명보건의료분야 예산심의대상 정부연구개발사업 중 신약개발분야에 해당하는 사업('17년 기준 30개)을 대상으로, 국가과학기술정보서비스(NTIS)에서 제공하는 국가연구개발사업 조사분석데이터('17)의 과제정보를 활용

- 신약개발분야 전문가를 통해 신약개발 목적의 과제를 선별, 신약개발단계, 의약품 종류, 타겟 질환 등의 분류기준에 따라 과제 분류

4. 결론 및 시사점

가. 바이오의료분야 과학기술정보 분석·활용 모형 고도화

ㅇ 분석·활용 모형 고도화 결과 사업 간의 관계성(유사정도) 분석, 특정 사업 속성변화, 분류모형의 범용화 등에서 유의미한 결과를 도출함

- 연구개발사업의 분석하는 과정에서 어떠한 사업이 서로 간에 유사한지는 사업의 기획, 평가 등에 고려되는 사항으로 동 분석·활용 모형이 시각적·정량적으로 분석 결과를 도출한다는 점에서 의미가 있다고 사료됨
- 의약과제 분류모형의 경우 범용화를 통해 사용자 접근성을 제고함은 물론, 딥러닝 방법론에 국한된 기 모형을 고도화하여 6개의 방법론을 적용하여 사용자 선택의 폭을 넓힘

  ※ 분류모형을 활용하여 신규과제를 선분류하고 이를 전문가에게 제공함으로써 의약과제 분류연구의 추진효율성 제고가 가능할 것으로 판단됨

ㅇ 그럼에도, 동 과제모형이 사용자의 의사결정을 지원하는데 적합한 모형(전문가시스템)임을 숙지할 필요

- 최신 기계학습 방법을 적용한 모형일지라도 사전에 등재되지 않은 단어의 경우 유사성, 연관성 분석에 개선이 요구되는 등 일부 한계점이 존재하는 것을 확인
- 이는 실질적으로 최신 기계학습 방법론을 적용한 모형이 인간과 같이 문맥이나 핵심키워드, 행간에서 정보를 이해한다기보다는 사용자가 입력한 질문에 대한 답을 과거보다 정확성 높게 찾아준다는 의미로 해석될 수 있음

ㅇ 임베딩 기술을 적용하여 비정형데이터인 텍스트데이터를 벡터화하여 정형화된 데이터로 변환하고 이를 바탕으로 클러스터링, 분류 연구를 시도한 점은 동 연구과제의 특이점으로 볼 수 있음

- 향후 추가적인 연구를 통해 동 연구과제에서 제시된 접근법의 유용성이 검증될 경우 텍스트 기반 정보서비스 제공분야에 활용이 가능할 것으로 판단됨

나. 딥러닝 기반 토픽 클러스터링 분석 방법론 연구

ㅇ Pubmed의 MeSH Subject Heading을 활용한 세계 바이오의료 과학기술정보의 심도있는 분석이 가능

- Pubmed를 운영하는 미국 NLM(National Library of Medicine)에서 개발한 MeSH Subject Heading은 매주 업데이트되는 일종의 바이오분야 키워드 사전임

ㅇ NTIS에 수집되는 논문 성과 데이터와 Pubmed를 연동한 과학기술정보분석을 제안

- 국가연구개발사업 과제 수행을 통하여 생산된 논문은 NTIS 성과정보로 입력되므로, 해당 논문이 Pubmed에 등재된 경우 국내 연구과제와 Pubmed 문헌의 직접적 연결이 가능

- 과제가 아닌 사업 단위로 분석을 수행할 경우 데이터가 충분하므로 논문, 특허 개수뿐만 아니라 결과물의 내용적 분석을 통해 연구성과가 사업의 목적에 부합했는지 여부를 분석하는 추적평가 차원의 활용이 가능

다. 17년도 정부 신약개발 R&D과제 DB구축 및 투자포트폴리오 분석

ㅇ 신약개발단계별 투자의 경우 인력양성, 인허가 등의 인프라 부분에 대한 투자를 보다 강화할 필요

- 후보물질 도출 및 최적화 단계(약 32.8%)의 비중이 높은 반면, 인프라 중 인력양성, 제도·정책, 인·허가 부분의 투자 비중은 4% 이하 수준

- 신약개발 촉진을 통한 동 분야의 성장을 모색하기 위해서는 제도·정책 및 인프라적인 부분에 대한 정부 차원의 지원이 필요

목 차

contents

표 목 차

contents

그림 목차

contents

제 1 장

서론

제1장 서론

제1절 연구 배경 및 필요성

가. 연구배경 및 필요성

□ 과학기술정보통신부는 최근 과학기술지식정보서비스(NTIS)에서 정보 접근성, 이용자 편의성 측면에서 서비스를 전면 개편하고 제공하는 과학기술정보의 공개 범위를 대폭 확대함

※ NTIS는 세계 최초의 국가R&D정보 지식 포털로 현재까지 정부 예산으로 지원된 과학기술과제정보를 제공(약 500만건)

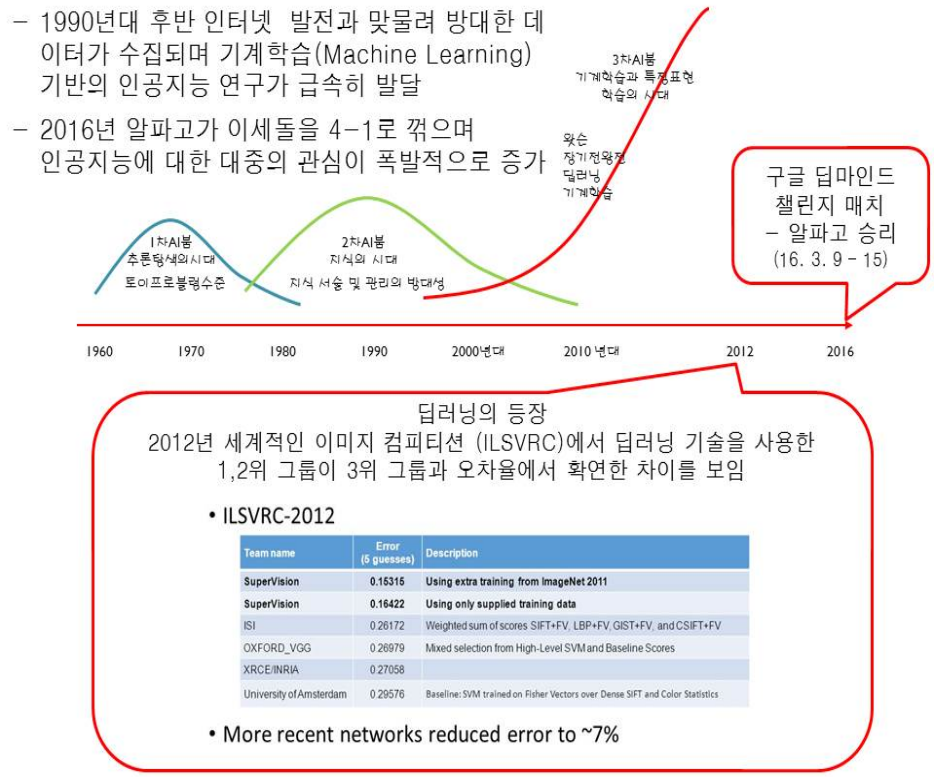
- ㅇ 정보 개방 항목을 개인정보 및 보안이 필요한 정보를 제외한 모든 항목을 확대하여, 당초 28%에서 '18년 78%까지 높임
- ㅇ 이에 따라 누구나 직접적으로 제공 정보의 가공이 가능해짐에 따라 과학기술지식 정보 분석 및 활용의 장이 마련됨

□ 2016년 개최된 다보스 포럼에서는 4차 산업혁명을 통한 변화와 혁신이 가속될 것으로 전망되었으며, 관련 요소기술인 빅데이터, 인공지능에 대한 사회적 관심도도 함께 증가

- ㅇ 인공지능이라는 용어는 1950년에 최초로 등장하였으며 인간 두뇌 모방(Artificial Intelligence, AI)에 대한 연구를 의미함
- ㅇ 인공지능은 1960년대에 붐이 일며 주목을 받은 시기가 있었으나, 당시의 인프라*로는 연구에 한계가 있었음

* 컴퓨팅파워가 부족하였고 충분한 데이터 확보 및 활용 측면에서 기술적 어려움 존재

- ㅇ 2010년대에 이르러 컴퓨터 하드웨어가 비약적으로 발전하고 인터넷 보급·확산으로 방대한 데이터가 축적되며 이러한 이슈가 점차 해소되었고 딥러닝과 같은 혁신기술이 개발·활용되며 인공지능 연구가 비약적으로 발전
  - 2016년 구글 딥마인드社의 알파고와 한국 프로기사 이세돌 9단과의 대국에서 알파고가 4:1로 승리하며 인공지능 기술의 전 세계적 관심이 증대됨



[그림 1-1] 인공지능 기술의 발전 동향

※ 출처: 마쓰오 유타카 (2015), 인공지능과 딥러닝-인공지능이 불러올 산업구조의 변화와 혁신, 동아엠엔비.(그림 재구성)

□ KISTEP 사업조정본부는 정부 R&D 예산배분조정을 심의하는 과정에서 NTIS 조사분석·성과데이터 등을 활용, 기술분야별 투자포트폴리오 구축 등의 업무를 매년 집중적으로 수행 중

o 바이오의료분야 투자포트폴리오 구축, 신규-기존과제 간 유사중복 검토, 신약개발 과제 통계브리프 발간 등을 위해 매년 전문가풀 구성·유지하며 정례적으로 인력기반 과제분석을 수행

□ 전술한 최근의 환경 변화를 감안, 과학기술지식정보의 분석·활용의 편의성 및 활용성 제고목적으로 과학기술지식정보(빅데이터)에 인공지능 연구의 기반이 되는 기계학습을 접목한 분석·활용 모형을 시범 개발*함

* 2017년 과학기술정보통신부가 공개한 국가과학기술지식정보(NTIS)를 바탕으로 '기계학습기반 바이오의료분야 과학기술정보데이터 분석·활용 모형 개발' 과제를 수행하였음

□ 기 수행된 과제에서 2008~2016년간 축적된 바이오의료분야 과학기술정보(약 14만 건)를 활용하여 모형을 구축함

○ 모형 구축과정에서 연구과제 내용 및 목표 등에서 제시되는 단어나 초록(문서)를 임베딩하여 내재된 공간상에서 벡터값을 할당하는 최신 기계학습 방법론(word2vec, doc2vec)을 적용

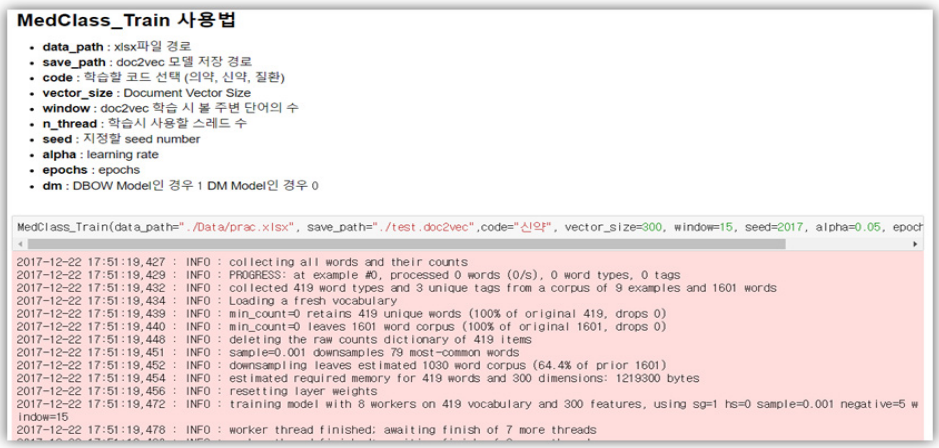
※ word2vec은 2013년 구글에서 제안한 자연어처리를 위한 기계학습 방법론으로, 단어 자체가 지니는 의미를 다차원 공간에서 벡터화하여 표현, 벡터연산을 통해 단어간 의미나 관계성을 추론

○ NTIS 조사분석데이터를 가공, 투자포트폴리오를 구축하는 과정에서 기초자료생성, 전문가풀의 구성유자관리, 결과물 교차검증 등으로 발생하는 실무진의 업무 부담을 완화하고자 실무적 사용이 가능한 기계학습 기반 업무지원 및 자동화 모형 개발을 목표 하였음

□ 본 연구진은 분석·활용모형의 실무적용 가능성을 확대하고자 동 과제를 통해 모형 고도화를 이루고 사용 환경 및 제반여건을 구축하고자 함

○ 현 분석·활용 모형은 커맨드 방식의 인터페이스*를 지니고 있어 프로그래밍에 익숙하지 않은 경우 사용이 여의치 않음(그림 1-2 참고)

* 사용과정에서 파이썬 등 같은 프로그래밍 언어의 이해가 필요



[그림 1-2] 의약과제정보 분류모형 시연 과정

※ 출처: 김한해 외 (2017), 「기계학습 기반 바이오의료분야 과학기술정보데이터 분석·활용 모형 개발」, 한국과학기술기획평가원.

○ 사용자가 분석·활용 모형을 보다 직관적으로 사용할 수 있도록(활용성 제고) 사용자인터페이스를 현 커멘트 입/출력 방식에서 마우스 등의 입력도구를 사용하는 방식으로 개선할 필요

- ㅇ KISTEP의 예산배분조정 기능이 세부사업수준에서 이루어지는 점을 감안, 과제 외에 세부사업 수준에서도 유사정도를 분석할 수 있는 수준으로 고도화 필요

□ 인공지능의 기반이 되는 기계학습을 접목한 동 연구의 추진은 연구자의 과학기술적 역량을 강화하고 업무효율성을 높임과 동시에, 데이터과학에 근거한 바이오의료분야 과학기술지식정보의 새로운 분석활용방안을 제시할 수 있을 것으로 기대

- ㅇ 기계학습 기반 명확한 척도에 근거한 바이오의료분야 과학기술정보데이터 분석·활용 모형 개발 시 실무적용이 가능할 것으로 예상
  - 과제정보의 경우 부처명, 세부사업명, 정부연구비, 연구목표 및 내용 등 차별성 있는 수십가지의 지표로 구성되어 기계학습을 적용하는데 최적의 데이터 포맷을 지니고 있음
- ㅇ 과학기술 핵심키워드 간 연관성 및 특이적 문맥을 이해할 수 있는 분석·활용 모형 구축으로 타 기술분야 확대 적용가능성 모색
  - 과학기술지식정보 분석·활용을 위해서는 특이적인 과학기술용어까지 이해할 수 있는 자연어 처리 모형 개발이 필수적
- ㅇ 또한, 동 과제 추진 과정에서 창출되는 성과물을 바탕으로 차년도 신약개발 정부 R&D 투자포트폴리오 분석 보고서 발간을 위한 비용 및 시간(전문가 활용, 업무가중 등)을 최소화 할 수 있을 것으로 사료됨

제2절 연구 목표 및 방법

가. 연구 목표

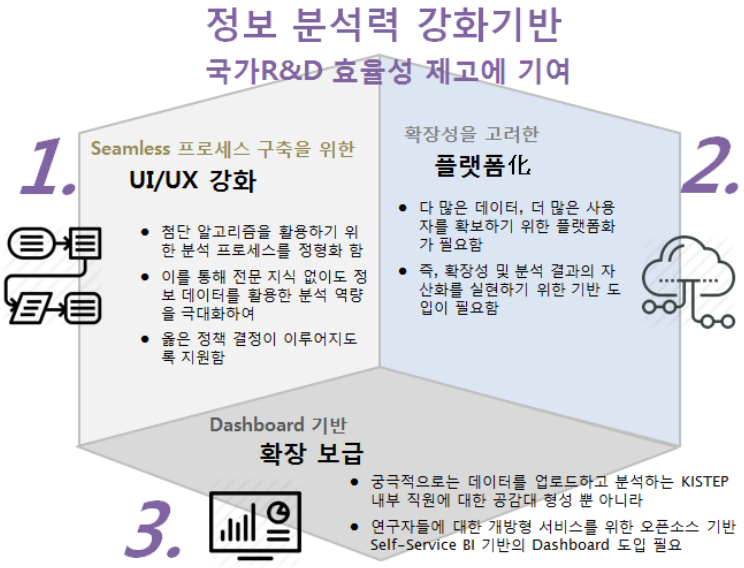
□ 바이오의료분야 과학기술정보 분석·활용 모형 고도화 및 신규 분석·활용 기능 추가

- ㅇ 딥러닝 등 최신 기계학습 방법론 기반으로 개발된 분석·활용 모형의 사용 편의성을 제고(그림 1-3참고)
- ㅇ 예산배분조정 업무에 활용 가능한 관계성(유사정도) 분석 및 과제분류 기능 고도화, 신규 분석·활용 기능 추가 발굴 및 적용

□ 17년도 정부 신약개발 R&D 과제 DB 구축

- ㅇ 17년도 신약개발연구과제DB 구축을 위해 의약과제정보 분류모형을 시범적 적용하고 지속 활용 방안 마련

※ 결과물은 차년도 투자우선순위 설정 및 2018년도 통계브리프 발간 기초자료 활용 예정



[그림 1-3] 바이오의료분야 과학기술정보데이터 분석·활용 모형 고도화 방안

※ 출처: 김한해 외 (2017), 「기계학습 기반 바이오의료분야 과학기술정보데이터 분석·활용 모형 개발」, 한국과학기술기획평가원.

나. 연구 내용

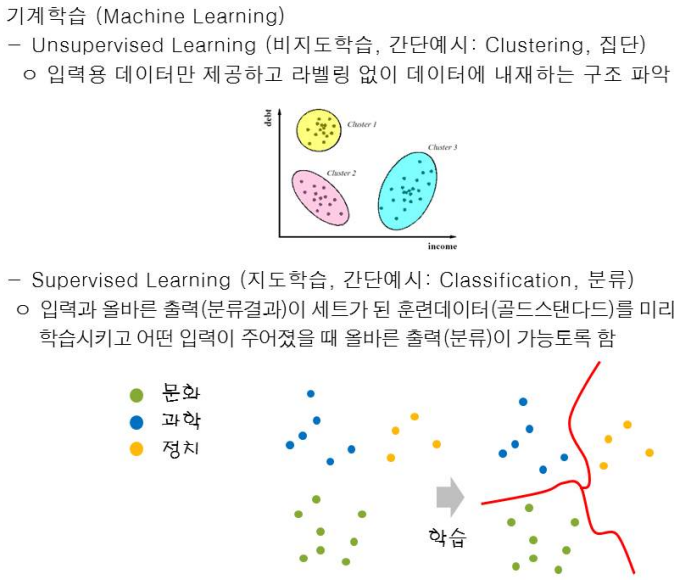
□ 동 연구는 연구 목표 실현을 기계학습을 활용하는데 특징점이 있음

○ 인공지능 기술의 토대가 되는 기계학습은 비지도학습(Unsupervised Learning)과 지도학습(Supervised Learning)으로 구분(그림1-4 참조)

※ 자율주행, 알파고 등의 등장으로 이에 기반이 되는 강화학습 또한 주목을 받고 있음

○ 비지도학습은 라벨링이 되어있지 않은 데이터를 기계에 제공하여 데이터의 내재된 구조를 파악

○ 지도학습은 준비된 훈련정보(골드스탠다드)를 기계에 학습시켜 분류모형(Classifier)을 구축하고 훈련정보를 기반으로 입력데이터를 분류



[그림 1-4] 기계학습 방법 구분 예시

※ 출처: 김한해 외 (2017), 「기계학습 기반 바이오의료분야 과학기술정보데이터 분석·활용 모형 개발」, 한국과학기술기획평가원.

□ 기계학습을 활용한 기 수행된 분석·활용모형 개발 과제에서 다음 3가지 분석 모듈이 개발되었음

○ 기계학습 기반 과학기술지식정보 자연어처리 모형

- 과학기술지식정보 분석을 위해 과학기술지식정보을 훈련정보(goldstandard)로 활용, 과학기술 특이적 단어와 문맥을 이해할 수 있는 자연어 처리 시스템을 구축

- 동 자연어 처리 시스템은 동 과제 내 연속되어 개발될 바이오의료분야 유사연구과제 및 신약개발연구과제DB 분류 시스템의 기반 기술로 활용

ㅇ 바이오의료분야 연구과제 간 유사·중복 검증 모형 개발

- 사용자가 학습에 사용되는 항목(연구과제명, 연구목표, 연구내용 등)을 직접 선택, 이를 바탕으로 모형을 학습하고 바이오의료분야 연구과제 간 유사정도(정량화된 수치를 제시)를 판단할 수 있는 의사결정 지원 환경 구축

※ (관련 유사연구) 국가연구개발사업 유사중복 검색 시스템 개발을 위한 실증연구, 홍세호, KISTEP 2013

ㅇ 신약개발 과학기술지식정보 분류 모형 개발

- 기 보유한 신약개발연구과제DB를 훈련정보로 활용, 주어진 신약개발과제의 개발단계, 의약품종류, 질환별 분류를 수행할 수 있는 모형을 개발

- 이를 위해 단순한 키워드 중심이 아닌, 과학기술적 문맥을 고려하거나 핵심키워드 간 연관성을 고려한 신경망 네트워크 구축

※ 전문가를 활용하여 2016년도 신약개발 정부 R&D 과제DB를 구축하고 해당 분류모형을 검증함

□ 동 연구는 기 수행 모형의 고도화를 위해 정부 바이오의료분야 과학기술정보데이터를 확장(17년도 조사분석 데이터 포함)하고 신규 분석·활용 기능 추가 및 사용자 편의성을 제고하고자 함

ㅇ 바이오의료분야 과학기술정보 분석·활용 모형 고도화 및 신규 분석·활용 기능 추가

[분석·활용 모형 인터페이스 개선]

- 현 커맨드 입력방식의 분석·활용 모형 인터페이스를 마우스 등의 입력도구 사용방식으로 개선하여 사용자 편의성 측면에서 범용성을 확보

[연구개발 사업 및 과제 간 관계성 분석 기능 고도화]

- 현 연구과제 관계성(유사정도) 분석기능을 세부사업수준으로 확대하고 바이오분야 R&D 투자포트폴리오 정보 분류기능 추가

[연구과제 분류기능 고도화]

- 바이오분야 R&D 투자포트폴리오 등 훈련정보 기반 범용적인 분류모형 구축을 통해 타 기술분야 확대적용을 위한 기반 마련

ㅇ 딥러닝 기반 토픽 클러스터링 분석 방법론 연구

- Pubmed 논문 초록정보*를 바탕으로 국내 바이오의료분야 연구과제 추진 현황, 글로벌 연구내용 간 종합적 비교분석 수행

* 바이오분야 모든 논문의 초록정보를 무료 제공

ㅇ 17년도 정부 신약개발 R&D과제 DB구축

- 17년도 정부 신약개발연구과제 정보 획득 후 의약과제정보 분류모형을 적용한 결과를 바탕으로 실무적 활용 방안 마련

다. 추진전략

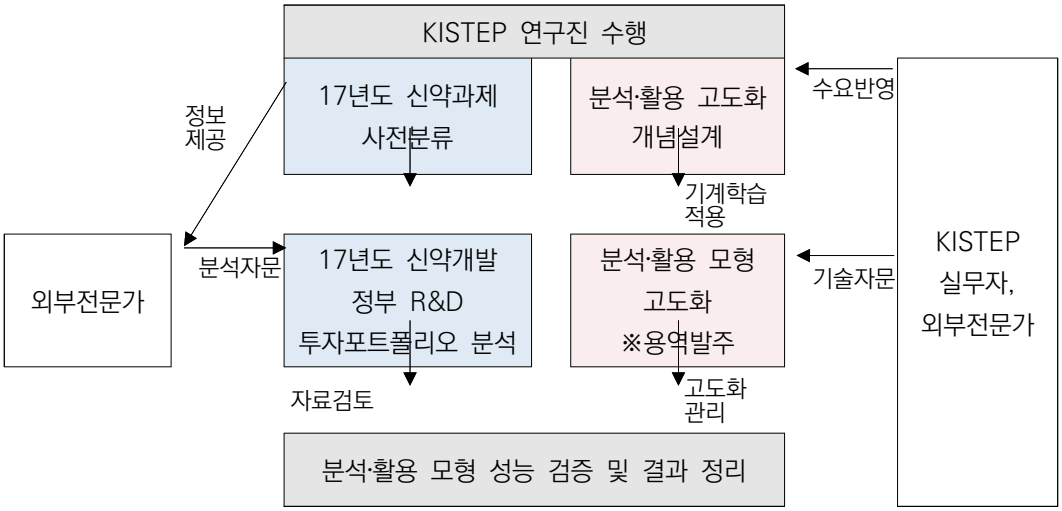
□ KISTEP 연구진은 바이오의료분야 과학기술정보 분석·활용 모형 고도화 및 신규 모형 개발을 위한 개념설계를 담당하고, 외부전문가 자문 및 기술용역 등 추진

※ 모형 고도화 시 개선된 인터페이스 구현, 최신 기계학습 방법론 심층적용을 위해 기술용역 발주

ㅇ 참여연구진의 개념설계 수준 및 기계학습 이해도 제고, 연구내용·성과 발표, 개선·활용 방안 도출을 위한 관련 교육 이수, 전문가 간담회 등 이행

ㅇ 분석활용 모형 사용자(KISTEP 연구진), 기계학습 전공자, 신약개발과제 분석 워킹그룹 등 관련 전문가를 활용하여 모형 고도화 및 의약과제정보 분류 시범연구 수행

〈표 1-1〉 바이오의료분야 과학기술정보 분석·활용 모형 고도화 추진전략



라. 연구방법

□ 바이오의료분야 과학기술정보 분석·활용 모형 고도화 및 신규 분석·활용 기능 추가

[인터페이스 개선]

- ㅇ 실제 사용자의 편의성 및 활용성을 제고하기 위한 개념설계를 추진하고 이의 적용 결과를 검토하여 개선사항 발굴 및 반영
- ㅇ 모형의 분석결과는 그래프(네트워크 등)나 설명 문구 등이 자동적으로 제공되는 방식으로 구현

[연구개발사업 및 과제 간 관계성 분석 기능 고도화]

- ㅇ 바이오의료분야 세부사업의 연구내용 및 목적, 세부사업별 과제리스트를 정리하고 이를 최신 기계학습 방법론에 접목, 세부사업/내역사업/연구과제수준에서 관계성(유사정도)를 정량적으로 제시*

  * 벡터값 기반의 코사인 유사도 수치 등

[연구과제 분류기능 고도화]

- ㅇ 기 의약과제 분류모형을 바탕으로 범용적 분류 모형을 개발하고 바이오분야 R&D 투자포트폴리오 등 별도의 훈련정보를 적용하여 범용적 활용 여부 검증

[조사·분석 과학기술표준분류 검증 자동화]

- ㅇ 과거 조사·분석 데이터 상의 각 과제의 과학기술표준분류 맵핑 결과를 학습, 기 훈련정보를 활용하여 분류성능을 점검하고 향후 활용 방안 모색

  ※ 의약과제 분류 범용화 모형과 유사 작동환경에서 구현되나 조사분석 데이터이 중요성을 감안 별도의 모듈 구성

□ 딥러닝 기반 토픽 클러스터링 분석 방법론 연구

- ㅇ 국내 바이오의료분야 과제와 Pubmed 논문 초록을 각각 클러스터링*하여 세부분야별 중점 연구주제와 연구수행 빈도 등을 분석하고, 글로벌 연구수행 내용 대비 국내 연구의 강점/약점을 분석

  * 비지도 기계학습 알고리즘으로 유사한 문서를 주제별로 묶는 분석법

- ㅇ Pubmed 논문초록 학습 시 Pubmed의 MeSH 체계*를 고려하여 국내 바이오의료분야 과제 간 분석 용이성 확보

  * MeSH(Medical subject headings)는 Pubmed에서 사용하는 바이오분야 표준 키워드 체계임

□ 17년도 정부 신약개발 R&D과제 DB구축

- ㅇ 17년도 과학기술지식정보 획득 후 KISTEP 신약개발연구과제DB 구축 기준을 적용하여 17년 신약개발 분석대상 연구과제를 선정하고 분류 연구를 수행
- ㅇ 분류는 의약과제 분류모형 및 전문가 그룹에 의한 두 가지 형태로 추진하되, 결과 비교를 통해 실무 적용 가능성을 검토
- ㅇ 또한, 모형이 사전 분류한 결과에 대한 전문가 검토 및 의견청취를 통해 의약과제 분류연구의 개선방안 모색

제2장

기계학습방법의 이론적 배경

제2장 기계학습방법의 이론적 배경

제1절 기계학습 정의 및 발전동향

가. 기계학습의 정의

□ Tom Mitchell은 기계학습을 경험(E), 성능(P), 태스크(T)의 관점에서 정의

- ㅇ 태스크와 관련된 경험이 늘어나면서 성능이 향상되는 프로그램을 기계학습 프로그램으로 정의함[1)]
- ㅇ 즉, 기계학습은 수집된 데이터로부터 태스크 해결하는 성능을 향상 시킬 수 있음

□ 통계학적 관점의 기계학습은 응용 통계학 기법의 일종[2)]으로 컴퓨터를 이용하여 복잡한 함수를 통계적으로 추정하는 것을 의미

나. 인공지능 대표성과를 통해 바라본 기계학습 발전 동향

□ 인공지능 태동 시기에는 자동번역 등의 분야에서 큰 기대를 받았으나 기대를 충족시키지는 못하였음

□ 초창기 인공지능/기계학습은 인간의 문제 해결 방식을 규칙으로 표현하고 이의 결합을 통해 추론이 가능한 전문가 시스템을 구현

- ㅇ 광물 탐사에 활용될 목적으로 개발된 프로스펙터(Prospector)[3)]는 광물 탐사과정에서 수집된 정보를 바탕으로 문제 해결책을 제시
  - 수집된 정보를 바탕으로 광석 매장 특성 및 매장 가능성 등을 추정하고 광석 데이터베이스를 통해 광물자원 분포 관련 분석 결과 제공
  - 인간 지질 전문가 지식을 통해 추론 규칙을 수집하고 베이즈 룰을 이용하여 가장 가능성이 높은 가설을 선택
  - 일부 도메인에서 인간 전문가 수준 혹은 그 이상의 문제해결 성능을 보여주었으나, 경험을 토대로 학습할 수 없다는 단점이 존재하여 발전 가능성에 제약

1) T. M. Mitchell. Machine Learning, McGraw-Hill Education, 1997.
2) I. Goodfellow et al. Deep Learning, MIT Press, 2016.
3) P. E. Hart and R. O. Duda, "PROSPECTOR - a computer-based consultation system for mineral exploration," Artificial Intelligence Center, SRI International, Technical Note, No. 155, 1977.

□ 1990년대 이후 관련 이론이 고도화 되며 실시간 응답이 가능한 초고성능의 전문가 시스템 결과물이 등장함

ㅇ 1987년 체스 그랜드마스터에게 승리하면서 놀라운 성능을 입증한 딥블루(Deep Blue)[4]는 전문가의 체스 플레이 과정을 규칙화하고 신속히 문제 공간을 탐색하는 초고성능 컴퓨터 기반 전문가 시스템으로 볼 수 있음

- 인간 전문가(체스 그랜드마스터)가 세운 체스 오프닝전략을 데이터베이스화 하고 위치 평가 함수를 계산, 휴리스틱 알고리즘을 토대로 탐색 공간의 크기 감소

ㅇ 2011년 유명 퀴즈 쇼에 출현한 IBM 왓슨(Watson)[5]은 인간과 동일한 조건으로 게임에 임하여 인간에게 승리하는 업적을 달성함

- 왓슨은 (1) 자연어로 주어진 질문 처리, (2) 오픈 도메인에 대한 질문 해결, (3) 인간 경쟁자보다 빠르고 정확한 답변 탐색 측면에서 놀라움을 선사
- 그러나 인간과 같이 질문을 이해한 것이 아닌 자연어 질문을 처리하고 그에 부합하는 답을 탐색하는 방식으로 작동
- 이를 위해 기계학습 기법을 활용하여 수백개의 알고리즘을 토대로 여러 가설과 증거를 결합하여 정확할 가능성을 계산

  ※ 기계학습을 통한 증거 결합 과정에서 성능향상 가능

- 확정 지식의 단시간 내 결합 능력 외에도 오픈 도메인 질의의 정확한 처리가 가능한 성능을 감안하였을 때 딥블루 대비 발전한 모델로 볼 수 있음

□ 전문가 시스템은 제한된 분야에서 문제해결의 성공가능성을 보여주었으나 경험을 기반으로 학습이 이루어지지 않기 때문에 문제를 해결한 경험이 성능 향상과 연계되지 못하는 한계점이 존재

□ 최근 등장한 알파고는 앞서 소개된 전문가 시스템과는 달리, 경험을 통해 스스로의 문제해결 성능을 향상시킬 수 있는 특징을 보여줌

ㅇ 구굴 딥마인드社 개발한 알파고는 방대한 크기의 문제 공간으로 인해 컴퓨터로 처리하기에 어렵다고 평가받은 바둑 인공지능으로,

4) M. Campbell, A. J. Hoane Jr., and F.-h. Hsu, "Deep Blue," Artificial Intelligence, Vol. 134, 2002, pp. 57-83.
5) D. A. Ferrucci, "Introduction to "This is Watson"," IBM Journal of Research and Development, Vol. 56, 2012, pp. 1:1-1:15.

ㅇ 2016년 이세돌과의 대국에서 승리하며 기계 학습 분야의 주목할만한 성과를 보여주었는데, 이는 기계학습 관점에서도 기존 시스템과는 차별화 되는 것이었음

- 딥뉴럴넷을 기반으로 작동하는 알파고는 인간의 기보 데이터를 활용하여 일차적으로 포석을 위한 판단훈련을 시킨 후 알파고 간자체 대국을 통해 판단 시스템을 강화하였음

- 알파고는 이세돌과의 대국에서 인간의 기보에서는 볼 수 없었던 새로운 포석을 보여주었는데, 이는 경험을 통해 스스로의 문제 해결 성능을 보여준 사례라고 할 수 있음

□ 향후 인공지능/기계학습의 발전에는 데이터 기반 학습의 제약을 얼마만큼 해소할 수 있을지가 관건이 될 수 있음

ㅇ 컴퓨팅파워 및 빅데이터 기술을 나날이 발전하고 있고 알파고에 활용된 딥뉴럴넷은 이미지 인식과 음성 인식 분야에서 이미 인간의 수준에 근접하거나 동등한 수준의 성능을 달성함

ㅇ 그러나 대상 분야에 따라 알파고와 같은 기계학습 접근이 항상 가능한 것이 아님을 염두 할 필요

- 가령, 개인 정보의 경우에는 데이터 수집 및 처리에 제한이 있기 때문에 이를 활용한 인공지능/기계학습 연구는 어려울 수 있음

제2절 기계학습 적용 방법론 탐색

가. 지도학습(Supervised learning)

1) 지도학습의 개념

□ 지도학습은 주어진 입력 데이터(**x**)와 연관된 결과(**y**)가 존재하는 상황에서 훈련데이터를 이용하여 입력-결과 매칭 모델을 학습하는 방법

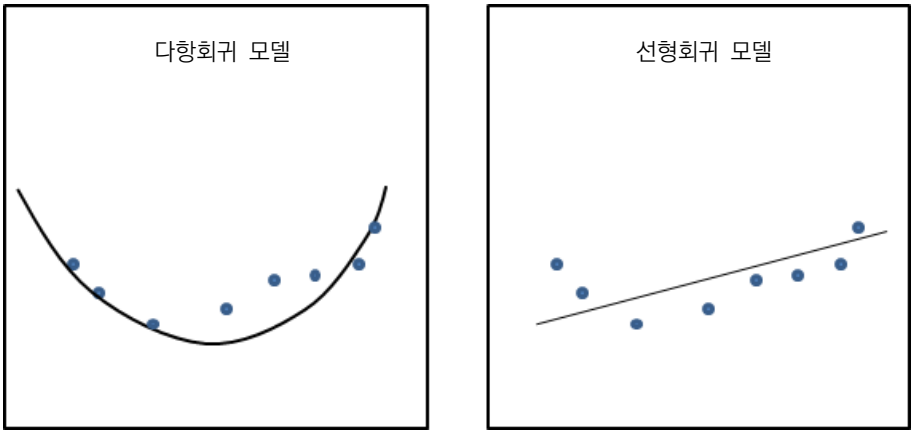
ㅇ 결과 값을 모르는 입력데이터가 제공되었을 때 해당 입력 데이터의 결과를 예측하는 학습 방법으로 분류(classification)와 회귀(regression)로 나뉠 수 있음

ㅇ 분류(Classification)

- 입력데이터 **x**의 $y \in \{1, \ldots, C\}$(여기서 C는 클래스나 카테고리 수 의미) 매칭 작업 의미
- 예를 들어, 남성에 해당 하는 사진을 수집하여 남성 클래스를, 여성에 해당하는 사진을 수집하여 여성 클래스 구성하여 남성과 여성을 분류하는 모델을 생성하고 모델의 학습에 사용되지 않은 사진이 입력되었을 때 남/여 여부를 구분하는 작업이 분류에 해당

ㅇ 회귀(Regression)

- 훈련 데이터로에서 입력데이터(**x**)와 연관결과(**y**)를 연결하는 모델 $\hat{f}$ $(f: \mathbb{R}^n \rightarrow \mathbb{R})$을 찾는 것으로, **x**,**y**는 실수로 주어짐
- 과거 주가변동 지수를 모델링하여 향후 주가 지수를 예측 모델 등이 회귀에 해당
- [그림 3-1]은 회기분석(regression)를 실시한 결과로, 원으로 표시된 훈련 데이터에 차수가 2차 다항함수와 1차인 선형 함수를 각각 적용한 것으로 (a), (b) 두 모델 모두 주어진 데이터를 완벽하게 설명하고 있지는 않음
- [그림 2-1]을 통해 학습을 통한 모델의 표현 능력에 따라 예측 정확도가 가변적임을 알 수 있음



[그림 2-1] 다항회귀(Polynomial regression)와 선형회귀(Linear regression) 모델 예시

- 데이터 기반 모델링을 위해서 분류와 회귀 모두 예측 성능을 측정할 수 있는 성능 측도가 요구됨
  - 평균제곱근에러(Mean Squared Error, MSE)는 m개 테스트 데이터($\boldsymbol{y}^{(\text{test})}$)와 이의 예측값($\hat{\boldsymbol{y}}^{(\text{test})}$)에 대해 다음과 같이 정의:

$$\text{MSE}_{\text{test}} = \frac{1}{m}\sum_i \left(\hat{\boldsymbol{y}}^{(\text{test})} - \boldsymbol{y}^{(\text{test})}\right)_i^2$$

  - 잔차제곱합(Residual Sum of Square, RSS)는 테스트 데이터($\boldsymbol{y}^{(\text{test})}$)와 이에 대한 회귀 함수 ($f_{\boldsymbol{\theta}}$)에 대해 (여기서 $\boldsymbol{\theta}$는 회귀 함수 결정 인자 벡터) 다음과 같이 정의:

$$\text{RSS}(\boldsymbol{\theta}) = \sum_i^m \left(\boldsymbol{y}^{(\text{test})} - f_{\boldsymbol{\theta}}\left(\mathbf{x}^{(\text{test})}\right)\right)_i^2$$

  - 쿨백-라이블러 발산(Kullback-Leibler divergence, KL divergence)는 두 개 확률 분포 p, q 차이 측정을 위해 사용되며 다음과 같이 정의:

$$\text{D}_{\text{KL}}(p||q) \equiv \sum_{k=1}^{K} p_k \log \frac{p_k}{q_k}$$

2) 지도학습 기반 알고리즘

□ 선형회귀(Linear Regression) 방법

- 선형회귀는 주어진 독립 변수 (x), 종속 변수 (y) 사이의 선형 관계 ($y \approx \beta_1 x + \beta_0$) 존재를 가정하는 방법

- 훈련 데이터로부터 선형모델 $\hat{y} = \hat{\beta}_1 + \hat{\beta}_0$를 추정 시에는 잔차제곱합(RSS)을 사용
- n개 훈련 데이터가 주어지면,

  $\text{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \cdots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$과 같이 계산됨
- 잔차제곱합 최소화를 위한 미분 시, 선형 관계 설명 계수 $\hat{\beta}_1$, $\hat{\beta}_0$는 다음과 같이 구할 수 있음:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

여기서 $\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$, $\bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$.

□ 로지스틱 회귀(Logistic Regression)

- 로지스틱 시그모이드(Logistic sigmoid) $\sigma(x) = \frac{1}{1+\exp(-x)}$는 0과 1사이 값으로 매핑되므로 확률을 도입을 통한 논리 전개 상황에서 사용됨
- 로지스틱 회귀는 로지스틱 시그모이드를 활용한 회귀기법으로, 다음과 같이 타깃 카테고리가 0과 1만 존재하는 문제에 대해 $p(X) = \Pr(Y = 1|X)$와 X의 관계 모델링 시 사용될 수 있음
  - 선형 관계를 이용하여 로지스틱 함수를 보다 구체적으로 표현하면 $p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$가 되며 해당 식은 $\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}$로 변형되어 표현할 수 있음
  - 로그 연산자를 적용하면 $\log\frac{p(X)}{1-p(X)} = \beta_0 + \beta_1 X$로 정리되며, $\log\frac{p(X)}{1-p(X)}$ 는 선형 관계로 표현될 수 있음을 알 수 있음
- 로지스틱 회귀는 회귀계수(regression coefficient) β_0, β_1의 추정을 위해 최대우도법(maximum likelihood)을 이용함

- 학습과정에서 $p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$ 를 이용하여 i번째 훈련 데이터를 확률로 나타낼 때, 출력 카테고리 0과 1에 대한 모든 훈련 데이터 확률의 최대값을 만족하는 회귀계수를 찾는 것이 훈련의 목적이며, 이 때 확률 값의 우도(likelihood)는 다음의 우도함수(likelihood function)으로 표현:

$$l(\beta_0,\beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1-p(x_{i'}))$$

- 회귀계수를 벡터 $\boldsymbol{\beta}$로 표기하고 최초의 우도에 로그 연산자를 씌울 경우 새로운 우도는

$$l(\boldsymbol{\beta}) = \sum_{i}^{N} \{y_i \log p(x_i;\boldsymbol{\beta}) + (1-y_i)\log(1-p\log p(x_i;\boldsymbol{\beta}))\}$$

로 표시할 수 있으며 각각 회귀계수에 대해 $\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_0} = 0$, $\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_1} = 0$을 계산, 회귀계수를 추정함

□ 다항회귀(Polynomial Regression)

○ 다항 회귀는 1차수를 이용하여 표현에 한계가 있는 선형회귀와는 달리 $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$를 최고 차수 d의 방정식으로 교체하여 다음과 같이 표현:

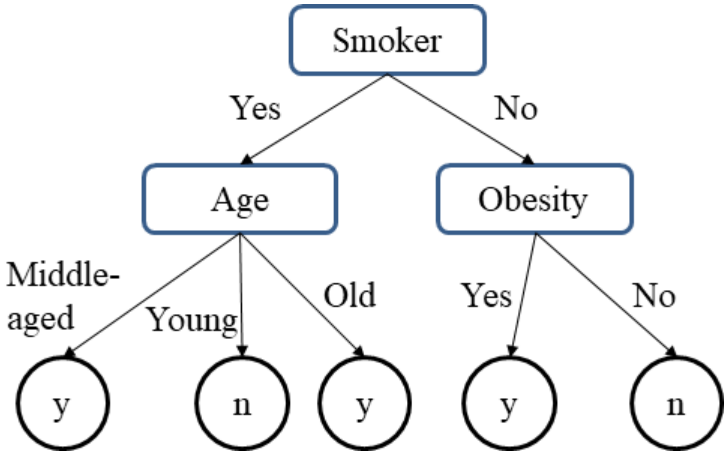
$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \cdots + \beta_d x_i^d + \epsilon_i$$

- 최고 차수 d가 충분히 큰 다항식을 사용할 경우 극단적 비선형 관계의 표현도 가능 (여기서 ϵ_i는 오류를 표시)
- 최고 차수 d의 크기가 과도할 경우 훈련 데이터 학습도 과도하게 이루어질 수 있으므로 지나치게 큰 차수 사용은 적절하지 않음

□ 의사결정트리(Decision Tree)

○ 의사결정트리는 [그림3-2]와 같이 트리구조의 모양을 지니고 있으며, 탐색하려는 공간을 중첩 분할 후 분류나 회귀 방법을 수행함

- 중첩 분할 시 가장 유용한 특성(feature)을 기준으로 공간을 분할하는데, 이때 생성되는 구조가 트리와 유사하여 의사결정트리라고 함
- 분류를 목표로 하는 의사결정트리는 가장 상위 루트 노드(root node)부터 말단 노드에 도달하는 경로를 구성하는 노드들이 데이터를 구성하고 있으며, 말단 노드(leaf node)가 카테고리에 해당하게 됨
- 카테고리에 부여되지 않은 임의의 데이터가 주어졌을 때 루트 노드부터 출발하여 해당 특성의 특성 값 확인 후 다음 노드로 이동하는 방식으로 말단노드에 도달하게 되고, 이 때 말단노드의 카테고리 값이 해당 데이터에 부여됨



[그림 2-2] 의사결정트리(Decision tree) 예시

- 의사결정트리 방법에서는 각 노드에 특성을 할당하는 것이 매우 중요한 과정이며 특성 선택에 따른 엔트로피 변화가 판단의 근거로 작용
  - 의사결정트리의 방법 중 C4.5 결정트리구성 방법은 아래와 같으며, 훈련 데이터에 존재하는 카테고리 수를 C라 할 때, 우선 전체집합의 엔트로피(Entropy)는 다음으로 정의됨:

$$\text{Entropy}(S) \equiv \sum_{i=1}^{C} -p_i \log_2 p_i$$

  - 여기서 p_i는 i번째 카테고리에 포함되는 데이터 비율 의미
  - 의사결정트리를 구성하는 각각의 서브트리 루트 부위에 해당하는 데이터 집합을 S라 하면, 루트 특성의 구성은 해당 특성을 알았을 때 발생되는 엔트로피 변화(정보이득, information gain) 값이 최대가 되는 특성으로 선택:

$$\text{Gain}(S, A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entorpy(S_v)$$

- 의사결정트리는 루트 노드에서부터 각 노드의 특성을 결정한 후, 정보이득을 이용하여 노드의 가지(branch)에 속하는 서브트리 루트 노드의 특성을 결정하는 방식으로 훈련됨

□ 서포트벡터머신(Support Vector Machine, SVM)

o 서포트벡터머신은 최대마진분류기(maximal margin classifier) 개념을 일반화시킨 분류(Classification) 방법으로, 서포트벡터(Support Vector)는 초평면(hyperplane)에 가장 근접한 데이터 인스턴스를 지칭

- 2개 클래스 A, B가 존재 할 때, 문제 표현이 올바르다는 가정 하에, 같은 클래스에 속한 데이터 인스턴스는 서로 가까이 있을 것으로 추정이 가능

- 데이터가 p차원의 벡터로 표시될 때, 다음 부등식을 충족하는 초평면(hyper-plane)*을 사용하여 두 개 클래스의 구분이 가능:

$$\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_p x_p > 0$$

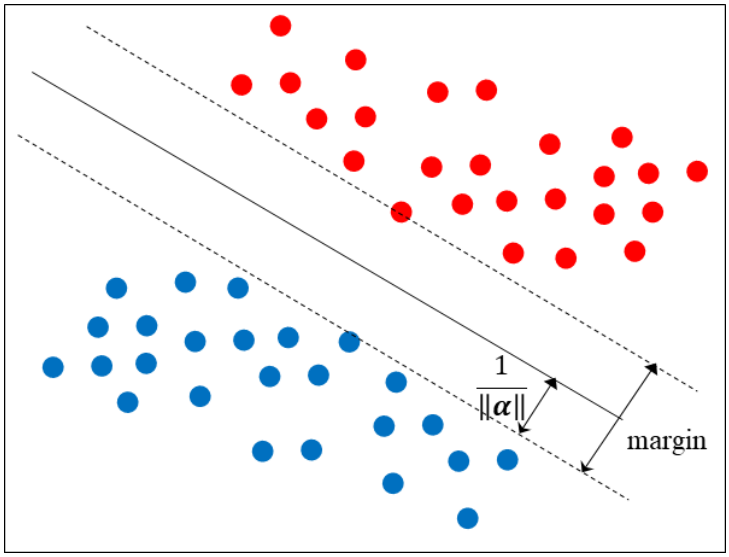
$$\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_p x_p < 0$$

* p차원 공간에서 클래스를 구분하는 $p-1$ 차원의 서브 공간으로 [그림3-3]에서 데이터를 구분하고 있는 직선이 해당

- 제시된 부등식을 일부 수정하여 클래스 A 레이블을 1, 클래스 B 레이블을 -1로 표기 시, 두 개 클래스를 분류하는 초평면은 다음 부등식을 충족하는 초평면으로 표현이 가능함:

$$y_i(\alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \cdots + \alpha_p x_{ip}) > 0$$

여기서 $i = 1, \ldots, |\mathcal{D}|$.



[그림 2-3] 서포트벡터머신(SVM) 예시

- [그림 2-3]과 같이 간단한 예의 경우에도 두 클래스를 분류하는 초평면의 수는 무한할 수 있으므로 하나의 초평면으로 특정될 필요
- 최대마진초평면(maximal margin hyperplane)은 각 클래스에 속한 데이터 인스턴스 중 초평면에 가장 근접한 원소의 거리가 가장 최대화 되는 초평면을 의미
  ※ 최대마진초평면 선택 시 미지 데이터 처리 가능성 제고
- 보다 특정하려는 초평면을 $\{\mathbf{x}: g(\mathbf{x}) = \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}\}$로 표현했을 때 해당 초평면과 가장 근접한 원소간의 거리(간격) ($\frac{y_i g(\mathbf{x}_i)}{\|\boldsymbol{\alpha}\|} = b$)를 최대화 하는 $\boldsymbol{\alpha}$를 찾을 경우 최대마진초평면을 설정할 수 있음
- 이때의 초평면을 특정하기 위해 $b\|\boldsymbol{\alpha}\| = 1$이라는 제약을 도입하고, 거리는 $b = \frac{1}{\|\boldsymbol{\alpha}\|}$로 표기
- 초평면과 가장 근접한 원소간 거리 $\frac{1}{\|\boldsymbol{\alpha}\|}$를 최대화하기 위해서는 $\|\boldsymbol{\alpha}\|$ 크기 최소화가 요구되는데, 현실적으로 위의 그림에서와 같이 모든 데이터 인스턴스가 완벽하게 구분되기는 어려움
- 따라서, 오분류 가능성을 표현하기 위해 $y_i g(\mathbf{x}_i) \geq 1 - \xi_i$에서와 같은 에러텀($\xi_i$)을 도입하게 되고, ξ_i의 도입과 더불어 학습의 목표가 모든 데이터 인스턴스에 대해 $\|\boldsymbol{\alpha}\|$ 크기를 최소화($\min\|\boldsymbol{\alpha}\|$)하는 $\boldsymbol{\alpha}$를 찾는 것으로 바뀌게 됨
- 이 때, $\boldsymbol{\alpha}$는 모든 데이터 인스턴스에 대해 제시된 $y_i g(\mathbf{x}_i) \geq 1 - \xi_i$ 를 충족하고 $\xi_i \geq 0$, $\sum \xi_i \leq C$ (C는 상수)라는 제약조건도 충족해야 함
- 제약 조건 존재 시 라그랑주 승수법(Lagrange multiplier)을 최적화 목적으로 사용하면 학습목표는 다음의 함수 L을 최소화 하는 것으로 바뀌게 됨:

$$L(\boldsymbol{\alpha}, \boldsymbol{b}) \equiv \frac{1}{2}\|\boldsymbol{\alpha}\|^2 - \sum_{i=1}^{|\mathcal{D}|} b_i[y_i \boldsymbol{\alpha}^{\mathrm{T}}\mathrm{x}_i - 1]$$

- 위 함수에서 다음의 라그랑주 듀얼 객체함수(Lagrangian dual objective function) L_D획득이 가능하며 L_D 계산을 통해 특정화하려는 초평면 탐색 가능:

$$L_D = \sum_{i=1}^{|\mathcal{D}|} b_i - \frac{1}{2}\sum_{k,j}^{|\mathcal{D}|} b_k b_j y_k y_j \mathbf{x}_j^{\mathrm{T}}\mathbf{x}_k$$

ㅇ 커널함수(Kernel function)

- 서포트벡터머신은 선형모델로서의 한계가 존재하므로, 선형 모델로서의 표현력 확대를 위해서는 특성 공간을 확장하는 방법을 선택할 수 있음
- 서포트벡터머신에 해당 방법을 접목할 경우 함수 L_D은 다음과 같이 변환하게 되는데, 이때에 특성 공간을 확장하는데 사용되는 함수를 커널함수(kernel function)라고 함:

$$L_D = \sum_{i=1}^{|\mathcal{D}|} b_i - \frac{1}{2}\sum_{k,j}^{|\mathcal{D}|} b_k b_j y_k y_j \langle h(\mathbf{x}_k), h(\mathbf{x}_j)\rangle$$

ㅇ 다양한 커널함수를 사용하여 특성 공간의 확장이 가능하나 $h(\mathbf{x}_k)$를 실제 계산하는데 어려움이 있을 수 있으며, 이 경우 서포트벡터머신은 커널트릭(kernel trick)이라는 방법을 사용하게 됨

- 커널트릭은 각각 데이터 인스턴스를 새로운 특성 공간에서 확장하지 않고, $K(\mathbf{x},\mathbf{x}') = \langle h(\mathbf{x}_k), h(\mathbf{x}_j)\rangle$를 만족하는 커널함수를 사용하여 벡터간 내적을 계산함으로써 특성 공간에서의 계산 효과를 대체할 수 있음
- 대표적인 커널함수는 예는 다음과 같음:
  a. d차 다항 커널(polynomial kernel): $K(\mathbf{x},\mathbf{x}') = (1+\langle \mathbf{x},\mathbf{x}'\rangle)^d$
  b. 레이디얼 베이시스(Radial basis): $K(\mathbf{x},\mathbf{x}') = \exp(-\|\mathbf{x}-\mathbf{x}'\|^2/c)$
  c. 신경망(Neural network): $K(\mathbf{x},\mathbf{x}') = \tanh(\kappa_1\langle \mathbf{x},\mathbf{x}'\rangle + \kappa_2)$

□ 나이브 베이즈 분류기(Naïve Bayes Classifier)

ㅇ 나이브 베이즈 분류기의 토대가 되는 베이지안 이론(Bayes theorem)은 주어진 입력 데이터($\mathcal{D}$)에 대해 가설(h)의 확률을 계산:

$$P(h|\mathcal{D}) = \frac{P(\mathcal{D}|h)P(h)}{P(\mathcal{D})}$$

- $P(h)$는 사전확률(prior probability), $P(h|\mathcal{D})$는 사후확률(posterior probability), $P(\mathcal{D}|h)$는 우도(likelihood)에서,
- 사전확률 $P(h)$는 데이터 관찰 전 가설(h)에 주어진 신뢰성의 정도로 해석 할 수 있으며, 사후확률 $P(\mathcal{D}|h)$는 데이터 관찰을 통해 가설(h)에 부여되는 가중치를 의미
- 이에 따라, 사후확률 $P(h|\mathcal{D})$는 데이터 관찰 결과 가설(h)에 부여되는 확신이나 신뢰도로 표현된다고 해석이 가능함

- 여러 개의 가설 집합(H)이 존재하고 사후확률 값이 가장 큰 가설을 선택하는 것이 가장 합리적인 결정일 때 이를 최대사후확률(maximum a posteriori, MAP)이라고 정의함:

$$h_{MAP} \equiv \underset{h \in H}{\operatorname{argmax}} P(h|\mathcal{D}) = \underset{h \in H}{\operatorname{argmax}} \frac{P(\mathcal{D}|h)P(h)}{P(\mathcal{D})}$$

- 각각의 데이터가 d개의 특성으로 표현되는 데이터($\langle a_1, \ldots, a_d \rangle$)인 경우 해당 데이터가 포함될 수 있는 카테고리 집합을 V로 표현하면,
- 결정이 가장 타당한 카테고리는 v_{MAP}, 즉 최대사후확률 값을 최대화 하는 카테고리로 정의되어야 함:

$$v_{MAP} \equiv \underset{v_j \in V}{\operatorname{argmax}} P(v_j|a_1, \ldots, a_d) = \underset{v_j \in V}{\operatorname{argmax}} \frac{P(a_1, \ldots, a_d|v_j)P(v_j)}{P(a_1, \ldots, a_d)}$$

- 일반적으로 결합확률분포(joint probability) ($P(a_1, \ldots, a_d|v_j)$)의 계산을 위해 많은 양의 데이터가 요구되나 나이브 베이즈 분류기는 목표 값(target value)이 주어질 경우 특성 값을 서로 조건부 독립이라 가정하고 결합확률을 간단히 계산:

$$P(a_1, \ldots, a_d|v_j) = \prod_i P(a_i|v_j)$$

- 타깃 카테고리는 다음 공식을 통해 결정됨:

$$v_{NB} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j) \prod_i P(a_i|v_j)$$

□ 학습 모델 평가 및 개선 방법

○ 검증집합(Validation set) 접근법

- 최고 차수가 d인 다항 회귀를 사용하여 주어진 입력에 대한 실수 값을 예측하기 위해서는 서로 다른 d값에 의한 다항식 생성 후 훈련 데이터를 바탕으로 다항식 계수를 추정해야 함
- 검증집합(validation set) 접근법은 우선 가용된 훈련 데이터를 무작위로 훈련 집합(training set)과 검증집합으로 구분하고, 훈련 집합을 통해 계수를 예측, 검증집합을 활용하여 서로 다른 d값에 대한 다항식 모델의 회귀 성능을 측정

○ 검증집합(Validation set)과 과적합(Overfitting)

- 기계학습 시 훈련 집합(training set)과 테스트집합(test set)은 서로 동일한 분포로 독립적인 샘플링이 되었다고 가정하나,

- 동일한 데이터 집합을 이용하는 경우를 제외하고는 관측 오류 등으로 인해 훈련 집단 목표 분포에 특이 분포가 혼재될 가능성이 있음
- 이에 따라, 가능한 모든 가설을 포함한 가설공간(H)에 속한 가설 h와 h'에 대해 훈련 집합에서는 h의 성능이 h' 보다 뛰어남에도 테스트집합에서는 h의 성능이 h' 보다 저조한 현상(과적합(overfitting))이 발생할 수 있음
- 검증집합은 과적합 문제를 해결하는데에도 활용이 가능한데,
- 훈련집합을 통해 모델을 훈련하는 과정에서 지속적으로 검증집합의 성능을 측정하면서, 검증집단의 성능은 약화되고 훈련 집단 성능이 계속해서 개선되는 시점에서 훈련을 중단할 수 있음

ㅇ 교차검증(Cross-validation)

- 교차검증은 학습 과정 시 활용 가능한 검증기법의 하나로,
- 기계학습 훈련과정에서 훈련데이터를 토대로 학습 커브를 생성하나, 태스크 해결 과정에서 보다 중요한 부분은 훈련과정에서 포함되지 않은 테스트데이터에 대한 훈련된 모델의 성능임
- 테스트 성능 측정 시 테스트데이터의 양이 충분할 경우에는 문제가 되지 않지만, 충분한 양의 데이터 확보가 용이하지 않을 경우 부족한 데이터 이슈를 해소하기 위해 교차검증(Cross-validation)을 활용

ㅇ 리브원아웃 교차검증(Leave-one-out cross validation, LOOCV)

- 리브원아웃 교차검증은 주어진 n개의 데이터집합에서 검증목적으로 1개 데이터를 사용하고 나머지 $n-1$개 데이터는 훈련에 사용하는 검증 기법임
- 예를 들어, 성능 측도로 평균제곱근에러(MSE)를 사용할 경우 (x_1, y_1)을 제외한 데이터를 훈련에 사용하면 $MSE_1 = (y_1 - \hat{y}_1)^2$로 정의할 수 있음
- 유사 방식으로 주어진 데이터에 대하여 (x_i, y_i)를 훈련에서 제외하였을 때 $MSE_i = (y_i - \hat{y}_i)^2$를 정의하면 이에 대한 성능은 다음과 같이 나타낼 수 있음:

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{MSE}_i$$

ㅇ k배 교차검증(k-fold cross-validation, k-fold CV)

- k배 교차검증은 주어진 훈련 데이터를 k개 수 그룹으로 분할하고 1개 그룹은 검증집합으로, 나머지 k-1개의 그룹은 훈련 집합으로 사용하여 모델을 훈련

- 검증그룹에 대해 훈련된 모델의 평균제곱근에러(MSE) 측정 시 총 k개의 평균제곱근에러가 산출되며 k배 교차검증 성능은 다음 식으로 측정됨:

$$\mathrm{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k}\mathrm{MSE}_i$$

ㅇ 부트스트랩(Bootstrap)

- 부트스트랩은 추정의 정확성을 제고하기 위해 훈련집합($\mathcal{D}$)로부터 n개 데이터를 무작위로 복원·선택하는 과정을 B번 반복함
- 이러한 과정에서 생성된 B개의 데이터 집합을 사용하여 통계치 θ는 다음의 방식으로 추정됨:

$$\hat{\theta}^{*(\cdot)} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{*(b)}$$

- $\hat{\theta}^{*(b)}$는 부트스트랩 샘플(bootstrap sample) (b)을 통해 확보된 추정치를 의미

ㅇ 배깅(Bagging; Bootstrap aggregation)

- 배깅(bagging)은 훈련 집합($\mathcal{D}$)로부터 n'개 샘플($n' < |\mathcal{D}|$)을 복원추출하는 과정을 통해 B개 훈련 집합을 새롭게 생성하고 각 훈련 집합에 대해 모델을 학습함
- 최종적으로 B개 모델의 예측 성능을 통합하고 다음과 같은 방식으로 최종 결과를 획득:

$$\hat{f}_{\mathrm{bag}}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{*b}(x)$$

- 여기서 $\hat{f}^{*b}(x)$는 b번째 부트스트랩 데이터 집합에서 획득된 모델의 예측 성능을 의미함

ㅇ 부스팅(Boosting)

- 부스팅은 여러 개의 분류기(classifier) 학습 후 개개의 분류기에서 도출된 예측결과를 종합하여 개별 분류기 보다 우수한 성능을 달성하는 것을 목표함
- 현재의 분류기 관점에서 가장 많은 정보량을 지닌 부분집합을 새로운 분류기를 이용하여 훈련
- 전체 훈련데이터(training data) ($\mathcal{D}$)에서 $n_1 < |\mathcal{D}|$개 훈련데이터를 무작위로 비복원추출 후 $\mathcal{D}_1$을 활용하여 첫 번째 분류기($\mathcal{C}_1$)를 훈련시킴

- 다음으로 훈련 집합(training set) ($\mathcal{D}_2$)는 1/2 확률 사건 A가 발생 시 $\mathcal{D}-\mathcal{D}_1$에서 샘플을 하나씩 취한 후 $\mathcal{C}_1$가 분류에 실패한 샘플을 $\mathcal{D}_2$에 추가하고,
- A^c가 발생할 경우 다시 $\mathcal{C}_1$가 분류에 성공한 샘플을 추가하는 방식으로 $\mathcal{D}_2$를 구성하고 이렇게 구성된 $\mathcal{D}_2$를 이용하여 새로운 분류기 $\mathcal{C}_2$를 훈련시킴
- $\mathcal{D}_1$,$\mathcal{D}_2$에 포함되지 않은 잔여 샘플들에 대하여 $\mathcal{C}_1$, $\mathcal{C}_2$의 예측결과가 어긋난 샘플을 모아 $\mathcal{D}_3$를 구성한 후 $\mathcal{D}_3$를 통해 새로운 분류기 $\mathcal{C}_3$를 훈련함

나. 비지도학습(Unsupervised learning)

1) 비지도학습의 개념

□ 비지도학습(unsupervised learning)은 특성을 나타내는 벡터 $\mathbf{x}$ 자체만을 제공한다는 측면에서 특성 벡터 $\mathbf{x}$와 대응되는 클래스 레이블/실수 값이 제공되는 지도학습(supervised learning)과 차이가 있음

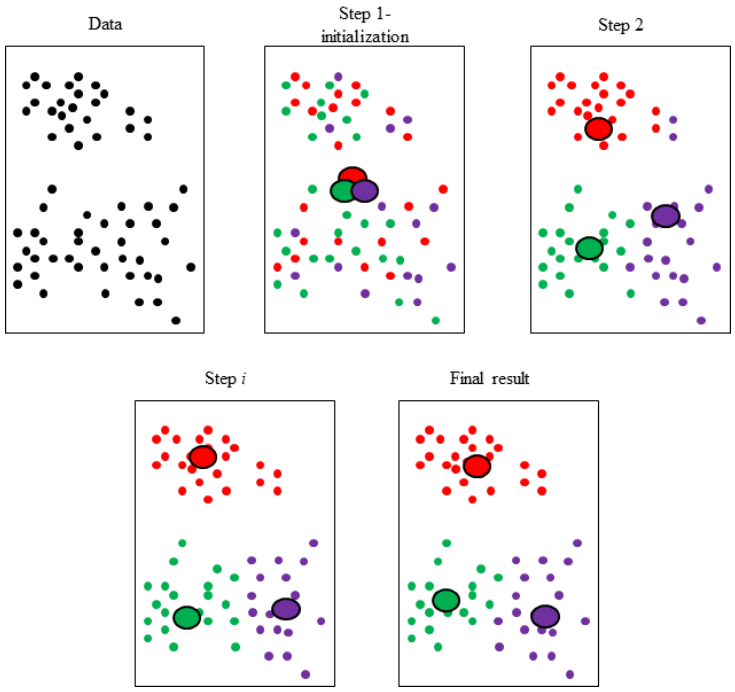
- ㅇ 학습 목표는 데이터에 내재된 구조적 특성을 파악하는 것으로, 비슷한 데이터를 무리짓는 군집화(클러스터링, clustering)나 문제 공간 축소 등에서 활용됨
- ㅇ 비지도학습의 경우 데이터 구조 학습을 위해 데이터 그룹의 유사도 측정이 가능한 측도가 요구됨:
  - 오차제곱합기준(Sum-of-Squared-Error criterion): i번째 그룹 D_i의 평균벡터(mean vector) $\mathbf{m}_i = \frac{1}{n_i}\sum_{\mathbf{x}\in D_i}\mathbf{x}$에 대해 다음으로 정의됨(여기서 c는 클러스터의 수):

$$J_e = \sum_{i=1}^{c}\sum_{\mathbf{x}\in D_i}\|\mathbf{x}-\mathbf{m}_i\|^2$$

  - 산점도 행렬(Scatter matrix): $\boldsymbol{S}_T = \boldsymbol{S}_W + \boldsymbol{S}_B$ 로 정의되는데, $\boldsymbol{S}_W$는 클러스터 내 산점도 행렬(Within-cluster scatter matrix)로 $\boldsymbol{S}_W = \sum_i^c \boldsymbol{S}_i$, $\boldsymbol{S}_i = \sum_{\mathbf{x}\in D_i}(\mathbf{x}-\mathbf{m}_i)(\mathbf{x}-\mathbf{m}_i)^t$로 표현되고, 클러스터 간 산점도 행렬(Between-cluster scatter matrix)을 의미하는 $\boldsymbol{S}_B$는 $\boldsymbol{S}_B = \sum_{i=1}^c |D_i|(\mathbf{m}-\mathbf{m}_i)(\mathbf{m}-\mathbf{m}_i)^t$, $\mathrm{m} = \frac{1}{|D|}\sum_{i=1}^c |D_i|\mathbf{m}_i$로 표현됨

2) 비지도학습 기반 알고리즘

□ k-평균 클러스터링(k-means clustering)



[그림 2-4] k-평균 클러스터링 동작 예시

- k-평균 클러스터링은 목표된 클러스터의 개수(k)에 도달할 때까지, 클러스터 분산 합(total within cluster variance)이 최소화 되는 방향으로 클러스터 중심을 계속 이동시키며 주어진 데이터를 군집화 하는 방법
  - 무작위로 클러스터를 부여하고 기댓값 최대화(Expectation-Maximization) 방법을 사용하여 목표 측도를 최소화하는 그룹을 탐색함([그림 2-4] 표기된 원은 클러스터 중심을 나타냄)
- k-평균 클러스터링은 $\boldsymbol{S}_W$의 최소화를 목표하는데, 주어진 $\boldsymbol{S}_W$ 수식을 원안 사용 시 규모가 큰 클러스터는 최소화에 어려움이 있을 수 있으므로 $\boldsymbol{S}_W = \sum_i^c \frac{1}{|D_i|}\boldsymbol{S}_i$, $\boldsymbol{S}_i = \sum_{\mathbf{x}\in D_i}(\mathbf{x}-\mathbf{m}_i)(\mathbf{x}-\mathbf{m}_i)^t$로 수정 후 다음 수식을 충족하는 클러스터로 결정:

$$C_1, \dots, C_k = \underset{C_1,\dots,C_k}{\operatorname{argmin}} \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{\mathbf{x}\in C_k} (\mathbf{x}-\mathbf{m}_k)(\mathbf{x}-\mathbf{m}_k)^t$$

- 기댓값-최대화(Expectation-Maximization): m개 데이터 $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$이 관찰되고 미관찰데이터 $\mathbf{Z} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m\}$ 존재 시 전체 데이터(Y)는 $\mathbf{Y} = \mathbf{X} \cup \mathbf{Z}$로 표현할 수 있음
- $\mathbf{Z}$를 미관찰 인자 $\boldsymbol{\theta}$의 확률분포를 따르는 확률변수(random variable)이라하면 $\boldsymbol{\theta}$의 현재 값은 h, 개정 값은 h'로 표기하면서 반복에 의해 주어진 조건을 충족하는 h를 찾는 것이 가능함
- 기댓값-최대화에서는 $\mathrm{E}[\ln P(Y|h')]$를 최대화하는 개정 값(h')을 찾음으로써 최대우도(maximum likelihood) 가설(h')을 찾게 됨
- 전체 데이터 (Y)는 미관찰 인자($\boldsymbol{\theta}$)를 통해 결정되어 Y의 분포를 정확히 알 수 없기 때문에 기댓값-최대화의 경우 $\boldsymbol{\theta}$ 에 대한 현재 추정 값(h)의 데이터 생성을 가정하고 데이터 생성 분포를 찾게 되는데, 이 때 Q함수를 다음과 같이 정의함:
- $\mathrm{Q}(h'|h) = E[\ln p(Y|h')\,|h, \mathbf{X}]$
- Step 1: 기대(expectation) 단계에서는 현재 가설(h)과 관찰된 데이터(X)를 활용하여 $\mathrm{Q}(h'|h)$ 계산 후 전체 데이터(Y)의 분포 추정
- Step 2: 최대화(maximization) 단계에서는 Q를 최대화시키는 개정 값(h')으로 가설(h)을 대체 함:

$$h \leftarrow \underset{h'}{\operatorname{argmax}}\, \mathrm{Q}(h'|h)$$

- 종료 조건 충족 시 까지 기대단계와 최대화 단계를 반복하며 가설을 갱신

□ 계층적 클러스터링(Hierarchical clustering)

o k-평균 클러스터링은 클러스터링 개수(k)를 입력하게 되어 있는데, 이는 클러스터 개수를 사전에 알고 있어야 한다는 의미로도 해석이 가능

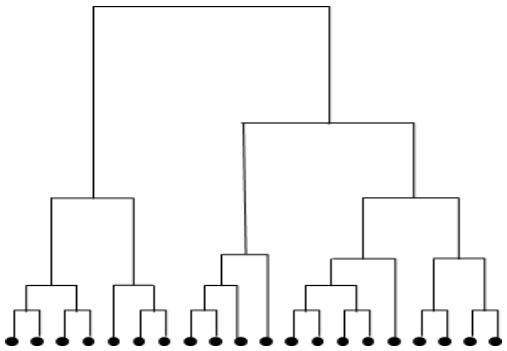
- 일반적인 경우 클러스터 개수의 사전 파악은 용이하지 않음

o 계층적 클러스터링의 경우 데이터 인스턴스 간 거리를 비교 측정하여 가장 인접한 데이터 인스턴스 쌍을 하나의 그룹으로 군집화함

o 이후 새롭게 생성된 데이터 그룹과의 거리를 비교 측정하여 가장 인접한 그룹을 다시 하나로 묶는 과정을 모든 데이터 그룹이 하나의 그룹으로 통합 시 까지 계속 반복함

- 개별 데이터 인스턴스 간 거리는 비교적 쉽게 계산이 가능하나, 그룹 간 혹은 그룹-개별 데이터 인스턴스 간의 거리 비교는 수행이 어려움

- 따라서, 그룹 혹은 그룹과 인스턴스 간 거리 비교 시 다음에 제시된 측도를 사용함:
  - $d_{\min}(\mathcal{D}_i, \mathcal{D}_j) = \min_{\mathbf{x}\in\mathcal{D}_i, \mathbf{x}'\in\mathcal{D}_j} \|\mathbf{x}-\mathbf{x}'\|$
  - $d_{\max}(\mathcal{D}_i, \mathcal{D}_j) = \max_{\mathbf{x}\in\mathcal{D}_i, \mathbf{x}'\in\mathcal{D}_j} \|\mathbf{x}-\mathbf{x}'\|$
  - $d_{\text{avg}}(\mathcal{D}_i, \mathcal{D}_j) = \frac{1}{|\mathcal{D}_i||\mathcal{D}_j|}\sum_{\mathbf{x}\in\mathcal{D}_i}\sum_{\mathbf{x}'\in\mathcal{D}_j}\|\mathbf{x}-\mathbf{x}'\|$
  - $d_{\text{mean}}(\mathcal{D}_i, \mathcal{D}_j) = \|\mathbf{m}_i - \mathbf{m}_j\|$



[그림 2-5] 계층적 클러스터링으로부터 획득한 dendrogram

- 제시된 4개의 거리 측도는 데이터 그룹 $\mathcal{D}_i$와 $\mathcal{D}_j$ 간의 거리 비교에 이용되는데,
- $d_{\min}$은 $\mathcal{D}_i$에 속한 멤버 인스턴스 $\mathbf{x}$와 $\mathcal{D}_j$에 속한 멤버 인스턴스 $\mathbf{x}'$ 간에 가능한 모든 조합의 계산된 거리 중 최소 거리로 설정하며, $d_{\max}$는 가능한 모든 조합의 계산된 거리 중 최대 거리로 설정함
- $d_{\text{avg}}(\mathcal{D}_i, \mathcal{D}_j)$에서는 각기 속한 멤버 인스턴스 간 모든 조합의 계산된 거리의 평균으로 설정하며 $d_{\text{mean}}(\mathcal{D}_i, \mathcal{D}_j)$은 $\mathcal{D}_i$ 중심점 $\mathbf{m}_i$과 $\mathcal{D}_j$ 중심점 $\mathbf{m}_j$ 간의 거리로 설정

□ 주성분 분석(Principal Component Analysis, PCA)

○ 주성분 분석은 실제 입력된 데이터보다 낮은 차원으로 공간을 표현할 때 사용*되는 방법으로 고유벡터(eigenvector)와 고유 값(eigenvalue)을 사용함

* 낮은 차원의 표현 공간 학습

- 정사각행렬(square matrix) A에 관해 $\mathbf{A}\boldsymbol{v} = \lambda\boldsymbol{v}$를 충족하는 0벡터가 아닌 벡터 $\boldsymbol{v}$를 고유벡터로, scalar λ는 고유 값으로 정의됨

- 차원 축소: 데이터가 n차원의 문제 공간으로 표현되고, 이를 $l < n$에 해당하는 l차원으로 축소하고자 할 경우 $\mathbf{x} \in \mathbb{R}^n$에 해당하는 $\mathbf{x}$를 $\mathbf{c} \in \mathbb{R}^l$에 해당하는 $\mathbf{c}$로 표현하는 것을 고려할 수 있으며,

- 이 때, 인코딩 함수 $f, f(\mathbf{x}) = \mathbf{c}$ 와 디코딩 함수 g $\left(g\text{의 조건 } g(f(\mathbf{x})) \approx \mathbf{c}\right)$ 를 도입하게 되면 디코딩 함수 $g(\mathbf{c}) = \boldsymbol{D}\boldsymbol{c}$ $(\boldsymbol{D} \in \mathbb{R}^{n \times l})$의 정의가 가능함

- $\boldsymbol{D}$ 가 쉽게 유도되도록 $\boldsymbol{D}$의 열이 서로 직교(orthogonal)하도록 제약하고 당초 벡터 $\mathbf{x}$와 $g(\mathbf{c})$의 차이를 최소화시키는 g를 찾으면 다음과 같이 표현됨:

$$\boldsymbol{c}^* = \underset{\boldsymbol{c}}{\operatorname{argmin}} \|\mathbf{x} - g(\mathbf{c})\|_2$$

- 여기서 L^2 노름(norm)*에 제곱을 취하게 되면 최소화가 되어야 할 함수는 $(\mathbf{x} - g(\mathbf{c}))^{\mathrm{T}}(\mathbf{x} - g(\mathbf{c})) = \mathbf{x}^{\mathrm{T}}\mathbf{x} - 2\mathbf{x}^{T}g(\mathbf{c}) + g(\mathbf{c})^{\mathrm{T}}g(\mathbf{c})$로 단순화됨

* 유클리디안노름이라고도 표현되며, n차원의 유클리디안 노름 공간상에 원점에서 해당벡터까지 직선거리를 의미

- 이후 $g(\mathbf{c})$를 포함하는 부분만을 추출하면 다음과 같이 표현됨:

$$\begin{aligned}
\boldsymbol{c}^* &= \underset{\mathbf{c}}{\operatorname{argmin}} -2\mathbf{x}^{\mathrm{T}}g(\mathbf{c}) + g(\mathbf{c})^{\mathrm{T}}g(\mathbf{c}) \\
&= \underset{\mathbf{c}}{\operatorname{argmin}} -2\mathbf{x}^{\mathrm{T}}\boldsymbol{D}\mathbf{c} + \mathbf{c}^{\mathrm{T}}\boldsymbol{D}^{\mathrm{T}}\boldsymbol{D}\mathbf{c} \\
&= \underset{\mathbf{c}}{\operatorname{argmin}} -2\mathbf{x}^{\mathrm{T}}\boldsymbol{D}\mathbf{c} + \mathbf{c}^{\mathrm{T}}\boldsymbol{I}_l\mathbf{c} \\
&= \underset{\mathbf{c}}{\operatorname{argmin}} -2\mathbf{x}^{\mathrm{T}}\boldsymbol{D}\mathbf{c} + \mathbf{c}^{\mathrm{T}}\mathbf{c}
\end{aligned}$$

- 여기서 $\mathbf{c}$는 $\mathbf{c} = \boldsymbol{D}^{\mathrm{T}}\mathbf{x}$로 유도가 가능하기 때문에 $f(\mathbf{x}) = \boldsymbol{D}^{\mathrm{T}}\mathbf{x}$, $r(\mathbf{x}) = g(f(\mathbf{x})) = \boldsymbol{D}\boldsymbol{D}^{\mathrm{T}}\mathbf{x}$로 설명할 수 있으며, 다음의 과정을 거쳐 D가 유도될 수 있음:

$$\boldsymbol{D}^* = \underset{\boldsymbol{D}}{\operatorname{argmin}} \sqrt{\sum_{i,j}\left(\mathbf{x}_j^{(i)} - r(\mathbf{x}^{(i)})_j\right)^2} \quad (\boldsymbol{D}^{\mathrm{T}}\boldsymbol{D} = \boldsymbol{I}_l)$$

- l=1로 간주하고 문제를 풀 경우 D가 아닌 단일 벡터 $\boldsymbol{d}$, $\boldsymbol{d}^* = \underset{\boldsymbol{d}}{\operatorname{argmin}} \sum_i \|\mathbf{x}^{(i)} - \mathbf{x}^{(i)\mathrm{T}}\boldsymbol{d}\boldsymbol{d}^{\mathrm{T}}\|_2^2$ $(\|\boldsymbol{d}\|_2 = 1)$를 찾는 문제로 변경 가능

- 이 때 주어진 m개 데이터 인스턴스의 반영이 가능한 행렬 X ($\mathbf{X} \in \mathbb{R}^{m \times n}$, $\mathbf{X}_{i,:} = \mathbf{x}^{(i)\mathrm{T}}$)를 설정할 경우 다음 식을 토대로 원하는 행렬d의 확보가 가능함:

$$\boldsymbol{d}^* = \underset{\boldsymbol{d}}{\operatorname{argmin}} \|\mathbf{X} - \mathbf{X}\boldsymbol{d}\boldsymbol{d}^{\mathrm{T}}\|_F^2 \quad (\boldsymbol{d}^{\mathrm{T}}\boldsymbol{d} = 1)$$

- 이는 다음과 같이 풀어 쓸 수 있으며:

$$\boldsymbol{d}^* = \underset{\boldsymbol{d}}{\operatorname{argmin}} -2\mathrm{Tr}(\mathbf{X}^\mathrm{T}\mathbf{X}\boldsymbol{d}\boldsymbol{d}^\mathrm{T}) + \mathrm{Tr}(\mathbf{X}^\mathrm{T}\mathbf{X}\boldsymbol{d}\boldsymbol{d}^\mathrm{T}\boldsymbol{d}\boldsymbol{d}^\mathrm{T}) \quad (\boldsymbol{d}\boldsymbol{d}^\mathrm{T} = 1)$$

- 고유값 분해(eigen-decomposition)을 적용하여 고유값 크기에 상응하는 $\mathbf{X}^\mathrm{T}\mathbf{X}$ 고유벡터를 선택하면 차원의 축소가 가능함

ㅇ 주성분 분석은 데이터가 가장 크게 변화하는 축으로 해당 데이터를 프로젝션하는 방법으로 다음과 같은 해석이 가능

- 유도 과정을 통해 도출된 상위 k개 고유 값에 상응하는 고유벡터들의 경우 데이터 변화의 정도가 큰 축에서 작아지는 축으로 정렬되는 축의 집합으로 해석 가능
- 또는 주성분을 데이터와의 거리 합이 가장 작은 축을 의미하는 것으로 해석 가능

□ 독립성분분석(Independent Component Analysis, ICA)

ㅇ 주성분 분석은 표현 공간(특성 공간) 상에 주어진 데이터를 가장 표현하는 방향을 찾고자하는 접근법인 반면 독립성분분석은 가장 독립적인 방향을 찾고자하는 비지도 학습 방법임

- 독립성분분석에서는 d개의 독립된 소스에서 생성된 데이터 $\mathbf{x} \in \mathbb{R}^d$ 존재 시, 관찰된 데이터 $\mathbf{s} = \mathbf{A}\mathbf{x}$는 혼합행렬A(mixing matrix $\mathbf{A}$)로 인해 변화가 발생한 상태로 가정함
- 가령 $\{\mathbf{s}^{(i)}: i = 1, \ldots, m\}$와 같은 데이터 관찰 시, 독립성분분석에서는 데이터 생성소스 $x^{(i)}$를 복원하고자 할 때,
- $\mathbf{W} = \boldsymbol{A}^{-1}$를 혼합되지 않은 행렬(unmixing matrix)이라고 간주할 경우 $x^{(i)} = \mathbf{W}s^{(i)}$의 계산을 통해 관찰 데이터에서 원래의 데이터 복원이 가능함
- 각 소스로 표현되는 x_i 데이터를 밀도(density) p_x에 의한 생산이라고 가정하게 되면 소스 x의 결합분포(joint distribution)는 다음 식과 같이 쓸 수 있음:

$$p(\mathbf{x}) = \prod_{i=1}^{d} p_x(x_i)$$

- 관찰된 데이터(s)에 대한 복원 데이터를 y라고 표현하게 되면:

$$p_\mathbf{y}(\mathbf{y}) = \frac{p_\mathbf{x}(\mathbf{s})}{|\mathbf{J}|}$$

- 로 나타낼 수 있으며, 여기서

$$\mathbf{J} = \begin{pmatrix} \frac{\partial y_1}{\partial s_1} & \cdots & \frac{\partial y_d}{\partial s_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_1}{\partial s_d} & \cdots & \frac{\partial y_d}{\partial s_d} \end{pmatrix}, |\mathbf{J}| = \left| |\mathbf{W}| \prod_{i=1}^{d} \frac{\partial y_i}{\partial s_i} \right|$$

- 데이터 복원 과정을 소스 신호(source signal)의 선형 변환(linear transform)으로 해석 시 $\mathbf{y} = f[\mathbf{W}\mathbf{s} + \mathbf{w}_0]$ (f는 일반적으로 시그모이드 함수를 사용)로 표현 가능하며 $\mathbf{W}$와 $\mathbf{w}_0$를 구하기 위해서는 결합엔트로피(joint entropy)를 사용:

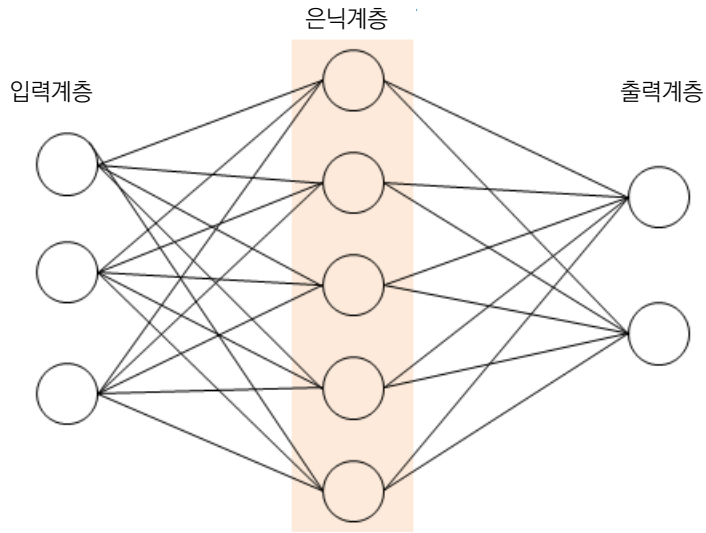
$$\boldsymbol{H}(\mathbf{y}) = -E[\ln p_{\mathbf{y}}(\mathbf{y})] = E[\ln|\mathbf{J}|] - E[\ln p_{\mathbf{s}}(\mathbf{s})]$$

다. 딥러닝(Deep Learning Network)

1) 딥러닝의 개념

□ 딥러닝 기술은 상당한 시간을 가지고 발전을 거듭해 온 인공신경망(Artifical Neural Network, ANN) 기술에 기반을 두고 있음

ㅇ 인공신경망 신경세포(Neuron) 간 조직적 연결로 구성된 인간 뇌 구조를 모방한 망(Netowrk)으로, [그림 2-6]에서 제시된 바와 같이 입력계층(input layer), 은닉계층(hidden layer), 출력계층(ouput layer)으로 구성됨



[그림 2-6] 인공신경망(Artificial Neural Network) 예시

□ 딥러닝 기술은 인공신경망의 은닉계층을 심층적으로 구조화하고 학습과정에서 개별 노드 간 연결된 엣지([그림 2-6]의 실선)에 할당된 가중치(weight) 탐색 과정을 의미

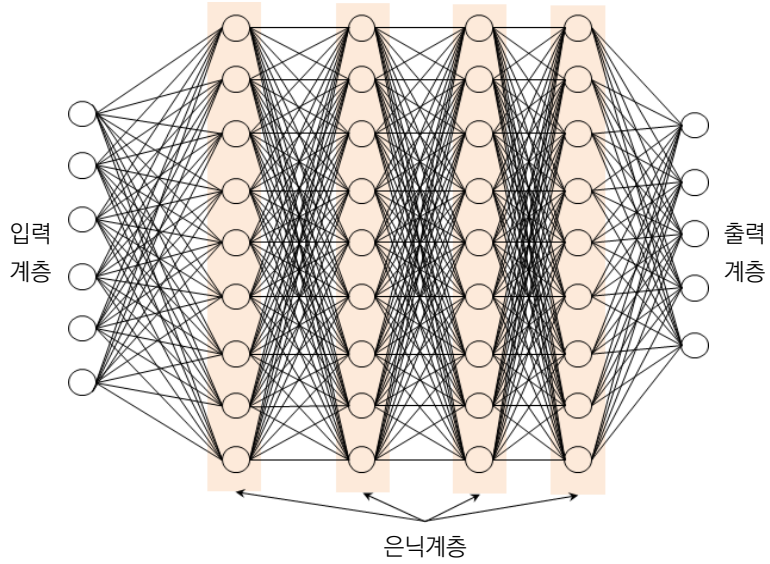
※ 주어진 데이터의 특징(특성)을 스스로 학습하는 비지도학습방법에 바탕을 둠

- ㅇ 인공신경망의 각 노드는 이 전 노드에서 받은 데이터를 처리한 후 다음 노드로 처리된 데이터를 전달하는 방식으로 작동함
- ㅇ 데이터의 노드 간 전달 시 해당 엣지의 가중치가 반영되어 처리되는데, 딥러닝 과정을 통해 엣지의 적절한 가중치를 학습하게 됨

2) 딥러닝 기반 알고리즘[6)]

□ 심층피드포워드망(Deep feedforward network)

- ㅇ 심층피드포워드망(Deep feedforward network), 피드포워드신경망(feedforward neural network), 다층퍼셉트론(multilayer perceptron)은 모두 상동의 개념임
- ㅇ 심층피드포워드망은 정보가 입력계층(input layer)에서 은닉계층(hidden layer)을 거쳐 출력계층(output layer)으로 전달되는 것으로 간주
  - 정보의 전달 과정에서 목표 함수 $f^*(\mathbf{x})$를 근사하는 것을 목표로, $\mathbf{y} = f(\mathbf{x};\boldsymbol{\theta})$의 매핑 관계 정의를 통해 가장 우수한 근사 결과를 담보하는 $\boldsymbol{\theta}$를 학습함



[그림 2-7] 심층신경망(Deep Neural Network) 예시

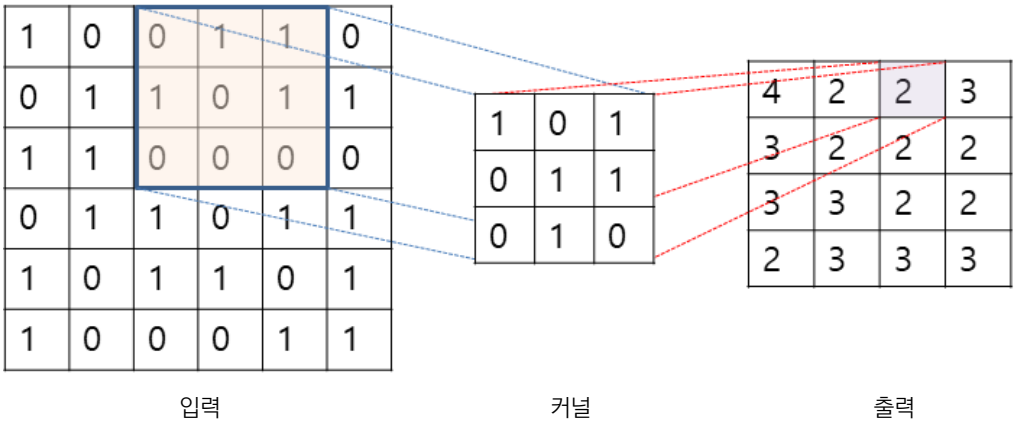
6) I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, The MIT Press, 2016. 내용을 인용함

- 피드포워드망은 여러 개 함수가 체인구조로 연결된 비순환 방향 그래프(acyclic directed graph)로 표현됨
  - 가령, $f^{(1)}, f^{(2)}, f^{(3)}$ 3개 함수가 체인으로 연결 시 $f(\mathbf{x}) = f^{(3)}\left(f^{(2)}\left(f^{(1)}(\mathbf{x})\right)\right)$로 표현되며 $f^{(1)}$, $f^{(2)}$는 각각 첫 번째, 두 번째 계층을 의미
  - 체인 길이는 "깊이(depth)", 동일 계층에 존재하는 노드의 수를 "폭(width)"이라고 하고 체인의 최종 레이어는 출력계층(output layer)에 해당함
  - 피드포워드망를 구성하는 각각의 노드는 입력 벡터를 스칼라값(scalar)으로 출력하는데, 사용되는 함수는 계산 용이성, 모델표현력을 감안하여 결정함
- 심층피드포워드망과 심층신경망의 경우 그레디언트(gradient)를 이용한 반복 계산에 의해 목표 함수를 개선시키는 방향으로 $f^*(\mathbf{x})$를 근사하게 됨
  ※ 노드 연산 시 비선형 함수를 활용하므로 선형모델의 최적화접근법 사용이 어려움
  - 딥러닝에서의 모델은 분포 $p(\mathbf{y}|\mathbf{x};\boldsymbol{\theta})$를 표현하여 최대우도(maximum likelihood)를 충족하는 파라미터 $\boldsymbol{\theta}$를 찾게 됨

□ 컨볼루션신경망(Convoluation Neural Network, CNN)

- 컨볼루션신경망은 이미지와 같이 격자형태의 토폴로지 구성을 보이는 데이터 처리 목적의 신경망 구조 네트워크를 지칭함
  - 최소 하나 이상의 계층에서 일반적인 행렬 곱셈이 아닌 컨볼루션(convolution) 연산이 사용됨
- 컨볼루션 오퍼레이션(convolution operation) 시에는 입력행렬(input matrix)과 커널행렬(kernel matrix)간 대응되는 셀의 값을 곱한 후 더한 결과가 출력행렬(output matrix)의 해당 셀 값이 됨:

$$s(t) = (x * w)(t)$$



[그림 2-8] 컨볼루션 오퍼레이션(Convolution operation) 예시

○ 컨볼루션 신경망에서는 함수(x)를 입력(input), 함수(w)를 커널(kernel)이라고 하며, 출력은 특징지도(feature map)라고 표현함

- 정수 t에 관하여 이산컨볼루션(discrete convolution)은 다음으로 표현:

$$s(t) = (x * w)(t) = \sum_{-\infty}^{\infty} x(a)w(t-a)$$

○ 2차원 데이터를 컨볼루션신경망에 적용할 경우 커널 K(kernel K)에 관한 컨볼루션 오퍼레이션은 다음으로 표현할 수 있음:

$$S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(m,n)K(i-m, j-n)$$
$$= (K * I)(i,j) = \sum_m \sum_n I(i-m, j-n)K(m,n)$$

- 보통의 신경망에서는 선형 변환(linear transformation) 시 행렬 연산을 사용하는데 이 때 모든 출력 유닛과 입력 유닛이 관련을 맺게 됨
- 반면에, 컨볼루션신경망의 경우 입력보다 작은 커널을 이용하여 sparse connectivity(연결성) 혹은 sparse interaction(상호작용)을 구현

□ 순환신경망(Recurrent Neural Network, RNN)

○ 순차적 데이터 처리에 특화된 순환신경망(RNN)은, 인자공유(parameter sharing) 아이디어를 활용하여 특정되지 않은 길이(length)의 시퀀스 처리를 가능하게 함:

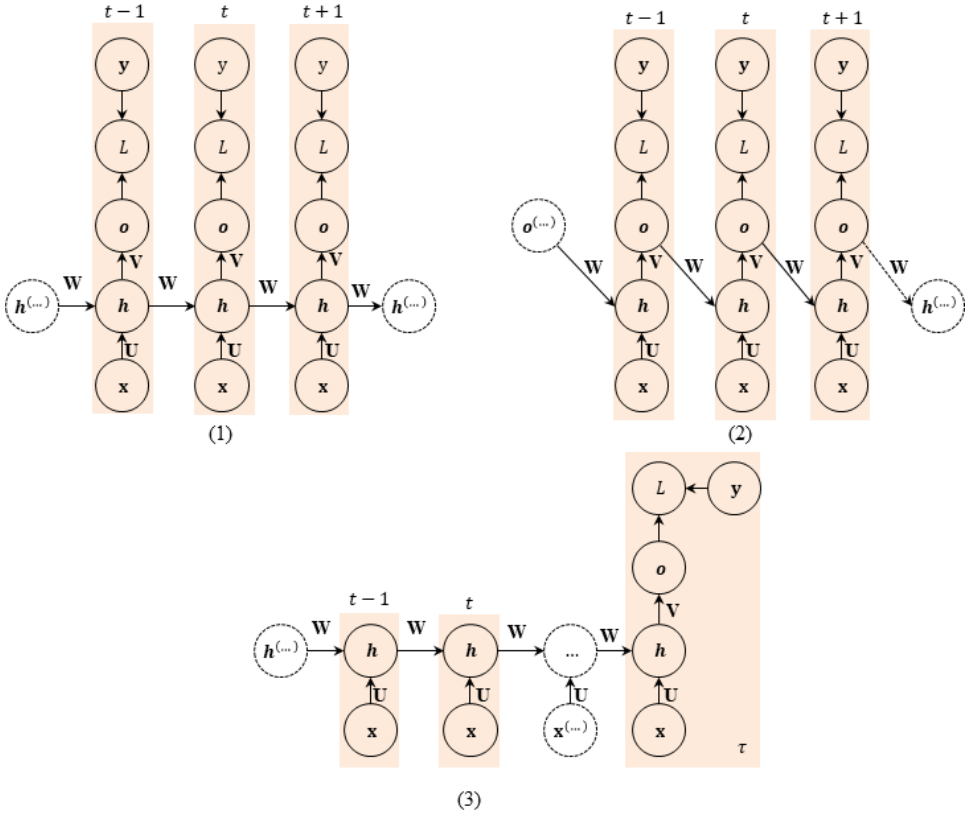
$$\boldsymbol{h}^{(t)} = f\left(\boldsymbol{h}^{(t-1)}, \mathbf{x}^{(t)}; \boldsymbol{\theta}\right)$$

- 위와 같은 식을 통해 은닉상태(hidden state)를 계산하게 되는데,
- 이 때, 인자 $\boldsymbol{\theta}$에 의해 동작이 규정되며, 현재 타임스텝(time step) (t)에 관한 입력($\mathbf{x}^{(t)}$)과 이전 타임스텝(t-1)에 관한 은닉상태($\boldsymbol{h}^{(t-1)}$)를 통해 현재 상태(state) ($\boldsymbol{h}^{(t)}$) 값을 결정함

※ 곱하기 연상으로 구성된 신경망의 설계 구조상 은닉 값을 추정하는 학습과정에서 시간을 거슬러 오를수록 그레디언트 소실(vanishing gradient) 문제 발생

ㅇ 순환신경망의 주요 유형은 다음과 같이 정리가 가능

- [그림 2-9]와 같이 (1) 매 타임스텝별 출력이 생성되고 은닉유닛(hidden unit) 간 순환연결(recurrent connection) 존재; (2) 매 타입 스텝별 출력이 생성되고 한 타임스텝 출력과 다음 타임스텝 은닉유닛 간 순환연결 존재; (3) 출력은 하나만 생성되며 은닉유닛 간 순환연결이 존재



[그림 2-9] 순환신경망(RNN) 유형

□ 장단기 메모리(Long Short-Term Memory, LSTM)

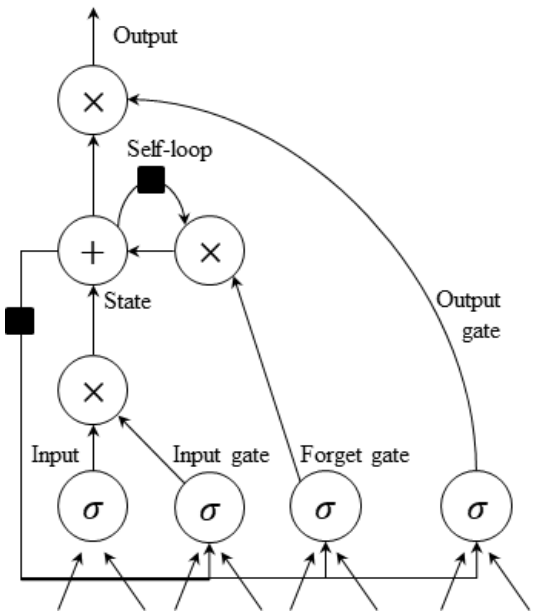
ㅇ 장단기 메모리는 게이트화된 순환 유닛(gated recurrent unit)을 활용하여 그레디언트 소실 문제를 해결한 발전된 형태의 순환 신경망(게이트화 순환신경망(gated RNN)의 일종)임

- 게이트화 순환 신경망은 미분값이 너무 커지지도, 소멸하지도 않는 경로(path)를 사용하고 연결 가중치는 매 타임스텝별 변경될 수 있도록 구현됨
- 그레디언트 지속 전달이 가능한 셀프루프(self-loop)를 도입하고 컨텍스트에 의해 셀프루프 가중치가 제어되도록 설정하여 성능을 향상시킴

※ 장단기 메모리 셀프루프를 보유하고 있는 장단기 메모리 셀(LSTM cell)을 사용

- 셀은 일반적인 순환망(recurrent network)와 동일한 입·출력을 지니고 있으나 [그림 2-10]에서와 같이 정보 흐름 제어를 위한 게이트 유닛*(gate unit)이 추가 존재

* 입력게이트(input gate), 망각게이트(forget gate) 출력게이트(output gate)



[그림 2-10] 장단기 메모리 셀(cell)의 블록 다이어그램

- 망각게이트(forget gate) ($f_i^{(t)}$) 제어를 받는 상태유닛(state unit) ($s_i^{(t)}$)은 장단기 메모리 셀의 가장 중요한 컴포넌트임(i는 i번째 셀, t는 타임스텝):

$$f_i^{(t)} = \sigma\left(b_i^f + \sum_j U_{i,j}^f x_j^{(t)} + \sum_j W_{i,j}^f h_j^{(t-1)} \right)$$

- 여기서 $\mathbf{x}^{(t)}$, $\boldsymbol{h}^{(t)}$는 각각 현재 입력벡터, 은닉계층 벡터로 $\mathbf{x}^{(t)}$, $\boldsymbol{h}^{(t)}$는 모든 장단기 메모리 셀의 출력을 포함하고 있음
- 망각게이트의 바이어스(bias), 입력가중치(input weight), 순환가중치(recurrent weight)는 각각 $\mathbf{b}^f, \mathbf{U}^f, \mathbf{W}^f$으로 표현되며, σ는 시그모이드 유닛(sigmoid unit)을 나타냄
- 장단기 메모리 셀의 상태(state)는 갱신은 다음과 같이 이뤄짐:

$$s_i^{(t)} = f_i^{(t)} s_i^{(t-1)} + g_i^{(t)} \sigma\left(b_i + \sum_j U_{i,j} x_j^{(t)} + \sum_j W_{i,j} h_j^{(t-1)} \right)$$

- 이 때 외부입력게이트(external input gate) ($g_i^{(t)}$)의 계산은 다음과 같이 수행됨:

$$g_i^{(t)} = \sigma\left(b_i^g + \sum_j U_{i,j}^g x_j^{(t)} + \sum_j W_{i,j}^g h_j^{(t-1)} \right)$$

- 장단기 메모리 셀의 출력 ($h_i^{(t)}$)은 출력게이트(output gate) ($q_i^{(t)}$)에 의해 제어가 가능한데, 구체적인 계산은 다음과 같음:

$$h_i^{(t)} = \tanh(s_i^{(t)}) q_i^{(t)}$$

$$q_i^{(t)} = \sigma\left(b_i^o + \sum_j U_{i,j}^o x_j^{(t)} + \sum_j W_{i,j}^o h_j^{(t-1)} \right)$$

- 출력게이트의 바이어스(bias), 입력가중치(input weight), 순환가중치(recurrent weight)는 각각 $\mathbf{b}^o, \mathbf{U}^o, \mathbf{W}^o$ 으로 표현됨
- 장단기 메모리 구조에서는 해당하는 셀에 새로운 입력 혹은 에러 신호가 존재하지 않을 경우 입력게이트 및 출력게이트에 의해 상태가 유지됨
- 엑티베이션이 낮은 상태에서는 게이트가 닫혀있어 관계없는 신호가 셀 내부에 입력되지 않아 셀 상태가 다른 게이트 상태를 변화시키지 않음
- 망각게이트가 없는 장단기 메모리는 임의 시간 동안 정보를 기억할 수 있는데, 이때 연속된 입력 스트림이 주어질 경우 셀 상태 값이 증가하여 출력 값 (h)의 포화 현상을 야기할 수 있음
- 망각게이트는 이러한 상황의 방지를 위해 메모리 블록을 초기화 할 수 있는 기능을 제공

제3장

텍스트마이닝의 개념과 최신 동향

제3장 텍스트마이닝의 개념과 최신 동향

제1절 텍스트마이닝의 개념

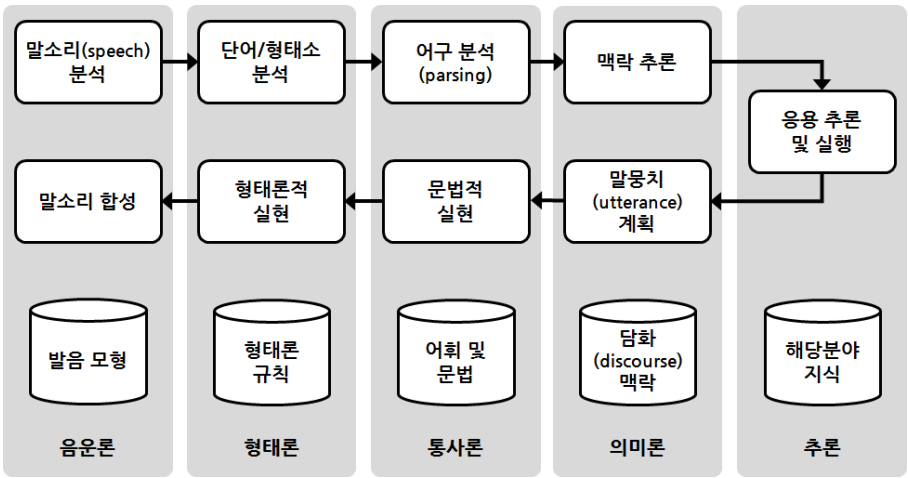
가. 자연어처리(NLP)에 대한 이해

□ 자연어처리(NLP: Natural Language Processing)는 기계가 인간의 언어를 이해하고 생성할 수 있도록 하기 위한 연구 또는 도구임

- ㅇ 최근 기술의 발전으로 컴퓨터를 활용하여 문자열을 토큰(token)화 하거나, 품사 태깅 및 번역까지 다양한 태스크를 수행할 수 있게 됨

□ 자연어 처리는 텍스트로 된 데이터에서 정확한 정보를 찾기 위한 검색엔진의 연구와 함께 발전해왔으며, 텍스트 내에서 어떻게 중요한 단어를 찾아낼지에 대해 연구하면서 발전

- ㅇ 언어학에 근간을 둔 자연언어처리는 음운론, 형태론, 통사론, 의미론 등 언어학의 세부분야로 표현할 수 있음[7)]



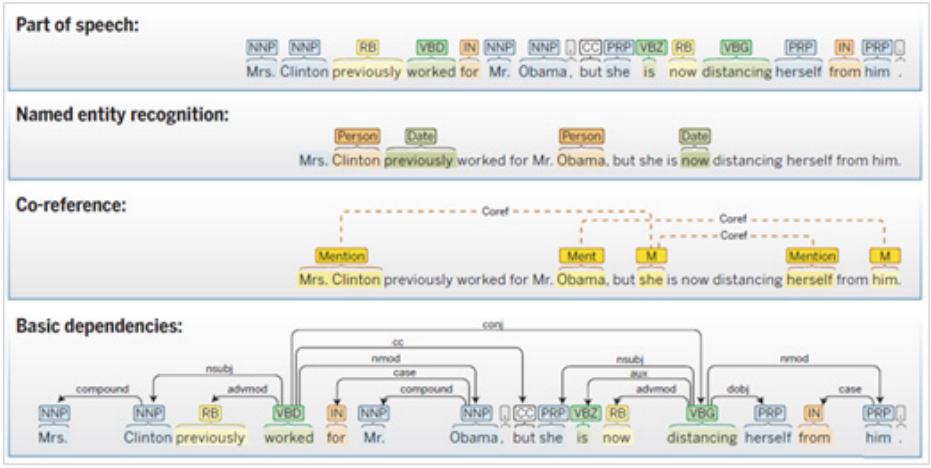
[그림 3-1] 자연언어처리의 전반적인 절차

※ 출처: NLP의 기본 절차와 Lexical Analysis (https://ratsgo.github.io/natural%20language%20processing/2017/03/22/lexicon/, 그림 재구성)

7) NLP의 기본 절차와 Lexical Analysis (2018.12.10. 접근) https://ratsgo.github.io/natural%20language%20processing/2017/03/22/lexicon/

ㅇ 어휘 분석(Lexical Analysis)은 품사태깅, 개체명인식, 상호참조, 의존관계 분석으로 구성[8)]

- 품사태깅(Part of speech, POS): 단어 품사정보를 추출함
- 개체명 인식(Named entity recognition): 인명, 지명 등 고유명사를 추출함
- 상호참조(co-reference): 선행 단어/구를 현재 단어/구와 비교하여 같은 개체인지를 판별
- 의존관계 분석(Basic dependencies): 단어 간 관계를 중시하여 문장구조를 분석함



[그림 3-2] 어휘분석의 종류

※ 출처: J. Hirschberg, C. D. Manning (2015) Advances in natural language processing, Science, 349(6245), p. 261-266

〈표 3-1〉 어휘분석의 절차

어휘분석 절차	분석 내용
1. 문장분리(Sentence splitting)	문서 내의 텍스트를 문장단위로 분리
2. 토큰화(Tokenizing)	의미를 가지는 문자열 단위로 문장을 분리
3. 형태소 분석(Morphological analysis)	분리된 토큰을 좀 더 일반적인 형태로 변환하여 단어수를 줄이고 분석의 효율성을 높이는 작업 ※ 대문자 → 소문자 변환, 동사원형으로 변환(Lemmatization) 등
4. 품사태깅(Part Of Speech; POS)	분리된 토큰에 품사를 부여하는 작업 ※ 전통적으로 의사결정트리(Decision Tree), 은닉 마코프 모델(Hidden Markov Model), 서포트벡터머신(Support Vector Machine) 등의 기계학습 방법론을 통해 수행

8) J. Hirschberg, C. D. Manning (2015) Advances in natural language processing, Science, 349(6245), p. 261-266

나. 형태소 분석

□ 자연어처리는 기본적으로 입력받은 문장을 형태소 단위로 해체는 태스크에서부터 시작함

- o 형태소는 의미를 가지는 최소한의 단위로, 형태소 분석 시에는 어휘적·문법적인 부분을 고려하여 입력된 문장이 해체 됨
- o 보통의 형태소 분석은 문장 내에서 형태소를 찾고 각각의 형태소에 다음과 같이 품사태깅(POS)을 적용하는 것 까지를 의미함

  - 예시) '하늘을 나는 새'라는 문장의 형태소 분석 예시

    하늘을: 하늘(일반명사)+을(조사)

    나는: 나(대명사)+는(보조사) // 날(동사)+는(관형형 전성어미)

    새: 새(일반명사) // 새(관형사)

다. 한국어 형태소 분석의 어려움

□ 한국어는 영어와 달리 단어 단위로 문장을 분리하는 것이 쉽지 않음

- o 보통 텍스트 분석 시 문장이나 글을 단어 단위로 나누어 입력받는 것이 편리하나, 한글은 조사가 단어에 붙어있고 어미 변환도 굉장히 불규칙함
- o 이에 따라, 대부분의 한국어 형태소 분석의 경우에는 형태소 분석기를 사용하여 문장을 구분하는 것이 일반적임

> (영어) My father enters the room. ⇒ my, father, enter, the, room
> ☞ 단복수 처리, 띄어쓰기 기준으로 단어 분리 정도만 해도 상당히 깔끔한 의미 단위로 구분된 단어들을 얻을 수 있음
>
> (한국어) 아버지가 방에 들어가신다. ⇒ 아버지가, 방에, 들어가신다 ⇒ 아버지, 방, 들어가다
> ☞ 간단한 룰로 조사를 분리하거나 동사의 기본형으로 바꾸는 것이 어려움
> (예시) "가"로 끝나는 것을 자르는 단순한 방법 적용 불가: 요가, 주가, 말라가, ...

- (불규칙한 용언 변화) 한국어의 일부 용언은 어간과 어미의 형태가 유지되지 않아 일정한 규칙으로 설명할 수 없는 경우가 있음[9)]

  ※ 불규칙한 용언의 변화를 일일이 사전화 시켜야하기 때문에 많은 작업이 필요

〈표 3-2〉 불규칙한 용언 변화의 예시

불규칙 용언 유형	예시
어간이 변하는 경우	짓다 – 짓고 – 짓지 – 지어 - 지은
어미가 변하는 경우	푸르- + -어 → 푸르러 이르(조(지))- + -어 → 이르러
어간과 어미가 둘 다 변하는 경우	하얗- + -아서 → 하얘서 파랗- + -아 → 파래

※ 출처: 용언 – 동사, 형용사 (https://brunch.co.kr/@adipoman/190)

- (어휘의 중의성) 한국어는 중의성을 갖는 단어가 특히 많아 풍부한 단어사전과 정교한 형태소분석 기술이 결합되어야 중의성의 해소가 가능

  ※ 예시) “나는”에 대한 분석은 1) 나(대명사 ‘나’의 의미) +는 2) 날다(동사 ‘날다’의 의미)+는 3) 나다(동사 ‘태어나다’, ‘발생하다’ 등의 의미) +는 으로 구분가능

- (외래어 및 사전 미등록 단어) 한국어는 외국어의 발음을 그대로 표기하는 경우가 많고, 사전에 미등록된 단어는 형태소 분석이 적절히 수행되지 않을 수 있음

〈표 3-3〉 외래어 및 사전 미등록 단어의 예시

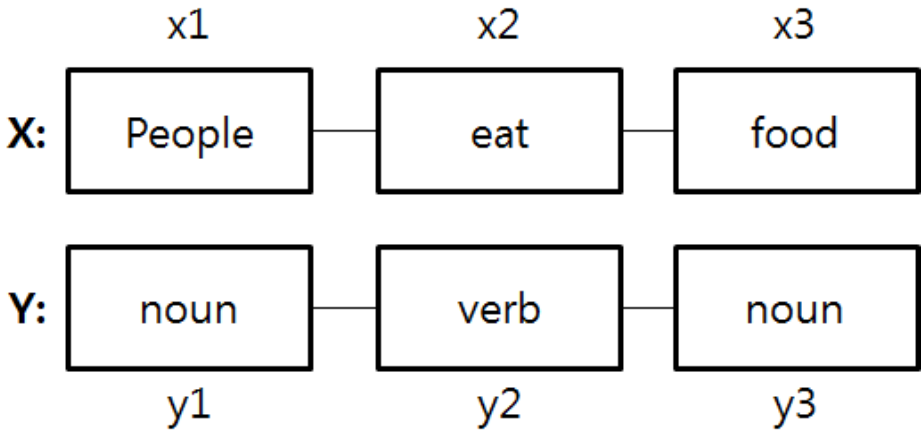
외래어 형태소 분석 오류 예시	사전 미등록 단어 예시
마이크로 → 마이크 + 로 드럼세탁기 → 드럼 + 세탁기 올레드(OLED) → 올 + 레드	드론 (새로운 단어) 화웨이 (회사명) 소확행 (인터넷 신조어)

라. 품사태깅(POS) 방법

□ 품사태깅은 순차적으로 입력되는 단어들의 앞뒤 순서를 확률적으로 계산하여 다양한 결과 값 중 문맥에 가장 적합한 하나의 결과를 선택하는 과정임

ㅇ 아래 그림에서 순차적으로 입력되는 단어(x_1)와 현재 품사(y_1)를 통해 다음에 올 수 있는 가장 적합한 품사(y_2)를 예측하는 문제라고 볼 수 있음

9) 용언 – 동사, 형용사 (2018.12.10. 접근) https://brunch.co.kr/@adipoman/190



[그림 3-3] 순차 데이터 레이블링 문제

□ 은닉 마코프 모델(Hidden Markov Model)

- 순차적인 데이터를 다루는 데 강점을 지녀 개체명인식(Named entity recognition), 품사태깅(POS)등 단어의 연쇄로 나타나는 언어구조 처리에 가장 널리 이용되고 있는 모델로 마코프 체인을 전제로 한 모델임
- 형태소 분석 뿐만 아니라 이전사건(앞서 입력된 단어)에서 현재사건(현재 입력된 단어)이 올 수 있는 확률과 현재사건 자체가 일어날 수 있는 확률을 연속적으로 계산 할 수 있음

□ 조건부무작위장(Conditional Random Fields, CRF)

- 조건부무작위장(CRF)은 레이블의 인접성 정보를 이용하여 레이블을 예측하는 방법론으로, 입력변수 X가 주어졌을 때 여러 특징함수를 통해 레이블 시퀀스 Y를 나타낼 확률을 구함
- 연구자가 유연하게 피쳐(Feature)*를 설정할 수 있고 최대엔트로피마코프 모델(MEMM)에서 발생하는 편향 레이블 문제(Label bias problem)의 극복이 가능함

* 기계학습 용어로, 데이터에서 사용자가 분석하고자 하는 측정가능한 속성을 말함

제2절 텍스트마이닝 분석 기법

가. 단어 추출

□ 문장의 형태소 분석을 통해 형태소를 구분한 후 가장 기본적인 단어 빈도를 추출하기 위해서는 가장 기본적으로 원-핫 인코딩(One-hot encoding)* 방식으로 문서단어행렬(Document term matrix, DTM)구조의 데이터를 생성함

* 변수들을 이진법 벡터로 표시하는 방법으로, 텍스트마이닝의 경우 해당 문서에 해당 단어가 존재할 경우 1, 없다면 0으로 표시하여 전체 문서의 단어 분포를 0과 1로만 이루어진 하나의 행렬로 나타낼 수 있으므로 프로그래밍 환경상에서 데이터 처리가 용이해짐

ㅇ 문서단어행렬(DTM)이란 각 문서가 행(Row), 각 단어가 열(Column)에 나열된 행렬 구조로서 문서 상 단어의 출현빈도를 통해 문서와 단어 간 관계를 나타냄

	특정문서에 집중적으로 나오는단어	혼합단어		잘안쓰는 단어	
	단어1	**단어2**	**단어3**	**단어4**	**...**
문서1	0	1	0	0	...
문서2	0	2	0	0	...
문서3	7	2	1	0	...
문서4	6	0	5	0	...
....	...	...	....	...	...

[그림 3-4] 문서단어행렬(DTM)의 예시

ㅇ 문서단어행렬(DTM)을 통해 크게 단어를 3가지 특징으로 분류할 수 있음

ㅇ 문서단어행렬(DTM) 단어 또는 문서 간의 상관관계 및 전체 단어의 출현빈도를 쉽게 구할 수 있어, 텍스트마이닝을 하기 위해 가장 기본적으로 만들어야 하는 형태의 데이터 구조임

〈표 3-4〉 DTM으로 알 수 있는 단어의 특징별 분류

단어 분류	특징
잘 사용 안하는 단어	문서단어행렬에서 한 열(단어)의 합이 매우 작은 경우
자주(흔히) 사용되는 단어	문서단어행렬에서 한 열(단어)에서 0인 cell이 적은 경우
특정 문서에 집중적으로 사용되는 단어	문서단어행렬에서 한 열(단어)에서 특정 Row(문서)의 Cell값이 매우 큰 경우

나. TF-IDF

□ 정보검색 등의 과정에서 가중치로 사용되는 TF-IDF(Term Frequency - Inverse Document Frequency)는 주어진 문서군의 특정 문서 내에 단어의 중요 정도를 나타내는 통계적 수치를 의미한

- ㅇ 단어 또는 문서의 중요도를 단어의 출현 빈도를 통해서만 판단하면 흔한 단어(common word)나 흔한 단어를 많이 포함하고 있는 문서가 중요하다고 판단될 수 있음
- ㅇ TF-IDF는 어떤 단어가 특정 문서에 얼마나 집중적으로 출현했는지를 나타내서 중요도를 가늠할 수 있음
- ㅇ TF-IDF는 특정 단어의 출현 빈도와 특정 단어가 출현한 문서 빈도의 역수의 로그 곱으로 계산됨

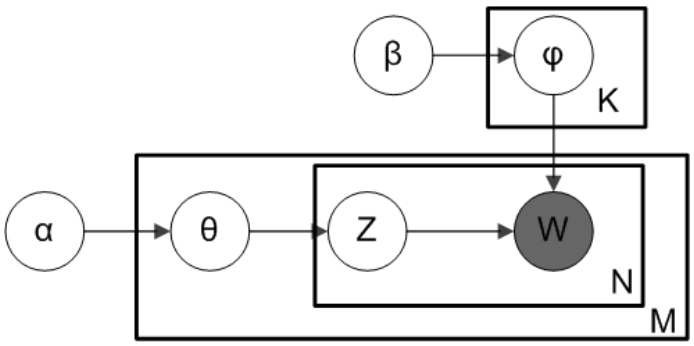
$$\frac{\text{특정문서내특정단어수}}{\text{특정문서내전체단어수}} \times \log \frac{\text{전체문서수}}{\text{해당특정단어가나타난문서수}}$$

다. 토픽 클러스터링(Topic Clustering)

□ 주제 분류를 의미하는 토픽 클러스터링은 각 문서가 어떤 주제들을 가지고 있고 각 주제별로 얼마만큼의 확률로 존재하는지 알아내는 방법임

- ㅇ 특히 LDA(Latent Dirichlet Allocation)라는 알고리즘이 가장 대표적이며, LDA는 사람이 글을 쓰는 과정을 반대로 진행하여 추론하는 방식과 비슷함
- ㅇ LDA와 같은 주제 분류는 정답지 없이 전체 문서들 내의 단어들의 동시출현 빈도의 분포에 따라 토픽을 구분[10]

- 각 문서별로 출현한 단어들을 보고 해당 문서가 어떤 토픽에 더 가까운지 0과 1사이의 값으로 나타냄



[그림 3-5] LDA 아키텍처

α : Dirichlet parameter, θ : 문서별 주제 분포도(topic distribution)
z : 각 문서 단어별 주제, w : 각 단어, β : Topic parameter
φ : 토픽별 주제어군, K : 주제의 개수, M : 문서 개수, N : 문서별 단어 수

※ 출처: Latent Dirichlet allocation (https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation)

ㅇ (LDA의 활용) 토픽 클러스터링은 태깅(레이블)되지 않은 문서를 효과적으로 묶어줄 수 있어 기계학습을 위한 태깅된 데이터를 만들기 위한 비용을 절약할 수 있음

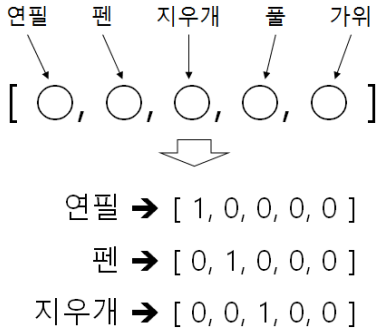
- 태깅되지 않은 데이터의 탐색적 데이터 분석(EDA) 작업, 이메일 주제 분류, 블로그/뉴스 주제 분류, 태깅된 데이터 생성을 위한 1차 작업 등에 많이 쓰임

다. 워드 임베딩(Word Embedding)

□ 워드 임베딩이란 원-핫 인코딩(One-hot encoding)을 사용하여 텍스트를 구성하고 있는 하나의 단어를 수치화하는 방법임

ㅇ 원-핫 인코딩은 단어에 해당하는 요소만 1이고 나머지는 0을 가지도록 벡터를 구성하여 텍스트를 숫자로 변경하는 방법으로, 단어 간의 관계가 전혀 드러나지 않는다는 단점을 내포하고 있음

10) Latent Dirichlet allocation (2018.12.10. 접근) https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation



[그림 3-6] One-hot encoding 예시

- 이를 해결하기 위해 단어 간의 관계까지 알 수 있는 형태의 벡터로 바꿀 수 있는 Word2vec, Glove, Fasttext등의 알고리즘이 개발됨

〈표 3-5〉 워드 임베딩 알고리즘의 종류

종류	설명
Word2vec	2013년 구글에서 발표된 방법론으로 가장 오래되었지만 가장 유용하게 쓰이는 방법론
Glove	2014년 스탠포드 연구팀에서 개발한 방법론으로 문서 내 단어 동시 등장 정보를 보존하기 위해 개발
Fasttext	2016년 페이스북에서 발표된 방법론으로 글이 짧고 노이즈가 많은 텍스트에 적합하도록 개발한 것이 특징임

□ 기계학습을 통해 사람과 같이 단어의 내재적 의미 파악을 통한 연관관계를 이해하려면 원-핫 인코딩 방식이 아닌 밀집표현(dense representation) 방식으로 표현되어야 함

※ 예를 들어, "사과"라고 하면 [과일, 달달함, 빨간 껍질, 주먹만 한 크기, 주스화 가능, 대구사과가 맛있지] 등과 같은 의미를 내재적으로 알고 있음

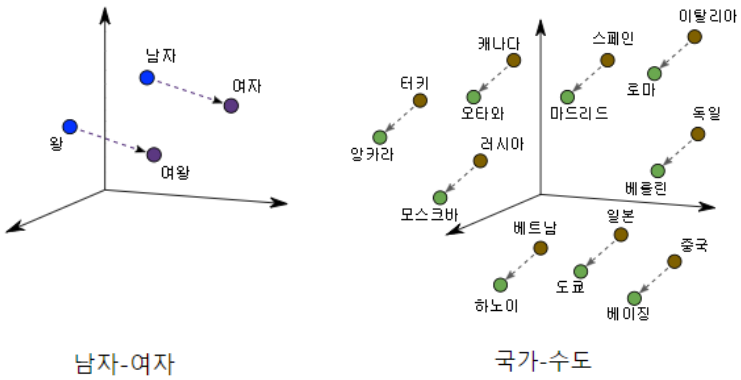
- 밀집표현은 원-핫 인코딩과는 달리 모든 벡터 성분이 값을 가지므로, 벡터의 모든 차원이 존재한다는 의미에서 dense라고 표현됨
- 밀집표현에서는 하나의 차원에 여러 속성들이 혼합되어 있기 때문에, 하나의 차원으로는 해석이 어렵고, 여러 차원을 조합하여 각 단어(벡터)간의 거리를 계산하여 단어의 연관정도를 계산할 수 있음

연필 ➔ [0.87, 0.12, 0.07, 0.52, 0.35]

펜 ➔ [0.27, 0.76, 0.12, 0.33, 0.21]

지우개 ➔ [0.29, 0.27, 0.92, 0.05, 0.55]

[그림 3-7] 밀집표현(Dense representation)의 예시



[그림 3-8] 밀집표현(Dense representation)에서 단어 간 연관정도 계산 예시

※ 출처: 임베딩:저차원 공간으로 변환 (https://developers.google.com/machine-learning/crash-course/embeddings/translating-to-a-lower-dimensional-space?hl=ko)

라. Word2vec

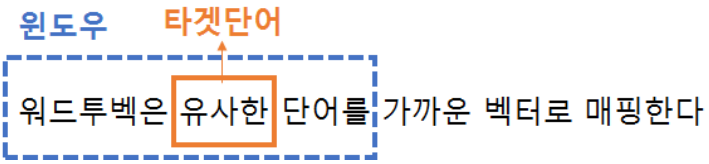
□ 대표적인 워드임베딩 방법론 중인 하나인 Word2vec은 분포 가설에 기반하여 의미상 유사한 단어를 동시 등장 정보(words of co-occurrence)를 사용하여 근거리 내에 벡터화 시킴

ㅇ Word2vec은 크게 Continuous Bag of Words(CBOW)와 Skip-Gram 두 가지 방식으로 구성

ㅇ CBOW 방식은 주변 단어의 맥락(context) 이용하여 타겟 단어(target word)를 예측함

- 주변 단어란 보통 타겟 단어의 직전 몇 단어와 직후 몇 단어를 뜻하며, 타겟 단어의 앞뒤에 있는 주변 단어의 범위를 윈도우(window)라고 함[11)]

11) 쉽게 씌어진 word2vec (2018.12.10. 접근) https://dreamgonfly.github.io/machine/learning,/natural/language/processing/2017/08/16/word2vec_explained.html (그림 재구성)



[그림 3-9] 윈도우와 타겟단어 예시

* 형태소 분석을 감안하면 실제는 [그림 3-9]와는 다르게 매핑됨

※ 출처: 쉽게 씌어진 word2vec (https://dreamgonfly.github.io/machine/learning,/natural/language/processing/2017/08/16/word2vec_explained.html, 그림 재구성)

ㅇ 반대로 Skip-gram방식은 타겟 단어로 주변 단어를 예측하는 방식으로 문제를 해결함

- Skip-gram은 단어 하나로 여러 단어를 예측해야 되기 때문에 정확도가 낮아 보이지만, CBOW보다 중심단어의 학습기회가 많이 주어지기 때문에 오히려 더 좋은 결과를 보여주는 장점이 있음

ㅇ 다만, 워드 임베딩 기법은 기본적으로 단어 수를 기반으로 계산하므로 학습량이 충분하지 않거나(학습대상 말뭉치 양 부족) 특정 주제에 치우쳐져 있을 경우 예상과 다른 결과가 도출되는 한계점이 있음

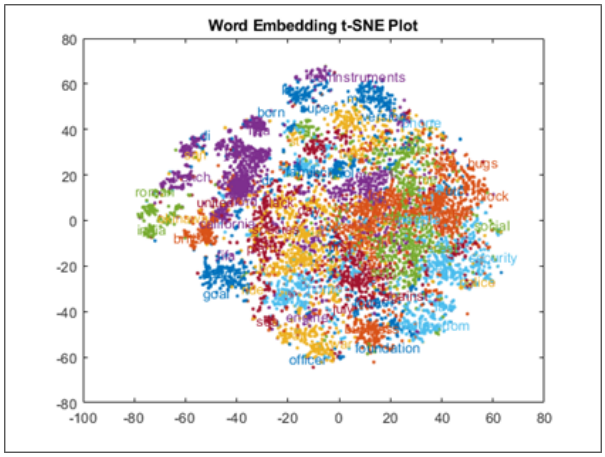
□ t-분포 확률적 임베딩(t-SNE) 방법론을 통한 word2vec 시각화

ㅇ t-SNE(t-distributed Stochastic Neighbor Embedding) 방법론은 고차원 데이터의 이웃한 구조를 유지하면서 낮은 차원으로 변환하는 방법 중 하나임

- 고차원 데이터는 [그림 3-10]과 같이 일반적인 2차원 그래프 형태로 출력[12]됨

ㅇ [그림 3-10]에서 같은 색깔의 점들이 뭉쳐 있음을 확인할 수 있는데, 이는 단어의 의미가 유사할 경우 같은 색깔의 점으로 표현되어 서로 가깝게 위치하기 때문임

12) Mathworks (2018.12.10. 접근) https://kr.mathworks.com/examples/text-analytics/mw/textanalytics-ex62579343-visualize-word-embeddings-using-text-scatter-plots



[그림 3-10] t-SNE를 활용한 시각화 예시

※ 출처: Mathworks (https://kr.mathworks.com/examples/text-analytics/mw/textanalytics-ex62579343-visualize-word-embeddings-using-text-scatter-plots)

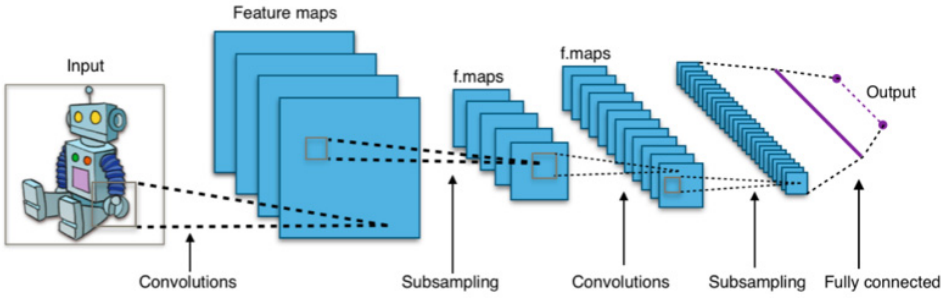
제3절. 딥러닝을 통한 텍스트마이닝의 진화

□ 딥러닝 기술이 컴퓨터 비전과 패턴인식 같은 분야에서 의미 있는 발전을 이루면서 자연어 처리 연구에도 딥러닝 기반의 연구가 늘어나고 있는 추세임

가. 컨볼루션신경망(Convolutional Neural Networks, CNN)

□ 이미지 분류에 많이 사용되던 컨볼루션신경망이 최근 텍스트에 적용되어 문서요약, 감성분석, 의미론적 검색 서비스 등에 활용 중

○ 전술한 컨볼루션신경망의 알고리즘상의 특성을 활용하여 단어가 결합된 결과물에서 높은 수준의 특징적 피처를 추출함



[그림 3-11] 이미지 처리를 위한 컨볼루션신경망 아키텍처

※ 출처: Convolutional neural network (https://en.m.wikipedia.org/wiki/Convolutional_neural_network)

□ 컨볼루션신경망 활용 시 문장의 긍정·부정 여부, 질문유형 분류 등의 판단 시에 뛰어난 성능을 보이며, 질의한 단어/문장의 가장 적합한 답변(문서)을 순위화 할 수도 있음

□ 컨볼루션신경망은 맥락윈도우(contextual window) 내 유의미한 단서를 추출하는데는 고도의 효율성을 지니고 있으나, 많은 데이터를 필요로 한다는 단점이 있음[13)]

○ 먼 거리의 문맥정보, 즉 문장 내에서 멀리 떨어져 있는 단어들은 모델링하기 어렵다는 단점도 존재하는데, 이는 문장의 순차적인 정보를 보존하지 못함을 의미하므로 기계번역 같은 용도에는 적합하지 않을 수 있음

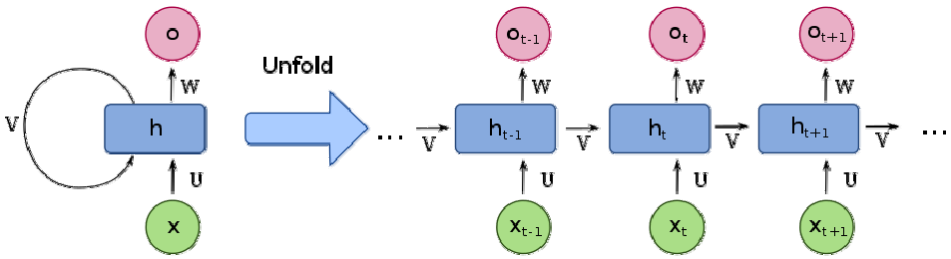
13) 딥러닝 기반 자연어처리 기법의 최근 연구 동향 (2018.12.10. 접근) https://ratsgo.github.io/natural%20language%20processing/2017/08/16/deepNLP/

나. 순환신경망(Recurrent Neural Networks, RNN)

□ 순환신경망(RNN)은 1990년에 연구된 아키텍처로서 데이터를 순차적으로 처리하여 언어 고유의 순차적 성격을 포착함

ㅇ 모든 입력 값을 독립적이라고 가정하는 순환신경망은 입력 시퀀스별 각 인스턴스에 동일한 작업을 수행하고 이전의 연산과 결과에 의존적*인 출력을 하는 특성을 지녀 "recurrent" 라는 용어로 표현됨[13)]

* 이전의 연산결과를 기억하여 현재의 연산과정에 이를 활용



[그림 3-12] 순환신경망 기본구조

※ 출처: Artificial neural network (2018.12.10. 접근) https://commons.wikimedia.org/wiki/Artificial_neural_network

□ 순환신경망은 다양한 길이의 문장, 문맥의 의존성*을 모델링 할 수 있어 망의 특성상 시간 분산 조인트 처리가 가능하여 텍스트 처리에 적합한 장점을 지님

* 시퀀셜로 데이터를 처리하기 때문에 언어 고유의 순차적 성격 포착이 가능한데, 그 결과'dog'와 'hot dog' 간의 의미 차이 구별이 가능함

ㅇ 가령, “hello”라는 단어가 한 글자씩 순차적으로 입력될 때 입력글자 바로 다음 글자를 예측하는 문제를 모델링 한다고 하면 다음과 같이 나타낼 수 있음[14)]

- “hell”을 입력하면 “o”가 반환되어 최종적으로 “hello”가 완성되는 모델로,
- “hell”이라는 글자를 원-핫 인코딩에 의해 벡터화하고 순전파(forward propagation)을 통해서 출력층(output layer)을 생성하게 됨

14) Andrej Karpathy (2015) The Unreasonable Effectiveness of Recurrent Neural Networks, http://karpathy.github.io/2015/05/21/rnn-effectiveness/

- 이 후 다음에 주어지는 글자를 답안으로 가정하여 역전파(back propagation)를 수행, 최적의 파라미터 값을 할당하게 됨

□ 순환신경망을 적용한 텍스트마이닝 연구로는 단어 분류, 문장 분류, 문장 생성 등이 있음

ㅇ (단어 분류) 개체명 인식, POS 태깅과 같은 문제 및 문자기반 모델링을 통해 중국어와 같이 복잡한 형태의 문자의 의미, 철자정보 추출 등에 활용

ㅇ (문장 분류) 문장 감성 분류, 문자 유형을 분류하고 적절한 답변을 매칭하는 문제 등에 적용

ㅇ (문장 생성) 서로 다른 시퀀스들을 매핑하여 기계번역 또는 응답메시지 자동 생성 문제에 활용하고, 이미지와 문장을 매핑하여 이미지 기반의 언어를 생성하는 등의 사례가 있음

□ 순환신경망은 언어모델링(Langurage modeling) 측면의 문제 해결에 적합한 알고리즘이나 문장 내 중요 단어 파악에는 컨볼루션신경망이 보다 우월성을 나타냄

다. 신경망 기계번역(Neural Machine Translation)

□ 구글 번역 및 국내 번역서비스의 활용도가 점차 높아지고 있고, TV방송에 해외에 나가서 한국어를 바로 현지 언어로 번역해서 말해주는 앱이 소개되기도 한 바 있음

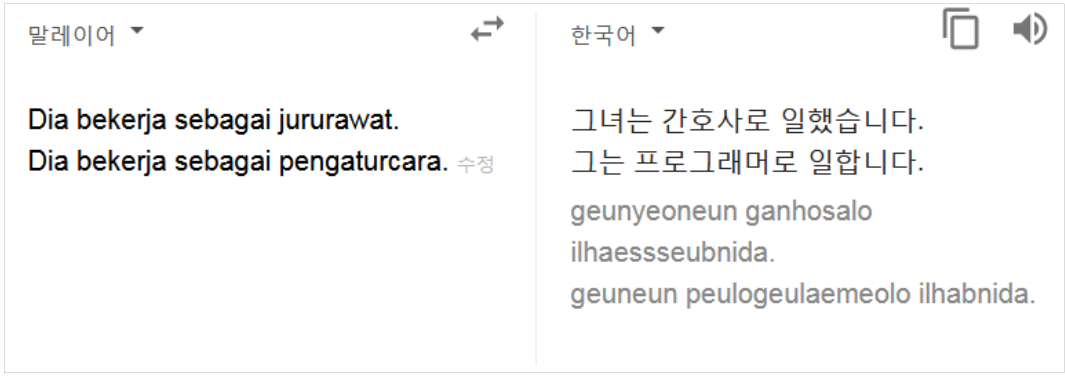
ㅇ 이는 번역의 정확도가 상당히 제고된 결과로 딥러닝 기술이 발전함에 따라 기존 통계방식의 기계번역이 신경망 기계번역으로 대체된 결과임

ㅇ 과거 통계방식의 기계번역은 구현에 있어 매우 복잡하고 절대적인 작업량이 많았지만, 신경망 기계번역은 end-to-end 모델로 많은 프로세싱 작업이 필요 없고 구현이 매우 단순하며 정확도가 높음[15)]

□ 다만 신경망 기계번역에 기대감이 점차 높아지고는 있지만 아직은 신뢰성 부족, 새로운 패턴인식 불가, 기억력 부재, 상식부족 등의 해결할 문제가 존재함

ㅇ (신뢰성 부족) 단순 철자나 문법오류가 아닌 문장의 뜻이 왜곡될 수 있음

15) Google AI Blog (2016) https://ai.googleblog.com/2016/09/a-neural-network-for-machine.html

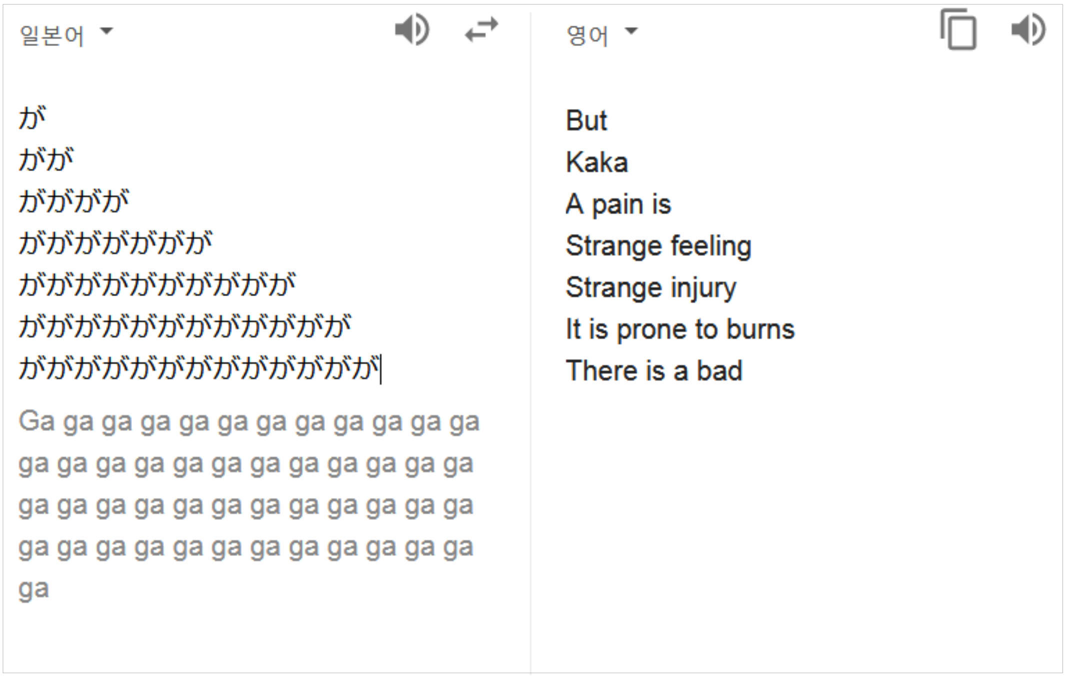


[그림 3-13] 말레시아어를 한글로 번역시 오류 예시

- [그림 3-13] 왼쪽의 말레시아어의 문장을 보면, 말레시아어는 성별("그", "그녀") 에 대한 발음을 구분하지 않고 하나의 단어 “sebagai”로 표현함
- 하지만 오른쪽 영어 번역 결과에서는 자동으로 성별을 구분하여 문장을 만듦
- 특히 간호사를 여자로 프로그래머를 남자로 자동으로 인식했다는 점이 문제인데, 이는 학습데이터의 대부분이 이러한 특징을 가지고 있었기 때문

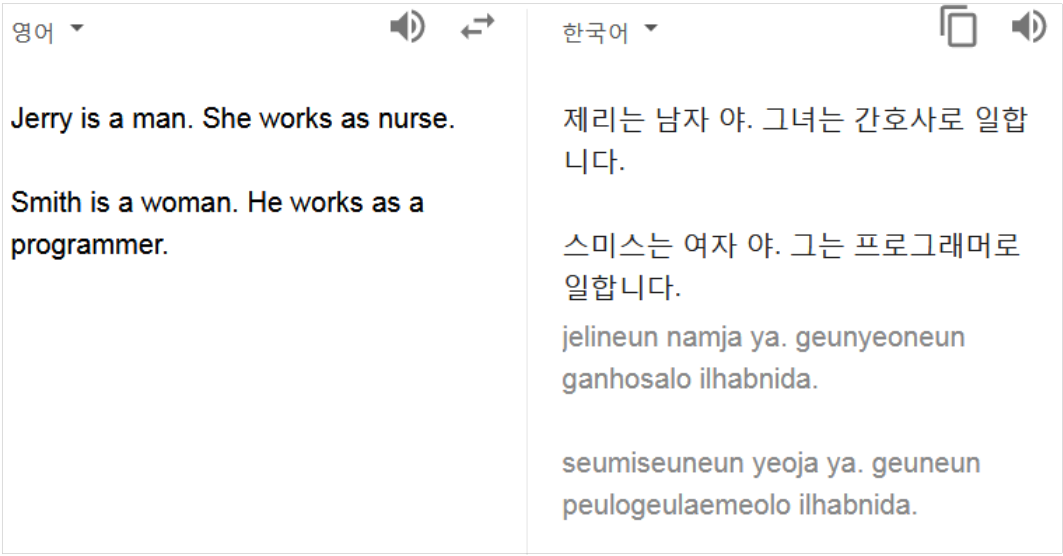
o (새로운 패턴 인식 불가) 새로운 패턴의 입력 값에 대해서는 적절하게 대응하지 못하는 문제가 있음

- [그림 3-14]와 같이 일본어 조사 が를 반복적으로 입력한 후 영어로 번역 시 유의미하지 않은 문장이 출력되는데 이는 모델이 이런 패턴을 단 한 번도 학습한 적이 없기 때문임
- 이는 컴퓨터가 학습하지 못한 새로운 문장 패턴이 입력될 경우 무의미한 문장을 반환하는 오류가 발생할 수 있다는 가능성을 나타냄



[그림 3-14] 반복적인 단어입력으로 인한 오류 예시

- o (기억력 부재) 신경망을 사용한 기계번역은 한 문장 단위로 번역하기 때문에 장문 번역시 많은 오류를 가져올 수 있음
  - 가령, 앞선 문장에서 제시된 인물의 성별, 장소와 같은 중요 정보를 다음 문장에 연결시키지 못할 수 있음
  - [그림 3-15]에서 영어 입력된 원문의 뜻은 "Jerry는 남자이다. 그는 간호사로 일한다." 이지만, 한국어로 기계번역 결과는 "Jerry는 남자야. 그녀는 간호사로 일합니다."로 표기됨
  - 이는 이어지는 문장을 번역할 시 Jerry가 남자라는 정보를 전혀 활용하지 못하기 때문에 아직 소설이나 통역과 같은 앞뒤 문장의 관계의 이해가 필요한 경우에는 기계번역이 부적합함



[그림 3-15] 앞선 문장을 무시한 오류 예시

ㅇ (상식부족) 시대, 트렌드, 글쓴이의 생각에 따라 글이 다양하게 해석되기 때문에 기계번역 시 판별하는데 한계점 존재함

- 가령, 영어 작성된 음악콘서트 기사 해석 과정에서 “I'm a huge metal fan!”이라는 문장을 한글로 기계번역 한다면 번역기는 해당 기사가 음악콘서트에 관련된 내용인줄 알 수 없기 때문에 전혀 다른 해석을 내놓을 수 있음

※ (입력) I'm a huge metal fan! → (출력) 나는 거대한 금속 선풍기다!

제4절 텍스트마이닝의 활용

□ (소셜미디어) 소셜미디어(블로그, 트위터, 페이스북 등)에 등록되는 엄청난 양의 텍스트를 수집하고 분석하여 사람들의 다양한 이야기를 통해 사회현상을 이해할 수 있음

- ㅇ 기업은 소셜미디어 분석을 통해 소비자의 심리, 행동을 이해하고 소비 트렌드를 센싱하고 그에 맞는 마케팅 활동이 가능

□ (온라인 쇼핑몰 상품후기 분석) 다른 사람들이 등록한 상품후기가 소비자의 구매 의사결정에 많은 영향을 미치면서 다양한 상품후기 분석을 하고 있음

- ㅇ 특히 긍정적인 상품후기, 부정적인 상품후기를 자동 분류하여 소비자에게 보여줌으로서 신뢰도를 높이고 부정적인 후기가 많은 상품을 자연스럽게 필터링 하는 효과가 있음
- ㅇ 상품의 새로운 감성정보(아기자기, 우아한, 컬러풀한, 편리한 등)를 찾아내어 소비자가 좀 더 쉽게 상품을 이해하고 탐색할 수 있도록 해줌

□ (문서 요약) 2012년 Summly라는 뉴스요약 서비스를 통해 텍스트 요약 기술이 알려지기 시작

- ㅇ 뉴스에 대한 주요 단어를 추출하여 추상적 요약(abstractive summarization)이라는 방식을 통해 자연스러운 요약 문장을 만들어 내는 수준까지 발전함
- ㅇ 현재 네이버, 다음, 중앙일보 등에서 특정 뉴스에 대해서는 문서요약 서비스를 제공하고 있음

□ (Image to text) 기존 OCR(Optical character recognition)처럼 이미지 또는 인쇄물 내 텍스트를 글자로 변환해주는 것이 아닌, 이미지 인식과 자연어처리 딥러닝 모델의 조합을 통해 이미지 내에 중요 정보를 인식하고 이를 사람이 이해 할 수 있는 텍스트로 변환

□ (챗봇) 메신저를 통해 들어오는 다양한 질문에 대해 사람이 일일이 대응하기 어려워지면서 자동으로 기계가 대화를 해주는 방식이 주목

- ㅇ 메신저의 사용시간이 늘어남에 따라 메신저를 통한 서비스(광고, 알림, 플친 등)가 점점 대중화되고, 소비자도 기존 채널인 전화, 문자가 아닌 메신저를 통한 소통을 선호하게 됨

- ㅇ 현재는 예상되는 질문에 따라 답변을 미리 정의하는 방식으로 여러 기업에서 도입하여 활용 중이나 점차 생성모델(Generative model)방식으로 답변을 자동으로 생성할 수 있는 형태로 개발 중

□ (기계번역) 딥러닝 기술이 발전하고 학습데이터가 점점 풍부해 지면서 점차 활용도가 높아지고 있음

- ㅇ 과거 통계기반 번역(SMT)의 낮은 정확도와 부자연스러운 번역 때문에 사용자가 적었지만 현재는 딥러닝 기반의 번역 기술(NMT)이 급속도로 발전하면서 서비스의 대중화를 이루고 있음[16)]

□ (음성비서·스피커) 2017년 기준으로 한국어 음성언어인식 정확도가 95%를 넘어가면서 여러 기업에서 스마트 스피커를 출시

- ㅇ 음성언어인식 기술 뿐만 아니라 제공할 수 있는 서비스가 무엇인지는 기업마다 다르기 때문에 제공 가능 서비스의 경쟁력도 중요한 문제임

〈표 3-6〉 주요 기업 음성비서 현황

기업명	제품명	출시일	주요 서비스 내용
아마존	에코	2014년 11월	제품주문, IoT, 음악재생 등
구글	구글 홈	2016년 10월	검색, 예약, 메시지전송, IoT 등
애플	Homepod	2017년 12월	애플 서비스와 연동
네이버	클로바	2016년 7월	네이버 서비스 연동(검색, 음악, 예약)
SKT	누구	2016년 9월	IoT, IPTV 제어 등
카카오	카카오미니	2017년 10월	카카오 서비스 연동(검색, 카톡, 음악)

- ㅇ 2018년 5월 구글 AI 어시스턴트의 레스토랑 예약 데모를 통해 스마트 비서·스피커 기술이 얼마나 빨리 발전하고 있고, 미래의 삶이 어떻게 변화될지 확인할 수 있음
- ㅇ 2018년 기준 한국은 글로벌 3위의 스마트 스피커 시장 점유율을 가지고 있을 정도로 큰 시장으로 성장 중

16) 배재경 (2017) 신경망 번역 모델의 진화 과정, 카카오AI리포트, https://brunch.co.kr/@kakao-it/155

□ (문서 분류 서비스) 제품 사용자 리뷰 점수 예측, 스팸메일 분류, 소비자 의견(VOC) 분류 등에 응용 가능

ㅇ (리뷰 점수 예측) 쇼핑몰에서 작성되는 소비자 리뷰의 점수를 예측

- 쇼핑몰의 상품 점수를 그대로 신뢰하기 어렵고 스팸이 존재하기 때문에 리뷰 내용을 보고 소비자의 만족도를 판단하여 리뷰 노출 순서를 변경하고, 낮은 점수 일부를 필터링 하여 사용함

ㅇ (스팸메일 분류) 이메일 스팸 분류 자동화는 가장 오래된 문제이기도 하고 과거에 많이 연구가 된 분야임

- 과거에는 단어 규칙을 등록하여 문서를 필터링 했지만 현재는 딥러닝 기술이 적용되어 더욱 정확한 분류 성능을 보여주고 있음

- Gmail은 받은 이메일 내용의 주요단어, HTML소스 형태, 이미지, 링크 등의 정보를 변수로 하여 크게 5개의 카테고리로 분류해주고 있음

ㅇ (VOC 분류) 기업의 가장 민감한 고객 접점인 고객센터에 접수되는 다양한 고객의 불만, 의견을 CS담당자가 수작업으로 분류하지 않고 자동으로 적절한 카테고리로 분류 및 적절한 답변 또는 대응 절차를 추천

- 일반적인 VOC(고객의 소리) 중 중요한 이슈 또는 장애 건을 자동으로 선별하기 위해 딥러닝 기술이 적용됨

- 중요한 이슈 또는 장애 건을 선별 한 후 비슷한 유형끼리 묶어주는 작업을 통해 고객 서비스 담당자의 수작업을 최소화

□ (AI변호사) 텍스트로 이루어진 방대한 판례 정보와 법 조항을 분석하여 사전 리서치 작업을 수분내에 처리

ㅇ 2016년 미국 AI 변호사 로스(ROSS)가 처음 출시된 이후 사람이 처리하는 영역의 상당부분을 대체할 수 있다고 평가됨

※ 국내에서는 유렉스라는 이름으로 출시

제4장

바이오의료분야 과학기술정보데이터 분석·활용 모형 고도화

제4장 바이오의료분야 과학기술정보데이터 분석·활용 모형 고도화

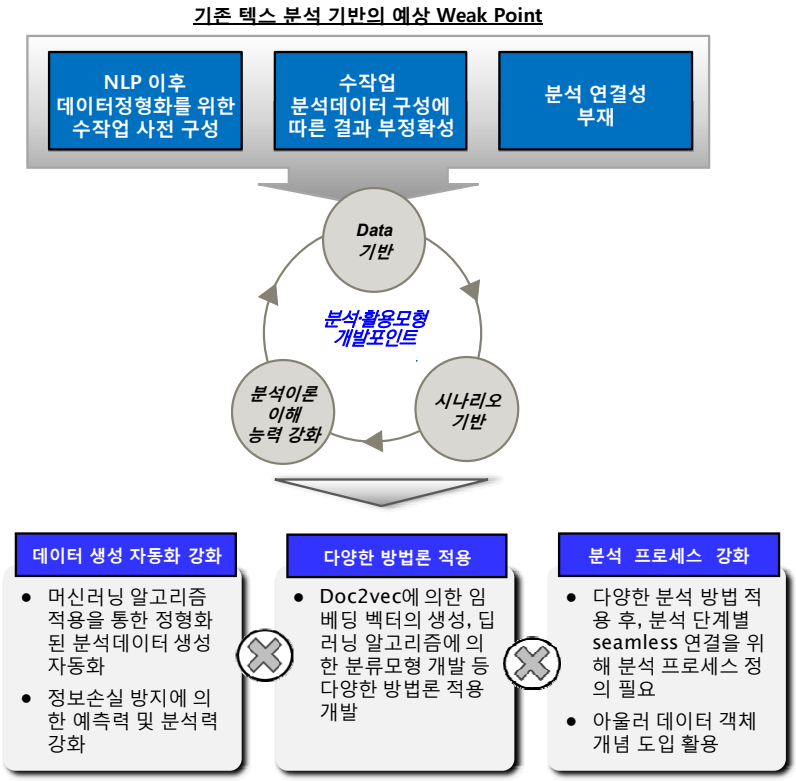
제1절 분석·활용 모형 개발 개요

가. 연구 추진 방향

□ 동 분석·활용 모형 개발 및 고도화 연구는 정성·정량적 분석을 수행하는 의사결정지원 모형을 업무 프로세스를 개선하는데 기여하고자 함

※ 사업분석 등 비정형데이터인 텍스트를 다루는 과정에서 사용자 편의를 개선

□ 이를 위해 앞서 소개된 기계학습 바탕의 텍스트마이닝 방법론(텍스트임베딩(embedding), 딥러닝 등의 알고리즘)을 활용 모형을 구축하고자 함



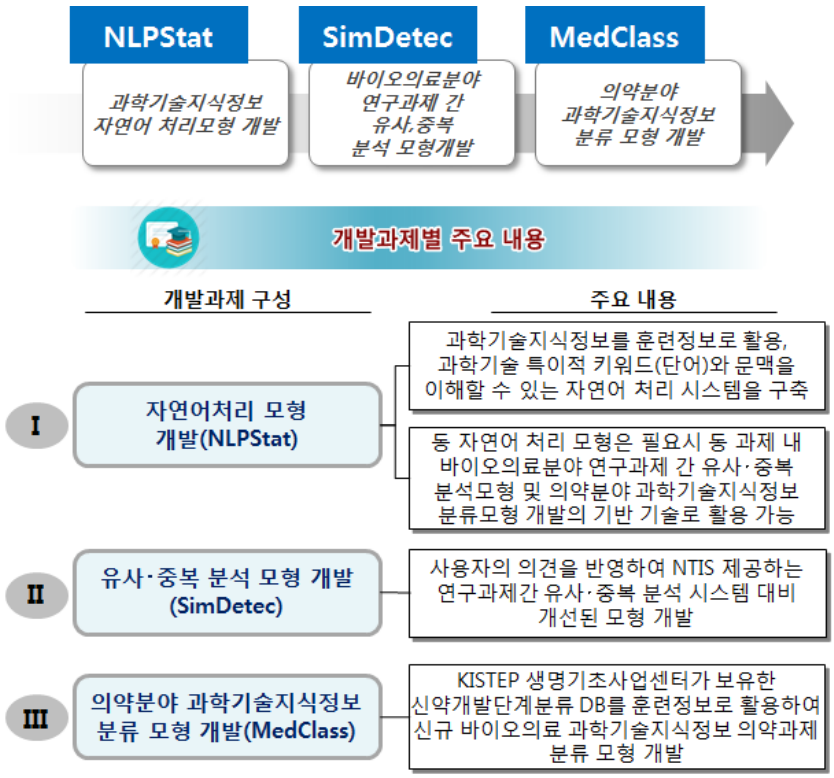
[그림 4-1] 분석·활용 모형 개발 추진방향

※ 출처: 김한해 외 (2017), 「기계학습 기반 바이오의료분야 과학기술정보데이터 분석·활용 모형 개발」, 한국과학기술기획평가원.

나. 분석·활용 모형 개발개요

□ 딥러닝 기반의 고도화된 기계학습 방법론을 활용하여 과학기술지식정보 자연어 처리, 바이오의료분야 연구과제 유사중복 탐색, 과제분류 모형을 개발

- o 과학기술정보는 주어진 항목이 동일한 장점은 있지만, 텍스트로 구성된 비정형 데이터에 해당함
- o 이를 고려하여 비정형 데이터를 원활히 탐색하고 정형화하는 기계학습 방법론(알고리즘)을 적용하여 사용자가 필요로 하는 기능을 구현하고자 하였음
- o 바이오의료분야 과학기술정보데이터 분석·활용 모형 개발 내용
  - (NLPStat) 과학기술지식정보 자연어처리 모형 개발
  - (SimDetect) 바이오의료분야 연구과제 간 유사중복 분석 모형 개발
  - (MedClass) 의료분야 과학기술지식정보 분류 모형 개발



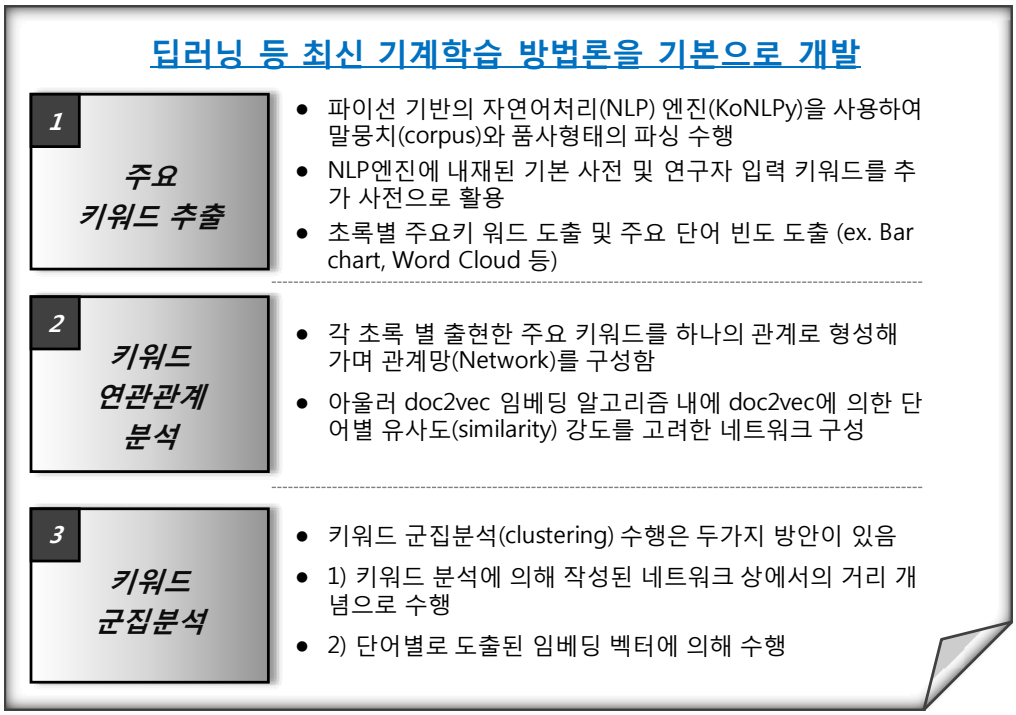
[그림 4-2] 분석·활용 모형 개발 개요

※ 출처: 김한해 외 (2017), 「기계학습 기반 바이오의료분야 과학기술정보데이터 분석·활용 모형 개발」, 한국과학기술기획평가원.

1) 과학기술지식정보데이터 자연어처리 모형(NLPStat)

□ 과학기술지식정보데이터 자연어처리 모형 개발 목표 및 기대효과

- ㅇ 동 모형은 과학기술지식정보를 훈련정보로 이용하여 특이적 키워드(용어)와 문맥의 이해가 가능한 자연어 처리모형 개발을 목표하였음
- ㅇ 매년 갱신되는 과학기술지식정보를 바탕으로 향후 학습을 통해 성능향상 및 데이터의 범주 확장을 계획함
  - 매년 신규 과학기술용어의 등장, 기술분야별 용어의 세분화 추세를 감안하여, 사용자 중심의 사전 구성 기능을 제공함
- ㅇ 과학기술지식정보를 임베딩 기술을 이용하여 벡터화하는 과정을 통해 비정형 데이터의 정형데이터화 기반을 마련함



[그림 4-3] 자연어 처리모형(NLPStat) 개발의 주요 목표 및 내용

※ 출처: 김한해 외 (2017), 「기계학습 기반 바이오의료분야 과학기술정보데이터 분석·활용 모형 개발」, 한국과학기술기획평가원.

□ 과학기술지식정보데이터 자연어처리 모형 개발 내용

o 과학기술지식정보데이터 자연어처리 모형은 본격적인 분석에 앞서 비정형 텍스트 데이터의 전처리 혹은 기초·탐색을 추진하는 단계에 해당

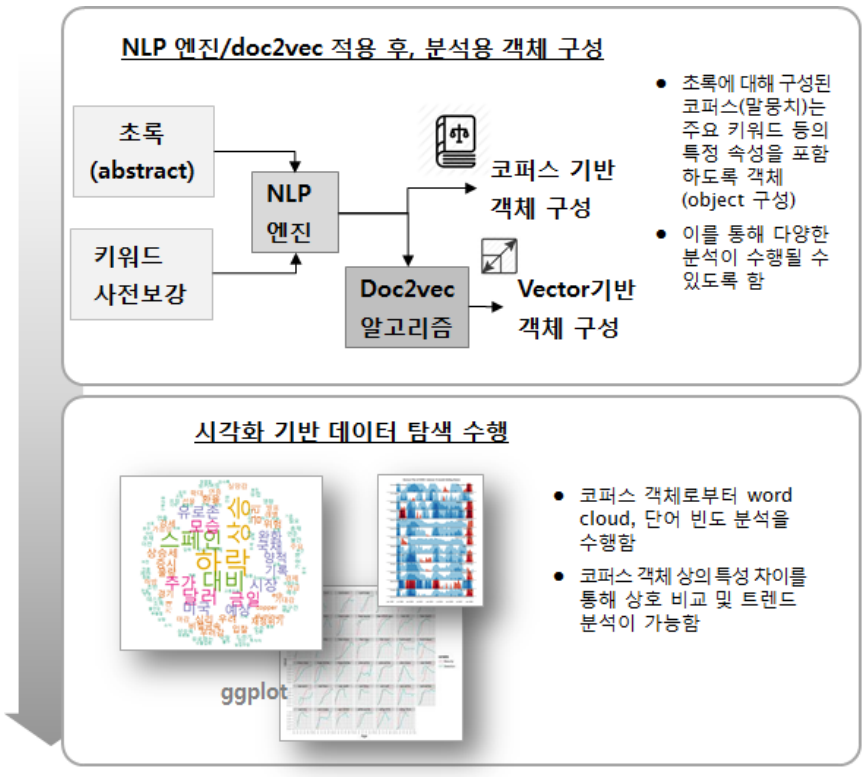
- 과학기술정보에서 주어진 용어들을 형태소 분석 과정을 거쳐 토큰화하고 관련 주요 키워드 추출, 용어(단어) 및 문서 수준에서 과학기술지식정보를 벡터화 함

o 주요 키워드 추출을 다음과 같은 두 가지 방식으로 구현되었음:

- 1) 사전 기반 단어 빈도수 카운트(일반적인 WordCloud)

- 2) 딥러닝 기반(word2vec) 알고리즘을 적용한 단어의 벡터 값 활용

o 이러한 두가지 방식을 활용하여 키워드에 근거한 데이터 경향 분석을 수행하고자 하였음

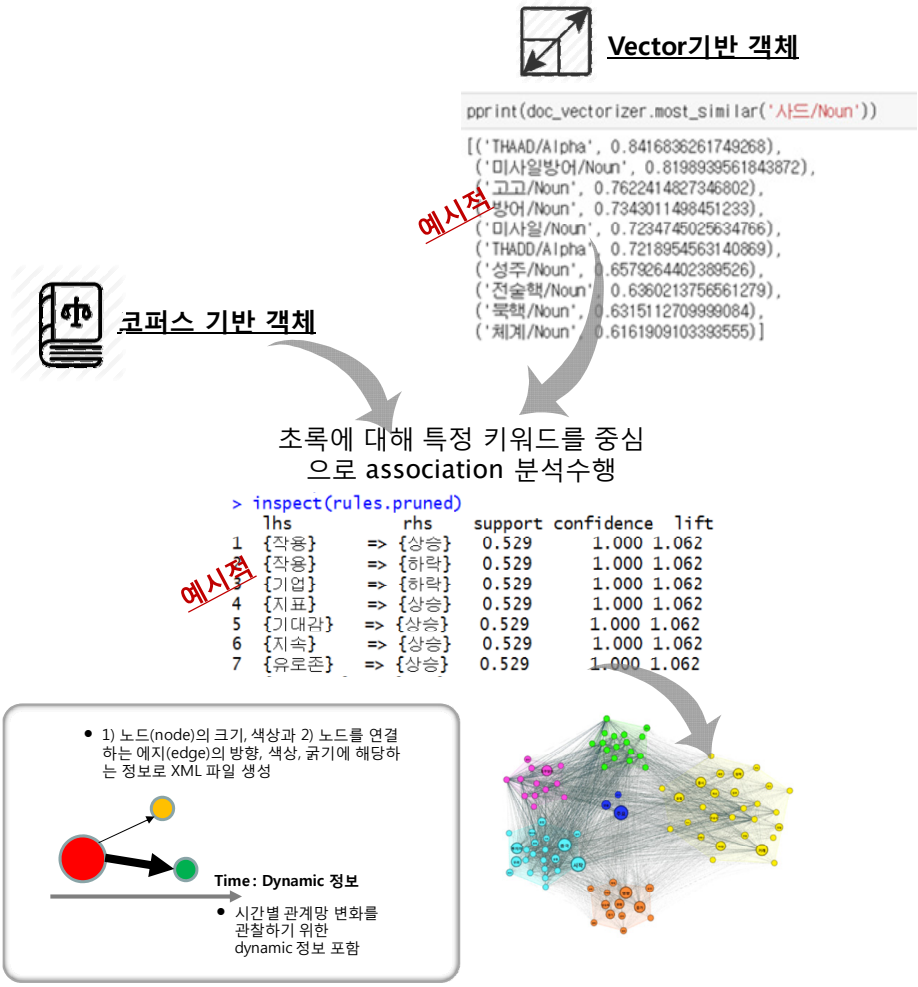


[그림 4-4] 주요 키워드 추출 방법 및 데이터 시각화 개요

※ 출처: 김한해 외 (2017), 「기계학습 기반 바이오의료분야 과학기술정보데이터 분석·활용 모형 개발」, 한국과학기술기획평가원.

ㅇ 주요 키워드 간 연관관계의 분석도 수행이 가능한데, 이는 1) 코퍼스 기반 객체, 2) 벡터 기반 객체 적용 방식으로 구분이 가능

- 1) 코퍼스 기반 객체 방식은 apriori 알고리즘을 기반으로 연관관계(association)를 분석한 후, 관계망을 구성함
- 2) 벡터기반 기반 객체 방식은 초록 간의 유사도나 주요 키워드 간의 유사도를 중심으로 관계망을 구성함



[그림 4-5] 주요 키워드 간 연관관계 분석 적용 방법

※ 출처: 김한해 외 (2017), 「기계학습 기반 바이오의료분야 과학기술정보데이터 분석·활용 모형 개발」, 한국과학기술기획평가원.

2) 바이오의료분야 연구과제 간 관계성(유사·중복) 분석 모형(SimDetect)

□ 바이오의료분야 연구과제 간 관계성(유사·중복) 분석 모형 개발 목표 및 기대효과

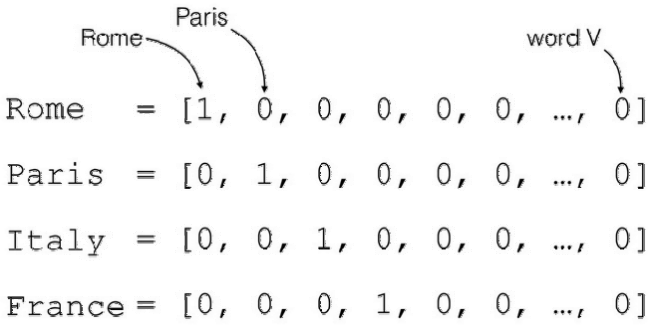
- ㅇ 국가과학기술정보(NTIS)에서 제공하는 연구과제 간 유사·중복[17] 분석 시스템과 다른 접근방식으로 작동되는 모형을 개발함
  - 당초와 같은 키워드 중복 빈도가 아닌 딥러닝 기반 텍스트마이닝 방법론을 활용하여 문맥적 흐름, 문장의 구성 등의 비교·분석을 통한 유사과제를 파악하고 이의 정도를 정량화 함
- ㅇ 기·신규과제 간 유사과제 탐색 시 키워드의 일치하지 않는 과제에 대해서도 유사과제의 판단이 가능한 환경을 제공

□ 바이오의료분야 연구과제 간 관계성(유사·중복) 분석 모형 개발 내용

- ㅇ 관계성 분석 모형의 경우 제시된 초록(연구내용) 간의 관계성(유사·중복성), 혹은 관련정도를 정량화된 수치[18]로 제시
  - 과학기술지식정보를 doc2vec 알고리즘에 의해 벡터화한 뒤 분석대상 과제 간 코사인유사도를 산출하여 관계된 정도를 가늠할 수 있음
  - 측정된 코사인유사도 값에 근거하여 시각화(네트워크) 기반 연관분석을 수행하고 이를 통해 과제 간 중요성 여부 등의 정성적 판단 수행
- ㅇ 관계성 분석 모형의 초록 임베딩 과정은 다음과 같은 프로세스로 수행됨
  - 과학기술정보데이터 자연어처리 모형(NLPStat)을 활용하여 연구과제별 용어(단어)의 원-핫 인코딩(one-hot encoding)을 수행함

17) 유사·중복의 의미는 관계성, 연관관계가 높다라는 의미로도 해석되므로, 동 연구에서는 유사·중복이라는 용어보다는 관계성이라는 용어를 사용함

18) 코사인유사도(Cosine similarity): 내적공간에서 두 벡터 간 각도의 코사인 값을 활용하여 두 벡터간의 유사성을 측정하는 방법으로, 동 모형에서는 과제를 벡터화하여 유사 과제 간 코사인 유사도를 측정

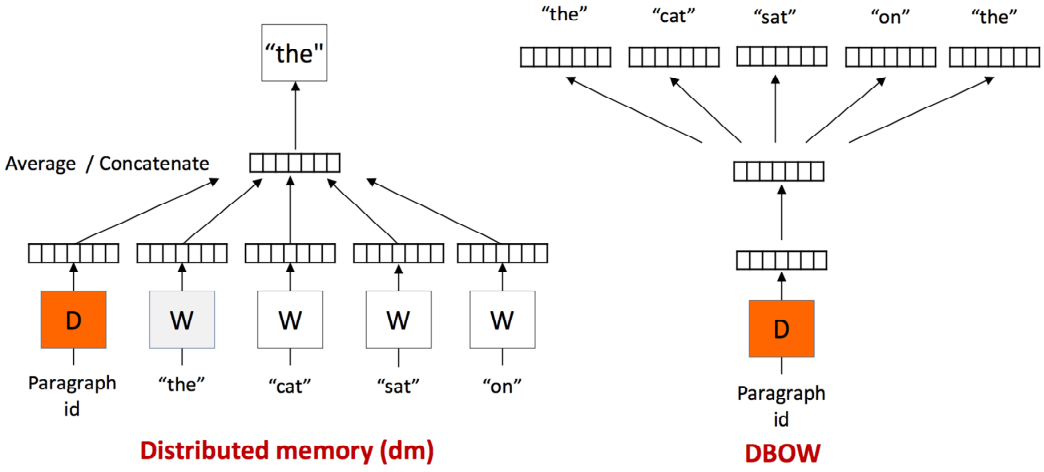


[그림 4-6] one-hot encoding 예시

※ 출처: 김한해 외 (2017), 「기계학습 기반 바이오의료분야 과학기술정보데이터 분석·활용 모형 개발」, 한국과학기술기획평가원.

- 윈도우사이즈(Window size)를 설정한 후, paragraph ID와 입력단어에 대한 대상단어(target word)를 예측하는 형태로 doc2vec을 수행(DBOW 방식)

※ word2vec 알고리즘을 적용하되, document ID를 벡터 길이 만큼의 unique ID를 부여하여 document 단위의 vector를 구함



[그림 4-7] doc2vec 알고리즘 개념도

※ 출처: 김한해 외 (2017), 「기계학습 기반 바이오의료분야 과학기술정보데이터 분석·활용 모형 개발」, 한국과학기술기획평가원.

3) 의약분야 과학기술지식정보데이터 분류 모형(MedClass)

□ 의약분야 과학기술지식정보데이터 분류 모형 개발 목표 및 기대효과

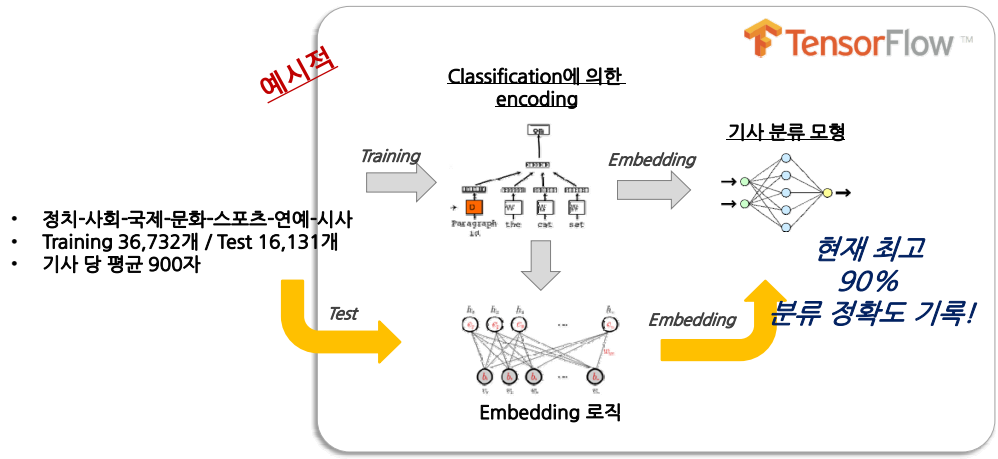
○ 그간 수집한 정부 신약개발연구과제DB를 훈련정보로 이용하여 신규 바이오의료분야 과학기술지식정보데이터 의약과제 분류 모형을 개발함

- 주어진 의약분야 연구과제의 사전 분류 정보(신약개발연구과제DB)를 토대로 사용자가 입력한 연구과제를 분류 기준에 근거하여 자동 분류함

- 매년 전문가 분류를 거쳐 신약개발연구과제DB를 갱신하고 있으므로, 동 모형은 신규정보를 반영하고 이를 학습하여 예측 성능이 향상될 수 있도록 개발

○ 딥러닝 기반 인공신경망 모델 구축을 통해 분류 성능의 우수성을 담보

- 전통적으로 이용되었던 기계학습 방법론을 대신하여 앞서 소개된 딥러닝 기반 인공신경망 모델을 활용, 과제분류의 성능을 고도화함



[그림 4-8] 의약과제 분류 모형(MedClass) 구현 및 작동방안 예시

※ 출처: 김한해 외 (2017), 「기계학습 기반 바이오의료분야 과학기술정보데이터 분석·활용 모형 개발」, 한국과학기술기획평가원.

□ 의약분야 과학기술지식정보데이터 분류 모형 개발 내용

○ 신약개발연구과제DB(훈련정보)를 doc2vec 알고리즘을 이용하여 벡터화하고 벡터화된 결과물에 딥러닝 신경망을 접목하여 분류 모형을 개발함

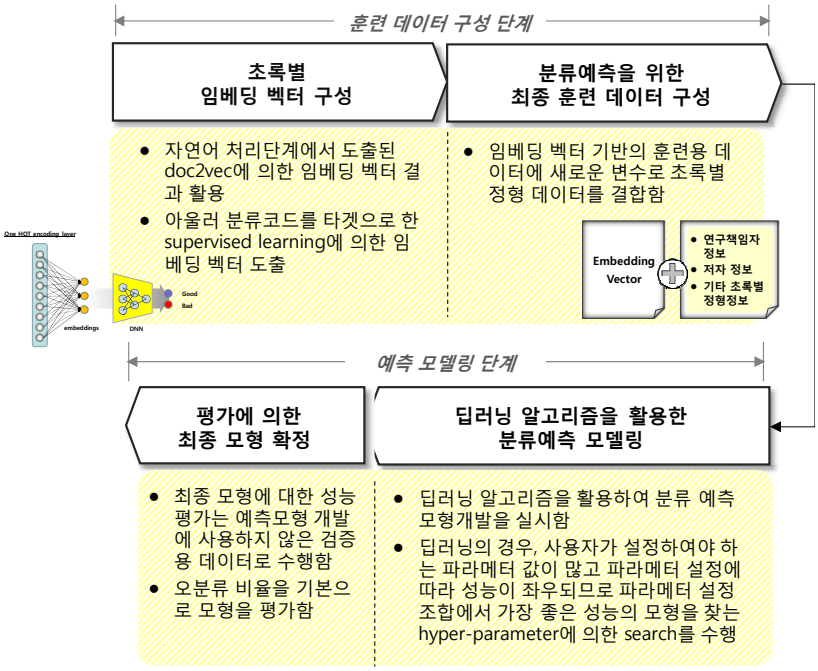
※ 분류 종류별 분류예측 모형 개발을 위한 지도학습(supervised learning)을 수행

- 2008-2016년까지 정부 신약개발 R&D연구 과제를 대상으로 모형을 개발함
- 과거 수행된 연구결과물*에 수립된 분류 기준을 적용, 관련 전문가를 통해 신약개발단계·의약품종류·질환별로 구분하여 분류

  * 신약개발 R&D 투자 효율화 방안(2012)

o 의약과제 분류모형 개발은 1) 훈련 데이터 구성, 2) 실제 예측 모델 개발 단계로 구분지어 추진됨

- 신약개발단계·의약품종류·질환별 분류 모형(Classifier)을 각각 구축하기 위해 주어진 분류 기준별로 개별적 학습을 수행함
- 기 구축된 2008-2015년 신약개발연구과제DB를 훈련집합/테스트집합으로 나누고 이를 활용하여 모형의 성능을 검증
- 최종적으로 기 과제 수행과정에서 신규 구축한 16년도 신약개발연구과제DB(전문가 분류)를 개발된 모형이 얼마만큼 잘 분류하는지 전문가 결과와 비교·분석함



[그림 4-9] 의약분야 과학기술지식정보 분류모형 개발과정 모식도

※ 출처: 김한해 외 (2017), 「기계학습 기반 바이오의료분야 과학기술정보데이터 분석·활용 모형 개발」, 한국과학기술기획평가원.

제2절 분석·활용 모형 고도화 방안

가. 연구 추진 방향

□ 동 연구과제는 기 개발된 분석·활용 모형을 바탕으로 1) 모형의 고도화 및 신규 분석·활용 기능 추가*, 2) 딥러닝 기반 토픽 클러스터링 분석 방법론 연구, 3) 17년도 정부 신약개발 R&D 과제 DB 구축 연구를 추진함

* 인터페이스 개선, 연구개발 사업 및 과제 간 관계성(유사정도) 분석 기능 고도화, 연구과제 분류기능 고도화, 조사분석 과학기술표준분류 검증 자동화 등

ㅇ 당초 분석·활용 모형에서 제공되던 3가지 모듈은 유지되고 새로운 사용환경 및 기능이 추가되는 방식으로 진행

나. 분석·활용 모형 고도화 방안

1) 연구과제간 관계성(유사정도) 분석모형 고도화

□ 동 과제에서 의미하는 관계성이란 기본적으로 “유사성”, “유사정도”를 의미하며, 이에 따라 두 가지 관점에서 관계성 모형을 고도화 하고자 하였음

□ 1) (유사중복성) 기 개발된 유사한 연구과제 검색(Simdetect) 기능은 인터페이스 개선으로 편의성을 제고함

□ 2) (연관·관련) “유사하다”는 의미는 “연관되어있다”, “관련이 있다”는 의미로도 해석 될 수 있으므로, 명확한 키워드를 입력하지 않더라도 연관된, 혹은 관련된 연구과제 검색이 가능한 모형의 특성을 활용하여 단시간 내 관심 분야의 연구과제를 검색하고 분석 기능을 제공

ㅇ 이를 위해, word2vec을 접목한 사용자 중심의 서치환경(검색엔진) 제공을 통해 사용자가 관심 있는 연구과제 검색을 가능토록 지원하고 관련 통계분석도 수행하도록 모형 고도화

ㅇ 분석·활용 모형은 doc2vec, word2vec 방법을 적용하여 문서정보(과제)와 단어(키워드)의 학습을 통해 구축되었으며, 이에 따라 관련 있는 단어(키워드)의 추출 및 시각화가 가능함

- 일반적인 wordcloud와는 다르게, 시각화 결과에서 관계성이 높을수록 큰 단어로 표기되어 사용자로 하여금 입력한 키워드와 연관된 중요 키워드의 선택적 추출이 가능(관계성(유사정도)의 정량화가 가능)

[그림 4-10] 기 분석활용 모형에 의해 줄기세포의 'stem'와 관련성이 있는 키워드로 제시된 단어 뭉치

※ 출처: 김한해 외 (2017), 「기계학습 기반 바이오의료분야 과학기술정보데이터 분석·활용 모형 개발」, 한국과학기술기획평가원.

ㅇ 당초 분석·활용 모형에서 doc2vec 방법론을 활용하여 연구과제 간 유사한 정도를 정량화 하고, 우선순위화가 가능한 점을 십분 활용하여 사용자가 우선순위화 한 과제목록을 사용자 의도대로 cut-off 할 수 있도록 설계

- doc2vec의 장점은 문맥의 파악으로 핵심키워드의 매칭이 있지 않더라도 관계가 되는 과제의 검색이 가능한 것에 있음

- 가령, '동물의 체내에서 사람의 장기를 생산하는 연구'를 입력하였을 때 출력되는 결과 〈표 4-1〉를 살펴보면,

- 줄기세포(stem cell), 유도만능줄기세포(iPSC), 배아줄기세포(ESC) 등의 키워드를 포함하지 않아도 줄기세포를 이용한 동물 기반 인공장기 연구의 탐색이 가능함

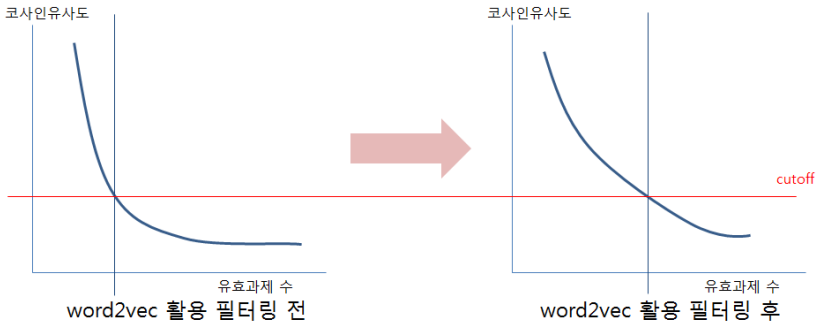
〈표 4-1〉 기 분석활용 모형에 의한 유사과제 분석 수행 예시

	유사도	과제수행년도	부처명	사업명	과제명-국문
1	0.923986868	2011	농촌진흥청	차세대바이오그린21	형질전환 복제돼지의 이식면역반응 특성규명 기술 확립(차세대바이오그린21)
2	0.923210918	2016	미래창조과학부	집단연구지원	인간화 돼지 연구센터
3	0.923134512	2015	농촌진흥청	차세대바이오그린21	Rag-2 유전자가 결핍된 면역결핍돼지를 활용한 환자 맞춤형 암 동물모델 개발
4	0.92236391	2007	교육과학기술부	우수연구센터육성〈SRC,ERC,MRC,NCRC〉	형질전환 돼지 생산을 위한 효율적인 초기배의 개발연구
5	0.919401263	2007	보건복지부	(보건의료기술연구개발)보건의료기술연구개발	돼지태아 줄기세포를 이용한 인슐린분비세포 분화와 이용에 관한 연구
6	0.918693227	2007	농촌진흥청	축산생명환경시험연구	형질전환가축 이용 바이오신약 생산기술 개발
7	0.918123561	2007	교육과학기술부	바이오신약장기사업	이종장기 이식 거부반응 억제를 위한 내피 및 상피세포의 조절
8	0.917363985	2009	교육과학기술부	미래기반기술개발	심혈관 질환 동물모델을 이용한 제대혈 및 지방조직 유래 줄기세포의 기능 연구
9	0.917051035	2012	교육부	일반연구자지원	돼지 심근경색모델에서 줄기세포 이식치료의 안전성 평가와 이식기술의 최적화를 통한 세포치료 기술 개발
10	0.917051035	2015	농촌진흥청	차세대바이오그린21	내분비호르몬이 돼지 및 인간 줄기세포의 EMT 조절을 통한 세포사멸에 미치는 영향 평가
11	0.91635837	2014	농림축산식품부	농생명산업기술개발	가축유래 전분화능 줄기세포를 이용한 고효율 형질전환/질환모델 동물 생산기술 개발
12	0.916288318	2010	농촌진흥청	바이오그린21	기세포 유래 면역세포 연구를 위한 바이오마커 및 기능조절 물질개발(바이오그린21)
13	0.916263344	2014	농림축산식품부	농생명산업기술개발	형질전환 돼지생산
14	0.916242148	2016	농림축산식품부	농생명산업기술개발	돼지 줄기세포를 이용한 유전자조작 기술의 확립

※ 출처: 김한해 외 (2017),「기계학습 기반 바이오의료분야 과학기술정보데이터 분석·활용 모형 개발」, 한국과학기술기획평가원.

[word2vec을 결과 활용 시 기대 효과]

- ㅇ 당초 키워드 기반의 빈도 혹은 필터링을 통한 과제 추출은 키워드 미싱 시 관련 과제를 추출할 수 어렵다는 한계점이 존재
  - 연구목적, 내용, 키워드 항목을 활용하여 추출한다면 보다 많은 과제를 추출할 수 있으나 이 경우에도 false-positive는 사용자가 직접 눈으로 가려내어 선별해 내어야 함
- ㅇ 동 분석 방법을 적용할 경우 doc2vec의 코사인유사도에 의해 정량화된 관계성 점수가 나오기 때문에 순위화가 가능하고 사용자가 일정 점수에서 cut-off하여 낮은 순위의 과제를 인위적 배제함과 동시에 관계성이 높은 과제는 최대한 포함할 수 있을 것으로 기대



[그림 4-11] word2vec 결과 활용 시 예상되는 기대효과

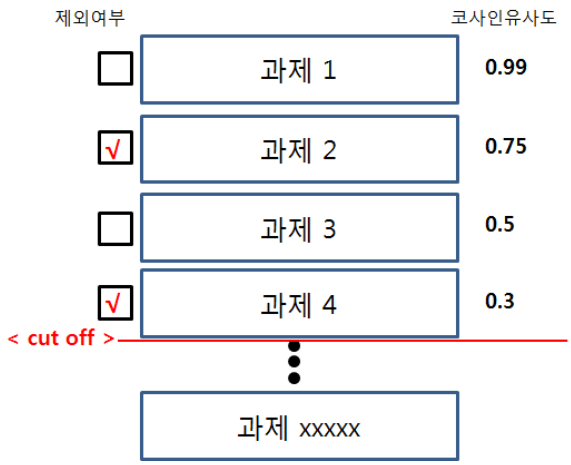
[활용 방안 예시: 감염병 관련 과제 3개년 투자 추이 분석]

- ㅇ 사용자가 감염병이란 키워드를 입력할 경우 감염병 관련 과제목록이 추출됨
- ㅇ word2vec에서 감염병과 관련된 키워드를 사용자에게 제공하고 사용자는 추천 키워드 중 적절한 키워드를 선택하여 관련 과제 목록을 도출

[그림 4-12] word2vec 기반 키워드 추천 예시

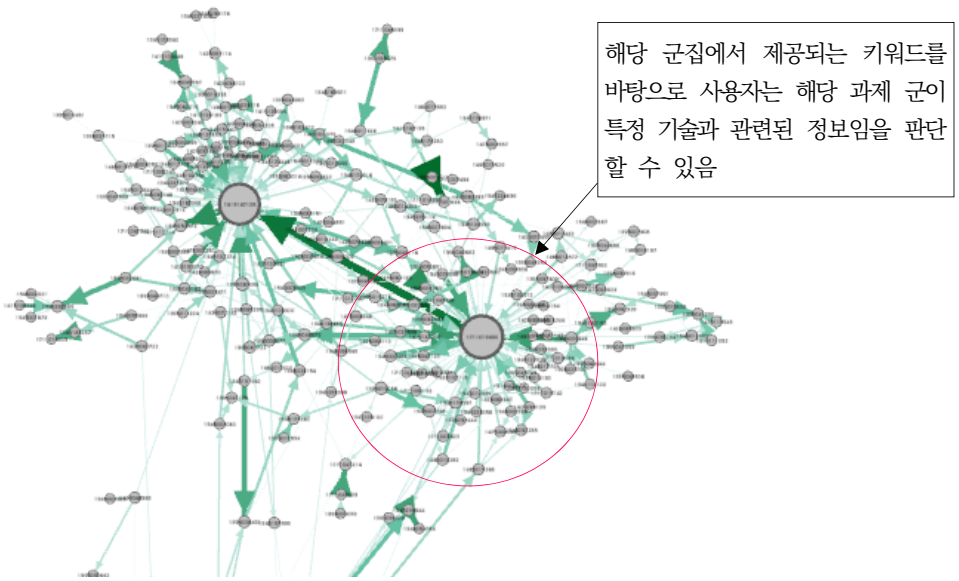
○ 사용자가 직접 관계성 점수(코사인 유사도)에 의해 우선순위가 결정된 과제목록을 보고 cut-off를 지정하여 연관성이 낮은 과제는 제외

- cut-off 내 과제들 중 사용자가 불필요한 과제를 제거할 수 있도록 장치 마련



[그림 4-13] 관련 과제 분석을 위한 과제목록 구축 예시

○ 관계성 분석 결과는 기 분석·활용 모형과 같이 네트워크 시각화가 가능하고 특정 노드들의 군집영역설정 시 해당 군집의 핵심키워드(토픽분석) 제공



[그림 4-14] 관련 과제 분석 결과 네트워크 시각화 예시

2) 정부 연구개발 사업 관계성(유사정도) 분석 모형

□ 동 모듈은 1) 사업 간 관계성(유사정도) 분석과 2) 특정 사업 속성 변화를 분석하는 것을 목표로 함

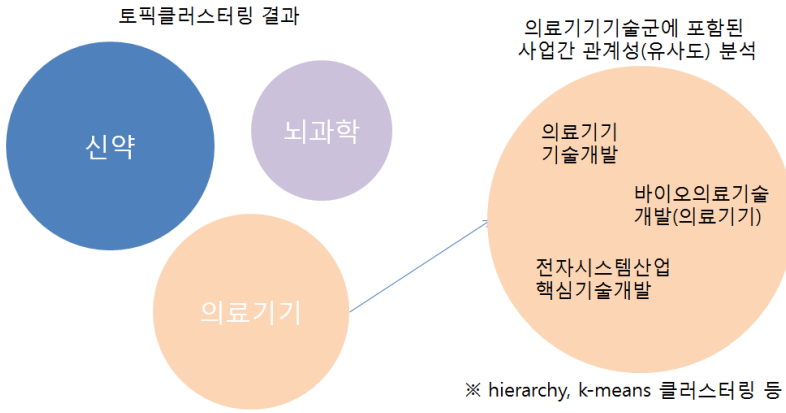
- ㅇ 사업 간 관계성(유사정도) 분석은 다수의 과제가 포함된 사업들 중 어떠한 사업이 서로 간에 유사한지 혹은 관련이 높은지 여부를 분석
- ㅇ 특정 사업 속성 변화는 특정 사업 내용이 시간이 지남에 따라 변화된 정도를 분석

[1. 사업 간 관계성(유사정도) 분석]

□ doc2vec 학습된 바이오의료분야 사업 내 연구과제들의 학습공간 위치를 활용하여 해당 세부/내역사업의 공간 상 위치를 특정화여 유사한 사업들이 서로 간에 우선 클러스터링 되도록 구현함

- ㅇ 예산배분조정과정에서 바이오분야는 신약, 의료기기, 줄기세포, 뇌과학, 바이오융복합, 임상보건과 같은 세부기술분야로 구분되고 있으며 각 사업들은 세부사업/내역사업 수준에서 세부기술분야로 매칭시켜 관리하고 있음
- ㅇ 세부사업/내역사업 수준에서 토픽 클러스터링을 통해 1) 추출되는 키워드를 분석

하여 현재와 같은 세부기술분야로 구분할 수 있는지, 어떠한 키워드에 근거하여 이러한 구분이 되었는지 분석한 후 2) 해당 세부기술분야에서 특정 사업간 관련성을 분석함



[그림 4-15] 사업 간 관계성 분석모형 예상분석결과 예시

o 해당 세부기술분야 내 특정사업간 관계성을 분석하기 위해 사업간 클러스터링을 실시할 경우 과제정보기반 doc2vec 결과물을 바탕으로 hierarchy 클러스터링 등을 시행

- single/average/complete distance를 바탕으로 계층화된 그래프화(혹은 네트워크식 레이아웃)가 가능한 hierarchy 클러스터링이 사용자에게 보다 직관적인 결과물로 활용이 가능할 것으로 판단됨

[2. 특정 사업 속성(사업내용 변화) 분석모형]

□ 1개의 분석 대상 사업에 속한 과제의 doc2vec 결과 값을 활용하여 클러스터링 방법론을 적용, 세부 및 내역사업 수준에서 내용(과제구성)의 변화를 탐지

※ 동 과제에서는 다차원에 투영된 과제를 -SNE 방법론을 적용하여 2차원화 하여 분석 결과를 제공

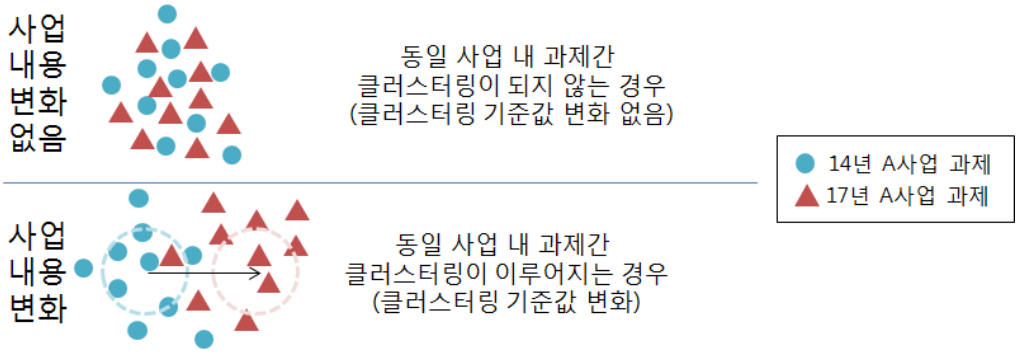
o 사용자가 특정 사업의 과제내용 변화 정도를 분석하고자 할 경우 동 모듈에 비교하고자하는 연도와 세부사업 및 내역사업 단위로 특정사업을 지정할 수 있도록 설계

- 세부사업 선택 시 해당 세부사업에 포함된 모든 과제가 분석대상이며, 내역사업 선택 시 해당 내역사업에 포함된 과제만 분석 대상에 해당

ㅇ 가령, [그림 4-15]와 같이 사용자가 2014년, 2017년, A세부사업을 선택할 경우 분석결과는 다음과 같이 제시가능

- (사업내용(과제구성)이 변화가 없는 경우) 14년 과제와 17년 과제 간 클러스터링이 이루어지지 않음(클러스터링 기준 값이 동일하거나 유사)

- (사업내용(과제구성)에 변화가 있는 경우) 14년 과제, 17년 과제로 구분되어 클러스터링이 이루어짐(클러스터링 기준 값이 이동)



[그림 4-16] 특정 사업 속성(사업 내용변화) 예상분석결과 예시

3) 딥러닝 기반 토픽 클러스터링 분석 방법론 연구

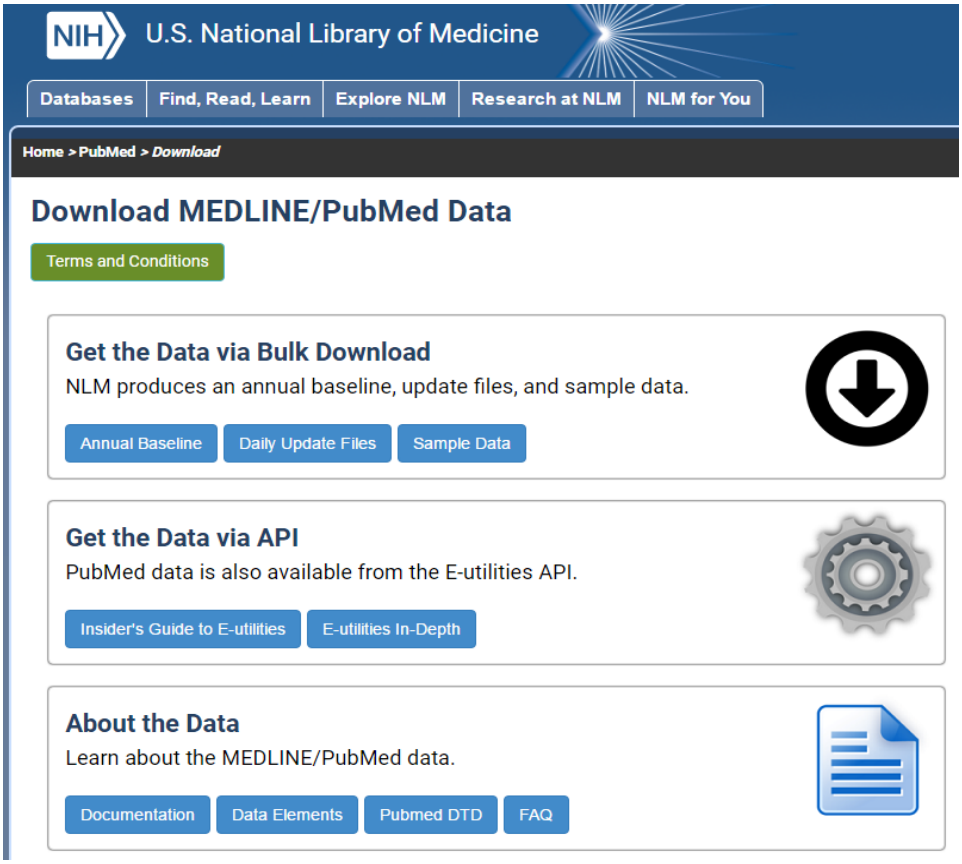
□ PubMed에 등재 된 모든 논문 초록정보를 다운로드하여 토픽 클러스터링을 수행하고, 국내 연구과제 정보 토픽 클러스터링 결과와 비교분석 추진

ㅇ PubMed 전체 서지데이터는 ftp 서버를 통해 제공되는 벌크 데이터를 다운로드를 통해 획득함(PubMed 서지데이터는 매일 업데이트됨, [그림 4-17] 참고)

- 다운로드 링크: https://www.nlm.nih.gov/databases/download/pubmed_medline.html

※ Annual Baseline(전년도 누적 데이터) : 1975년~2017년 11월 말까지의 서지데이터로, 파일 하나당 논문 30,000개씩 총 928개 파일이 업로드 되어있음(논문 2,784만개)

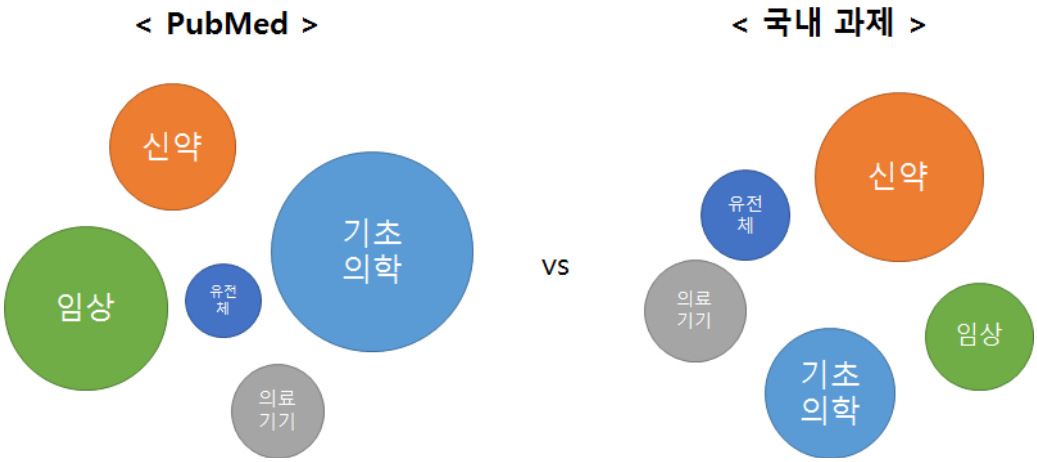
※ Daily Update Files(당해년도 일별 추가 데이터) : 2017년 11월 말 이후 업데이트분으로, 파일 하나당 논문 수 천 개 수준이며 총 300여개 파일이 업로드 되어있음



[그림 4-17] Pubmed 문헌정보 제공 웹사이트 정보

※ 출처: U.S. National Library of Medicine (https://www.nlm.nih.gov/databases/download/pubmed_medline.html)

- o 특정연도(또는 연도별) PubMed 서지데이터와 국내 연구 과제 정보(조사 분석)를 가지고 토픽 클러스터링을 수행하여 연구 분야의 분포 및 시계열 트렌드를 비교 (그림 5-16참고)
  - PubMed 서지데이터 다운로드(2008~2017년)를 다운로드하고 Abstract, Title, Year, Institution, Keywords, MeshHeading 등을 기준으로 DB화
  - PubMed의 각 논문 또는 국내 조사 분석 데이터의 각 과제를 하나의 노드로 간주하여 LDA 또는 doc2vec & k-means 방법론을 이용한 토픽 클러스터링 실시
- o 유저가 분석 범위 및 토픽의 개수를 정할 수 있게 하며, 토픽의 실제 주제를 추정하기 위해 인터페이스에서 각 클러스터별 최빈도 키워드 출력
  - 엑셀파일 형태의 토픽 분석 결과 다운로드 환경을 제공하여 사용자가 직접 추가적인 분석을 수행할 수 있도록 함

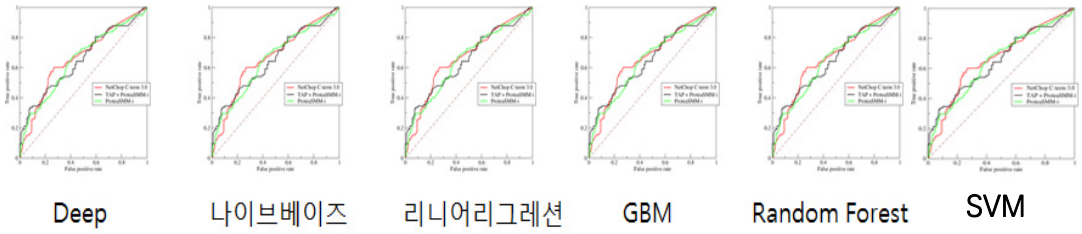


[그림 4-18] Pubmed 문헌정보 예상분석결과 예시

4) 의약과제 분류모형 범용화

□ 기 추진 분석·활용 모형에서 제작된 의약과제 분류모형을 범용화하여 데이터에 관계없이 입력 양식만 준수할 경우 활용이 가능한 범용적 분류모형 구축

- o 사용자별 보유한 자료의 분류 클래스 수 및 네이밍이 상이하므로 고도화 과정에서 이점이 고려될 필요
- o 현재 활용되는 딥러닝 기반 분류모형 이외에 Naive Bayes, Linear Regression, GBM, Random Forest, SVM 등 적용 가능한 5개의 방법론[그림 4-19]을 추가
- o 최종분류 결과 도출에 앞서, ROC Curve Analysis 등과 같은 예측결과 시뮬레이션을 통해 사용자가 최적의 알고리즘을 선택할 수 있는 의사결정 지원환경 구축
  - 데이터마다 분류되는 특성이 다를 수 있으므로, 알고리즘별 예측성능이 다를 수 있음
  - 딥러닝의 경우 별도의 학습이 요구되거나 소요시간에 따라 결과가 달라질 수 있으나, 그 외의 알고리즘은 비교적



[그림 4-19] 과제분류 모형에서 제공하는 6개의 방법론 및 예측결과 예시

- o 모형고도화 과정에서 사용자 의견을 반영하여 분류결과 통계정보 분석 및 그래프 시각화기능을 중심으로 旣 분류 모형의 사용 편의 및 활용 수준 제고

  ※ 매년 발간되고 있는 신약개발 통계브리프에서 제공되는 그래프 등의 양식을 준용

- o 이외에 NTIS 조사 분석 분류 모듈도 추가로 생성함

  ※ 기본적으로 의약과제 분류모형과 작동방식이 상동함

제3절 분석·활용 모형 활용 결과

가. 바이오의료분야 분석·활용 모형 고도화

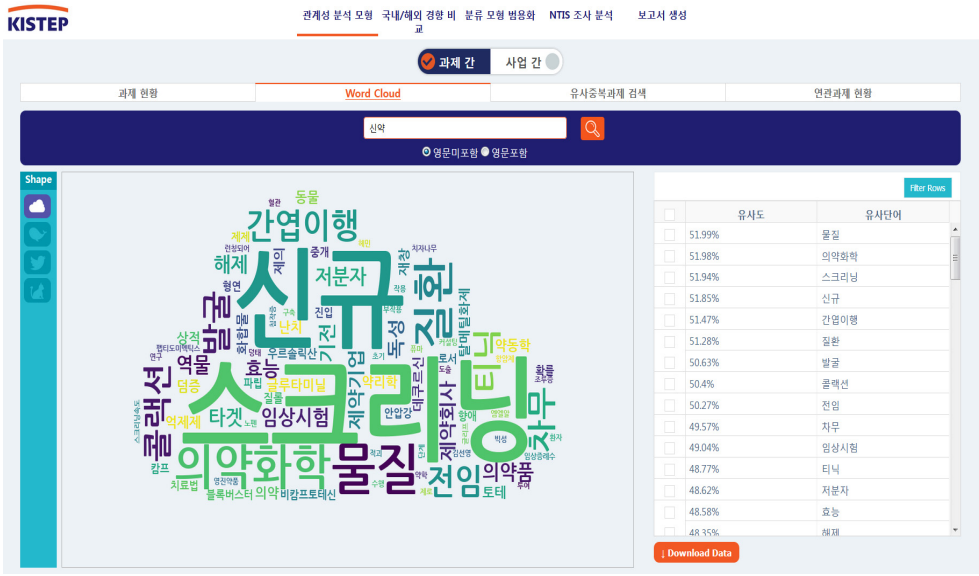
1) 연구개발과제 간 관계성 분석모형

[1. 연관 단어 분석]

□ word2vec 기반의 연구개발과제 내 연관 혹은 관련된 단어(키워드) 분석은 다음과 같은 방식으로 수행할 수 있음

o 키워드 입력 시 연관·관련 있는 단어가 시각화되어 표현됨은 물론, 코사인유사도에 근거하여 우선 순위화 된 단어의 목록도 함께 제시됨

- 사용자 기호에 따라 시각화 형태 및 영문 키워드 포함/미포함 여부를 선택 가능



[그림 4-20] '신약' 키워드를 활용한 연과 단어 분석 결과

□ 이를 활용한 결과에서 관계성이 높은 키워드를 검색 시 유의미한 핵심어가 가중치가 고려되어 출력되는 것을 확인할 수 있으나, 형태소 분석과정에서 둘로 분리될 수 있는 복합명사의 경우 유의미성이 다소 낮은 단어뭉치가 출력됨을 확인할 수 있었음

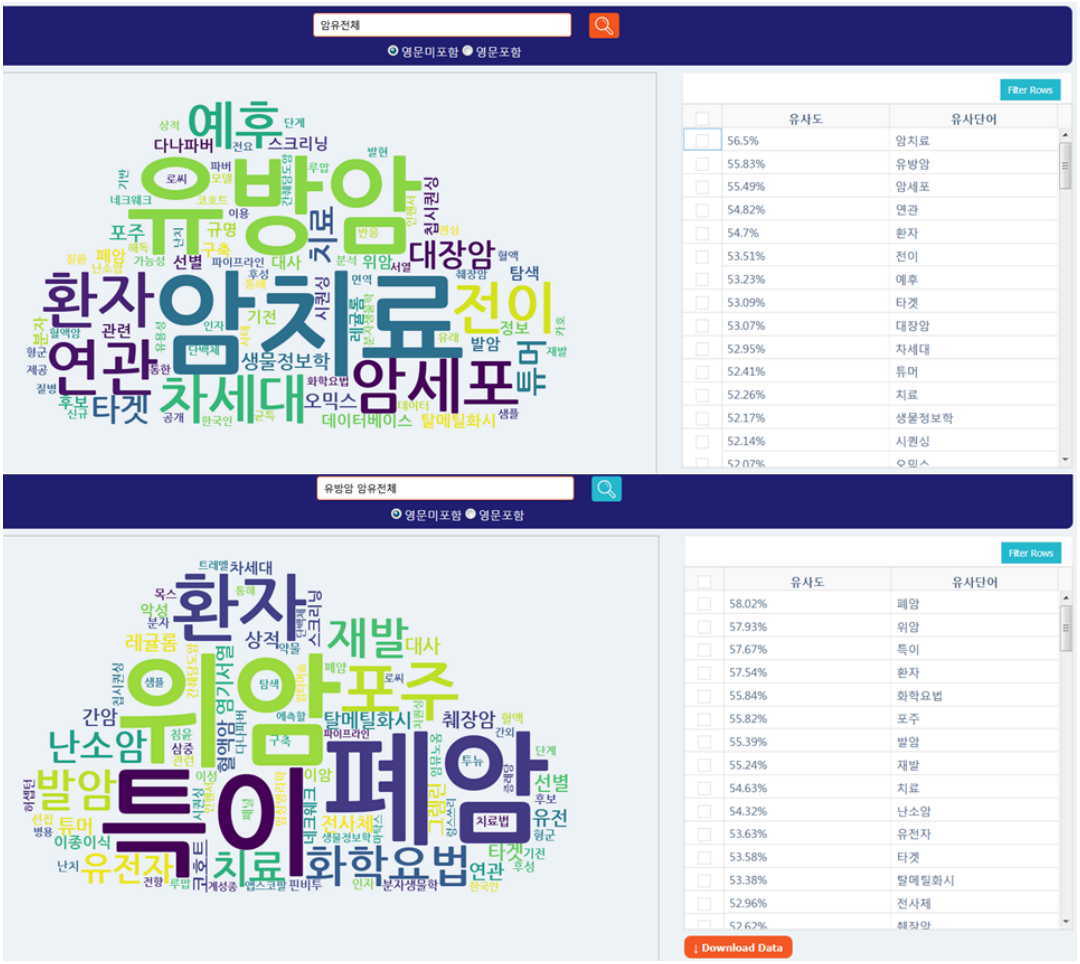
ㅇ 줄기세포의 경우 ‘줄기세포’라는 단어로 동 모형 내에 하나의 단어로 학습이 되어 있으나, ‘인공장기’의 경우 형태소 분석과정에서 ‘인공’ 및 ‘장기’로 분리되어 관계성 있는 키워드를 탐색함

- 동 분석활용 모형은 신규데이터 확보 시 재학습이 가능하도록 설계되어 있기 때문에 인공장기를 하나의 단어로 학습시켜 모형을 재구성할 경우 다른 결과가 도출될 수 있음

[그림 4-21] ‘줄기세포’, ‘인공장기’ 키워드로 관계성이 높은 단어를 도출한 결과

□ ‘암유전체’ 키워드를 활용해 연관 단어를 검색한 후 연관 단어 중 ‘유방암’ 키워드를 추가하여 검색한 결과[그림 4-22], 유방암 키워드가 추가됨으로 인해 보다 유방암 특이적인 연관 단어가 도출됨을 확인할 수 있었음

ㅇ 유방암, 암유전체 검색결과 유방암의 전이가 가장 흔한 폐(폐암)가 가장 연관이 높게 나타났고, 암유전체 관련된 키워드는 10위권 밖에서 나타남을 확인할 수 있었음



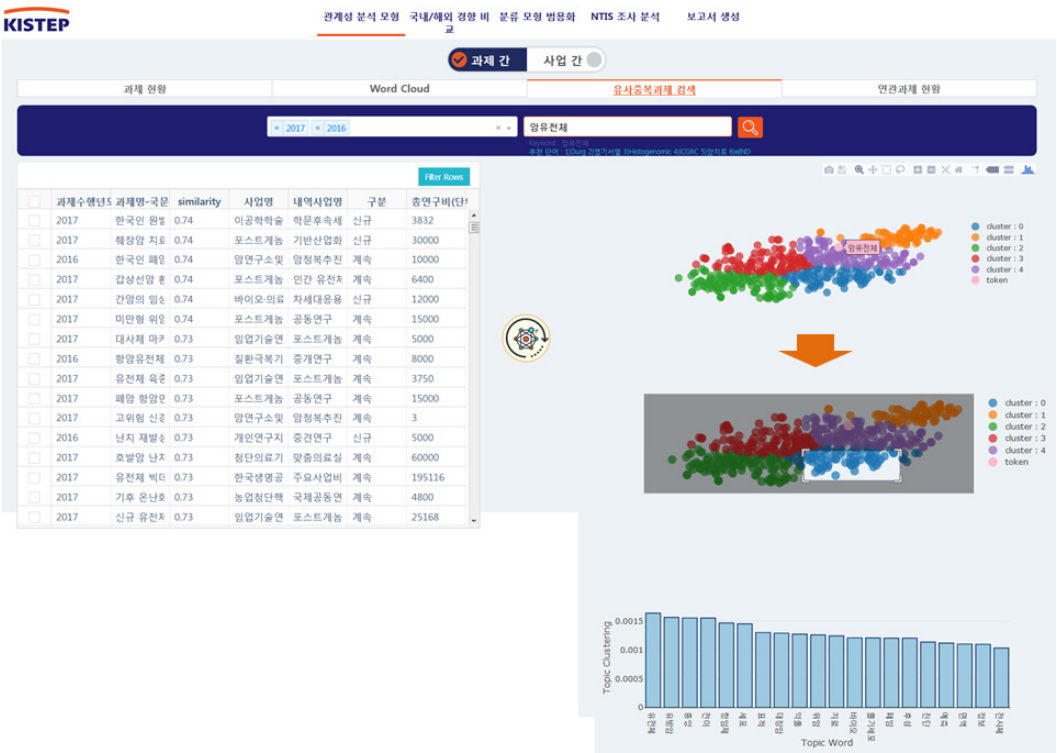
[그림 4-22] '(좌) 암유전체'와 '(우) 유방암, 암유전체' 키워드에 의한 연관 단어 분석 결과

[2. 관계성(유사중복성) 과제 분석]

□ doc2vec 기반의 연구개발과제 유사중복 과제 분석은 다음과 같은 방식으로 수행할 수 있음

ㅇ 유사중복과제 검색 모듈에서 검색대상연도 및 중복성 분석을 위한 키워드나 과제명, 연구내용 등을 입력할 경우 유사중복성이 높은 과제를 코사인유사도에 근거하여 출력하고 이를 네트워크로 시각화로 연계함

- [그림 4-23]는 '암유전체' 키워드를 입력한 예시로, 사용자가 키워드 입력 시 모형에서는 word2vec 결과와 연계하여 추천 키워드를 제공함
- 네트워크 시각화 시 입력된 키워드 혹은 문장의 클러스터 상 위치를 별도 확인할 수 있음
- 또한, [그림 4-23] 에서와 같이 색으로 구분되는 특정 클러스터를 마우스로 영역을 설정할 경우 연관 키워드를 토대로 토픽 클러스터링(주제분석)이 가능함

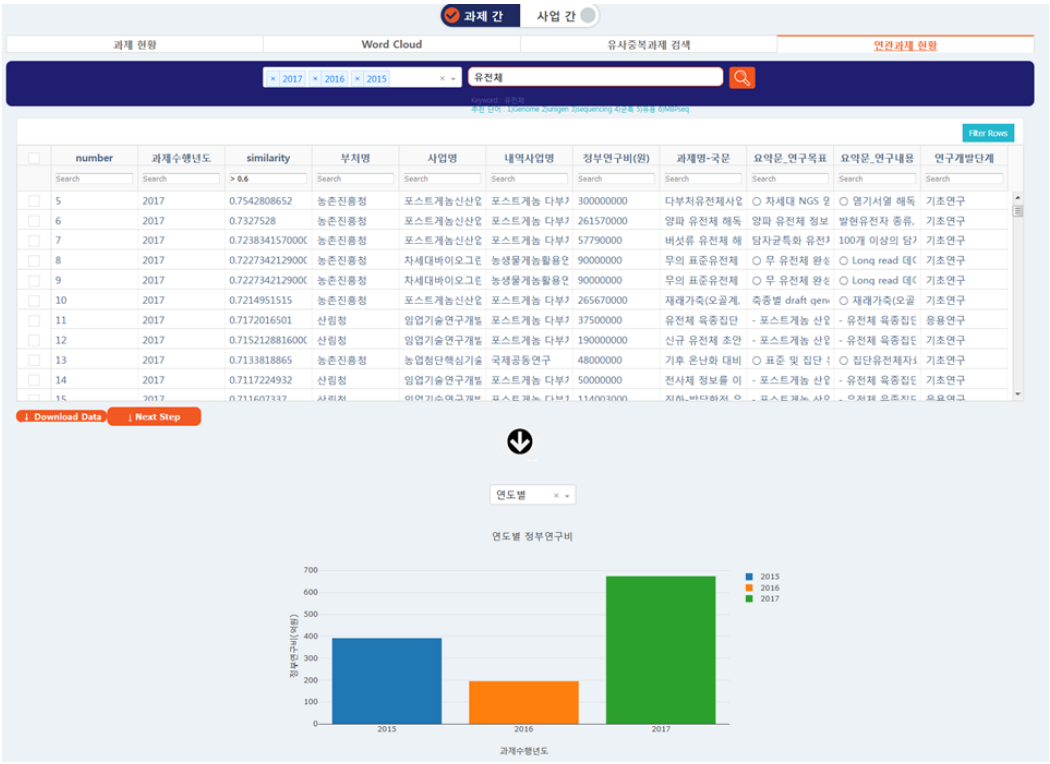


[그림 4-23] 분석 활용모형을 활용하여 암유전체 유사과제를 분석한 결과

[3. 관계성(연관) 과제 분석]

□ doc2vec 기반의 연관된 혹은 관련된 연구과제분석은 다음과 같은 방식으로 수행할 수 있음

o 연관과제 현황 모듈에서 검색대상연도 및 관심 분야의 키워드나 연구내용 등을 입력, 전술된 제안 방식을 토대로 코사인유사도 cut-off를 지정하고 통계 분석 등을 수행함



[그림 4-24] 분석 활용모형을 활용하여 유전체 연관 과제를 분석한 결과

o 유전체 단일 키워드 검색의 경우와 word2vec을 통해 제시된 관련 키워드 중 전장유전체, SNP를 추가적으로 입력할 경우 유사도 및 제시되는 과제의 차이가 발생함을 확인할 수 있었음

o 그러나 입력 키워드와 연관성이 높은 일부 과제가 낮은 유사도 값을 할당 받아 하위 랭킹에 제시되는 등의 문제가 발견되어 보다 정밀도를 제고할 수 있는 필터링 방안이 강구될 필요가 있다고 판단됨

- 가령, GWAS, SNP 관련 연구의 경우 유전체 분석 기본에 해당되므로 활용범위가

넓어 분야를 보다 특정하기 위해서는 특이적인 예시 발굴을 통한 추가적인 분석이 필요함을 확인할 수 있었음

- 필터링 결과에 따라 통계분석의 결과 또한 달라지기 때문에 정밀도 측면에서 개선이 요구됨을 확인할 수 있었음

〈표 4-2〉 검색어 차이에 따른 출력결과의 변화 예시

구분	유사도	부처명	사업명	과제명
검색어: 유전체	0.763983	농촌진흥청	포스트게놈신산업육성을위한다부처유전체사업(농진청)	국화 구조유전체 해독 및 정보분석
	0.759122	농촌진흥청	포스트게놈신산업육성을위한다부처유전체사업(농진청)	왕지네 신규 유전체 해독 및 조립
	0.754281	농촌진흥청	포스트게놈신산업육성을위한다부처유전체사업(농진청)	다부처유전체사업 NGS 분석 및 정보분석 지원
	0.732753	농촌진흥청	포스트게놈신산업육성을위한다부처유전체사업(농진청)	양파 유전체 해독
	0.723834	농촌진흥청	포스트게놈신산업육성을위한다부처유전체사업(농진청)	버섯류 유전체 해독에 특화된 유전체조립 스캐폴딩 기술과 해독 파이프라인 개발
	0.722734	농촌진흥청	차세대바이오그린21	무의 표준유전체 정밀화 및 유전체 활용 인터페이스 구축
	0.721495	농촌진흥청	포스트게놈신산업육성을위한다부처유전체사업(농진청)	재래가축(오골계, 진돗개) 표준 유전체지도 작성 및 특성 규명
검색어: 유전체, 전장유전체, SNP	0.70716	농촌진흥청	작물시험연구	육종가 친화형 유전체 변이 관련 인터페이스 개발
	0.702073	농촌진흥청	작물시험연구	주요 콩 유전자원의 전장유전체 재분석
	0.693935	농촌진흥청	포스트게놈신산업육성을위한다부처유전체사업(농진청)	다부처유전체사업 NGS 분석 및 정보분석 지원
	0.688697	농촌진흥청	차세대바이오그린21	유전체육종가 정확도 개선을 위한 프로그래밍
	0.681117	농촌진흥청	차세대바이오그린21	한국형 콩 haplotype map-GWAS 통한 유용 유전자 발굴
	0.677541	농촌진흥청	차세대바이오그린21	SNP칩 정보를 활용 친자 및 반형매감별, MS동일성 감정마커연계 및 유전질병검진 기술 개발
	0.67688	농촌진흥청	축산시험연구	한우 집단의 기능변이정보를 활용한 형질 예측 연구
	0.675991	농촌진흥청	차세대바이오그린21	벼 유전체 정보를 활용한 '백일미'의 극조숙 관련 유전자위 규명

2) 정부 연구개발 사업 관계성 분석 모형

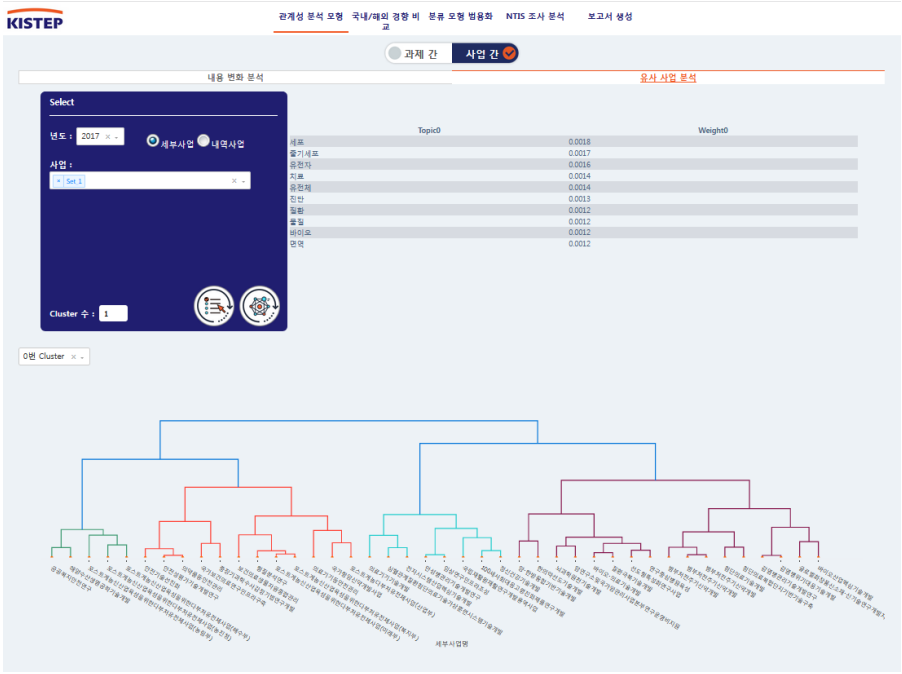
[1. 사업 간 관계성(유사정도) 분석]

□ 정부 연구개발사업 간 관계성(유사정도) 분석은 다음과 같은 방식으로 수행할 수 있음

- o 사업간 유사사업 분석 모듈에서 분석하고 하는 해당 연도 입력 후 세부사업/내역사업 중 분석 사업 수준을 선택
- o 다음으로 분석 대상 사업들을 고르고 희망하는 클러스터의 수를 입력하면 클러스터별 토픽클러스터링 결과와 함께 사업의 관계성*을 계층적 클러스터링(Hierarchical Clustering) 방법론으로 표현함

* 여기서 관계성은 덴드로그램 상의 거리로 파악할 수 있음. 거리상에, 계층적으로 가까울 경우 유사한 사업이며, 멀리 떨어져있을 경우 관련이 적은 사업임

- 동 연구의 접근방법은 세부/내역사업에 포함된 연구 과제를 딥러닝(doc2vec)을 통해 벡터화하고 특정 세부/내역사업이 포함하는 연구과제들의 중심 값을 해당 세부/내역사업의 값으로 할당하고 사업 간의 거리를 측정하여 관계성을 표현함



[그림 4-25] 사업 간 관계성(유사정도) 분석 수행 예시

□ 본 기능을 활용하여 17년도 기준 생명의료전문위 소관 생명보건의료분야의 사업을 바탕으로 분석대상사업(표 5-2 참고)들을 선정(출연연 등 제외)하고 관계성 분석을 진행하였음

- ㅇ 생명보건의료분야 및 농림수산식분야를 구분하여 진행하였으며, 1차적으로는 전체 세부사업을 기준으로 계층 클러스터링을 적용하여 사업 간의 관계성을 파악하고,
- ㅇ 2차적으로는 내역사업을 기준으로 토픽 클러스터링을 적용 후 클러스터 별로 포함된 내역사업의 유사성/차별성을 계층 클러스터링 방법으로 분석해보았음

〈표 4-3〉 사업간 관계성 분석 대상사업

구분	부처명	사업명
생명보건의료분야	과학기술정보통신부	공공복지안전연구
		뇌과학원천기술개발
		바이오·의료기술개발
	보건복지부	100세사회대응고령친화제품연구개발
		감염병관리기술개발연구
		감염병위기대응기술개발
		국가보건의료연구인프라구축
		국가항암신약개발사업
		국립재활원재활연구개발용역사업
		글로벌화장품신소재·신기술연구개발지원
		만성병관리기술개발연구
		보건의료생물자원종합관리
		선도형특성화연구사업
		심혈관계질환첨단의료기술가상훈련시스템기술개발
		암연구소및국가암관리사업본부연구운영비지원
		양·한방융합기반기술개발
		연구중심병원육성
		의료기기기술개발
		임상연구인프라조성
		정신건강기술개발
		질환극복기술개발
		첨단의료기술개발
		첨단바이오의약품글로벌진출사업
		첨단의료복합단지기반기술구축
		한의약선도기술개발

구분	부처명	사업명
		형질분석연구
	산업통상자원부	바이오산업핵심기술개발
		전자시스템산업핵심기술개발
	식품의약품안전처	안전기술선진화
		안전성평가기술개발연구
		의료기기등안전관리
		의약품등안전관리
	해양수산부	해양수산생명공학기술개발
	행정안전부	중장기과학수사감정기법연구개발
	다부처 사업	범부처전주기신약개발
		포스트게놈신산업육성을위한다부처유전체사업

□ 생명보건의료분야 전체 세부/내역사업을 기준으로 상당 수준의 관계성이 분석되는 것을 확인할 수 있었음

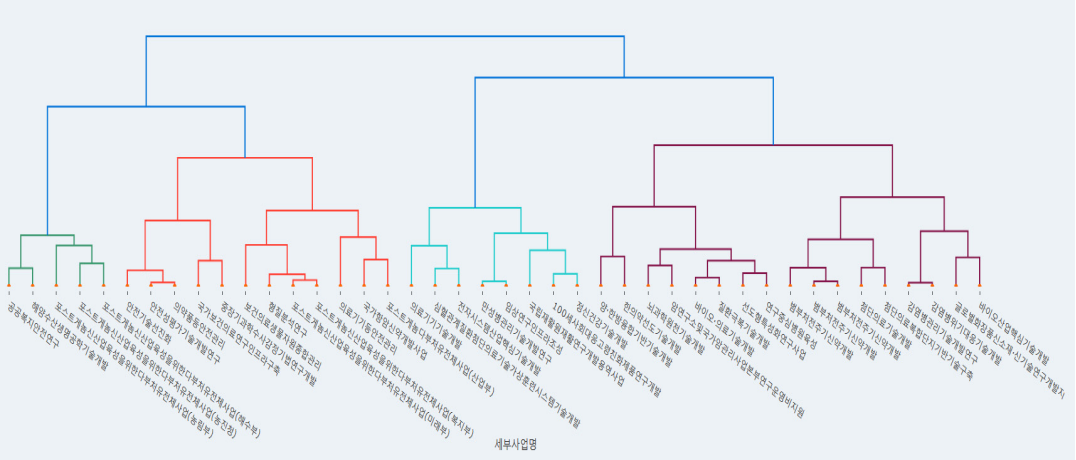
o 클러스터를 1개로 지정하고 모든 생명보건의료분야 세부사업간의 관계성을 분석하였을 때, 키워드는 다음과 같이 제시됨

- 생명보건의료분야는 신약/의료기기/줄기세포/유전체/뇌과학/바이오융복합/임상보건의분야로 세분화할 수 있는데, “물질”, “진단”, “줄기세포”, “유전체”, “면역”, “치료” 등 세부기술분야 연관된 유의미한 키워드가 도출됨

Topic0	Weight0
세포	0.0018
줄기세포	0.0017
유전자	0.0016
치료	0.0014
유전체	0.0014
진단	0.0013
질환	0.0012
물질	0.0012
바이오	0.0012
면역	0.0012

[그림 4-26] 클러스터를 1개로 가정하였을 때 생명보건의료분야 사업의 주요 토픽(키워드)

o 세부사업 수준에서의 덴드로그램 확인 시 [그림 4-27]과 같이 유사한 성격을 지닌 사업을 중심으로 클러스터링이 적절히 이루어짐을 확인할 수 있었음



[그림 4-27] 17년도 생명보건의료분야 세부사업수준 관계성 분석 결과

- 포스트게놈다부처유전체사업의 경우 인간 대상의 과기정통부/복지부/산업부의 사업과 비인간을 대상으로 하는 농식품부/농진청/해수부 사업 간에 구분되었고 형질분석연구와 같이 유전체 연구를 하는 사업은 인간 대상의 포스트게놈다부처 유전체사업(과기정통부/복지부)과 거리가 가까운 것으로 분석됨
- 의료기기기술개발, 전자시스템산업핵심기술개발 등 의료기기 관련 사업들 간의 계층이 가깝게 표현되었으며, 선도형특성화연구사업과 유사중복 이슈가 꾸준히 제기된 연구중심병육성사업은 서로 간에 가장 가까운 것으로 분석됨
- 이외에 고령·재활·정신건강과 관련된 사업들도 가까운 거리에 배치되었으며, 감염병 관리기술개발·감염병 위기대응기술개발도 서로 간에 가장 가까운 것으로 분석됨

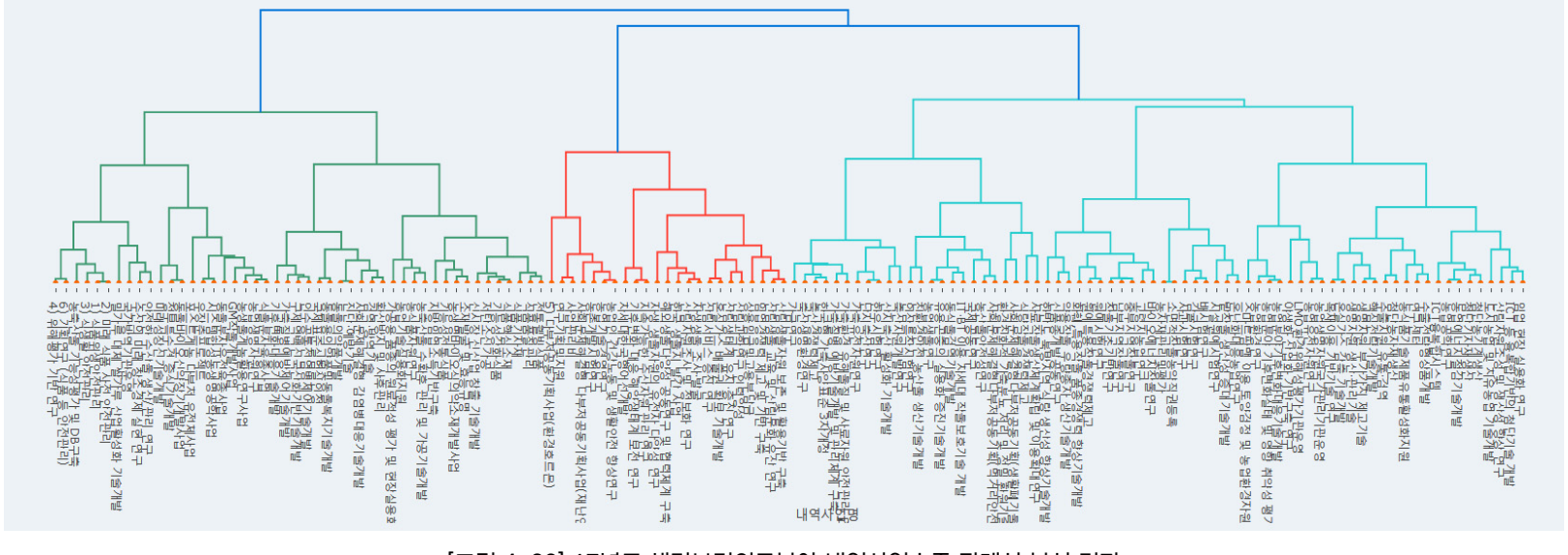
o 생명보건의료분야는 7개의 세부기술분야로 구성되므로 7개의 클러스터로 구분해 본 결과, 세부기술분야의 키워드가 적절히 도출되더라도 두 개 이상의 세부기술분야가 한 클러스터에 속하는 경우가 있었음

Topic0	Weight0	Topic1	Weight1	Topic2	Weight2	Topic3	Weight3	Topic4	Weight4	Topic5	Weight5	Topic6	Weight6
줄기세포	0.0017	임상시험	0.0011	표준	0.002	줄기세포	0.0016	유전체	0.0018	해양	0.0047	백신	0.0018
시험	0.0013	치료	0.0011	줄기세포	0.0019	세포	0.0016	의료기기	0.0017	and	0.0034	진단	0.0015
세포	0.0012	시스템	0.0009	생약	0.0017	유전자	0.0015	표준	0.0014	of	0.0032	결핵	0.0015
임상시험	0.0012	영상	0.0009	독성	0.0016	진단	0.0012	약물	0.0014	유전체	0.0031	바이러스	0.0014
독성	0.001	재난	0.0008	시험	0.0016	치료	0.0012	품목	0.0013	자원	0.0027	항원	0.0012
후보	0.001	질환	0.0008	품목	0.0016	바이오	0.0012	평가	0.0013	산업	0.0027	면역	0.0012
생산	0.001	정신건강	0.0008	모델	0.0015	물질	0.0012	가이드라인	0.0012	생물	0.0026	항체	0.001
물질	0.0009	지원	0.0008	의약품	0.0015	영상	0.0011	진단	0.0012	소재	0.0025	후보	0.001
평가	0.0009	의료기기	0.0008	항체	0.0014	질환	0.0011	시험	0.0012	바이오	0.0024	물질	0.001
면역	0.0009	환자	0.0008	성분	0.0013	조절	0.001	유전자	0.0012	정보	0.0024	감염	0.0009

[그림 4-28] 17년도 생명보건의료분야 세부사업을 7개의 클러스터로 구분한 결과

- 이는 세부기술 간의 중요성을 정책적 판단에 의해 구분한 결과로 사료되며, 이러한 점을 반영하여 클러스터링을 하기 위해서는 별도의 지도학습이 필요할 것으로 사료됨
- 가령, 어떤 사업이나 어떠한 키워드는 신약, 의료기기 등에 해당한다는 정보를 학습시킬 경우 그러한 구분이 가능할 것으로 보임
- [그림 4-28]의 Topic5 결과에서 and, of 키워드의 가중치(weight)값이 높게 나왔는데, 이는 추가학습을 통해 개선이 필요한 부분으로 사료됨

- 내역사업 수준에서의 덴드로그램을 확인했을 경우에도 [그림 4-28]과 같이 유사한 성격을 지닌 내역사업을 중심으로 클러스터링이 적절히 이루어짐을 확인할 수 있었음
  - 가령, 한약제제 개발, 양한방 융합전임상연구, 양·한방 융합기반 기초연구, 한양방 협력치료기술개발 등 한의약 관련 내역간의 관계가 가장 가깝게 나타났으며
  - 줄기세포·재생의료실용화연구, 줄기세포연구사업, 줄기세포사회밀착형지원, 줄기세포은행운영및표준화기반구축, 줄기세포연구·산업화 촉진지원, 조직재생기술개발사업, 세포재생기술개발사업 등 줄기세포분야 내역사업들이 군집화 됨을 확인할 수 있었음



[그림 4-29] 17년도 생명보건의료분야 내역사업수준 관계성 분석 결과

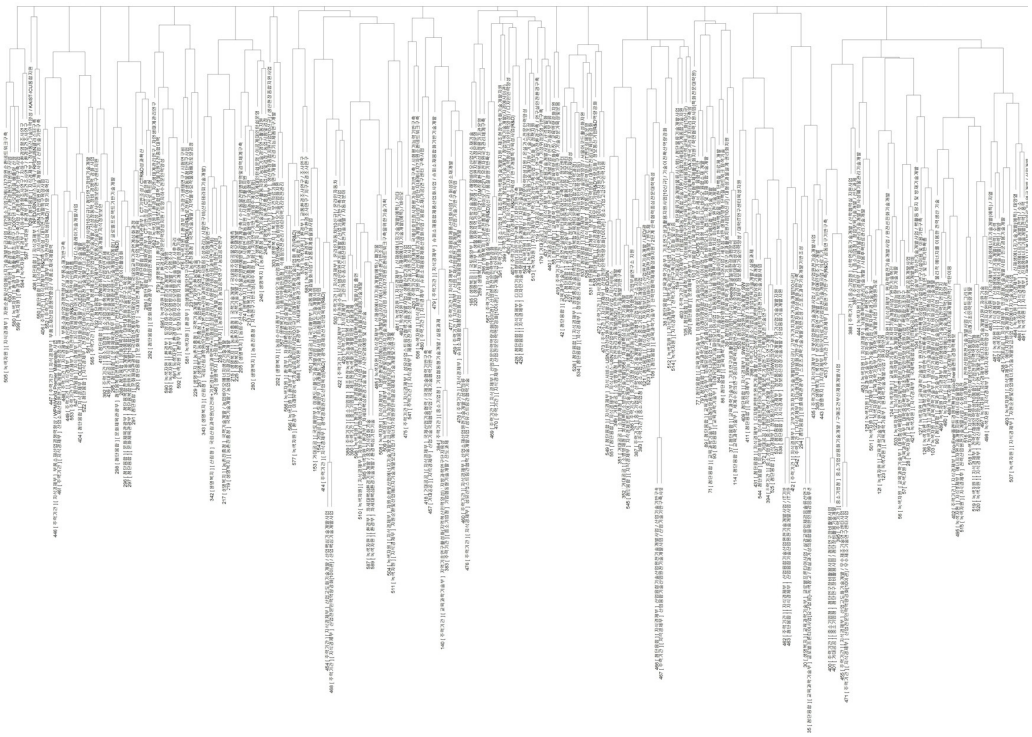
□ 정부 연구개발사업의 평가 등의 업무를 수행하는 KISTEP에서는 과거 정부 연구 개발사업 유사중복 검색 시스템을 개발함[19)]

ㅇ 해당 과제의 경우 동 사업과 같이 정부 연구 개발사업 내 과제에서 특징점을 파악하고 이를 바탕으로 세부/내역사업 유사중복 알고리즘을 개발하였음

- 기술분야, 사업목적, 연구 개발단계, 연구수행주체, 사업추진체계 등의 검토기준 항목을 토대로 사업 간의 유사중복 시스템을 개발함

ㅇ 기본적으로 해당 시스템은 사업과 관련 있는 키워드의 빈도 개념을 도입하여 유사중복시스템을 구축하였다는 점에서 벡터화 된 연구 과제를 활용한 동 분석·활용 모형과 작동 방식에서 차이가 있음

- 클러스터링 결과물의 취득은 사용자의 의도가 일부 반영될 수 있기 때문에 다른 방법론을 적용하였을 경우 결과가 다를 수 있음



[그림 4-30] 정부연구개발 내역사업간 유사중복 검색 시스템 결과 예시

※ 출처: 홍세호 (2013),「국가연구개발사업 유사중복 검색 시스템 개발을 위한 실증연구」, 한국과학기술기획평가원.

19) 홍세호 (2013),「국가연구개발사업 유사중복 검색 시스템 개발을 위한 실증연구」, 한국과학기술기획평가원.

o 해당 시스템 역시 상당히 유의미한 결과를 도출하고 있어 사용자 입장에서 본인의 활용목적에 부합하는 선택을 하는 것이 바람직 할 것으로 사료됨

- 가령, 동 분석 활용모형의 경우 딥러닝 기반의 학습 및 작동과정에서 컴퓨터 리소스가 상당히 요구될 수 있음

[2. 특정 사업 속성 변화(사업내용 변화) 분석모형]

□ 특정 사업의 속성 변화 분석은 다음과 같은 방식으로 수행할 수 있음

o 사업 간 내용변화분석 모듈에서 분석하고자하는 사업명 및 비교연도를 선택하면 해당 사업의 과거연구과제와 최근연구과제가 공간상에 투영됨

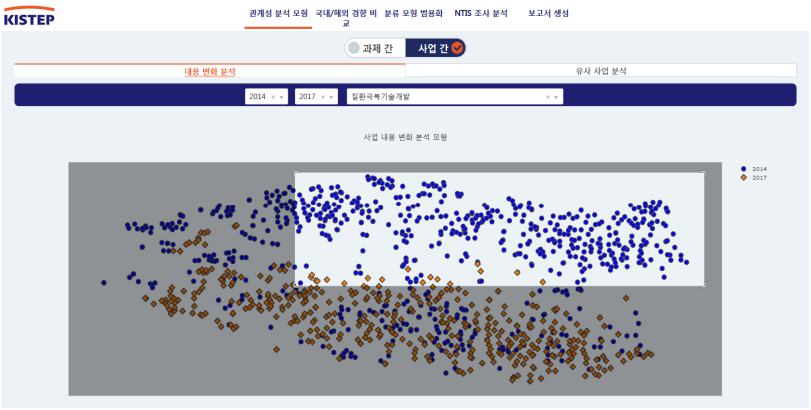
- 비교 연도 간 시각적인 구분(클러스터링이 명확히 구분)외에 사업의 중심 좌표 값 이동정도에 따른 정량적인 수치도 확인 가능

- 또한 사용자는 노드로 대표되는 연구과제 클릭 시 해당 과제의 개괄적 내용 파악이 가능함

[그림 4-31] 특정 사업 속성 변화분석 수행 예시

o 비교 연도의 과제 간 클러스터링이 명확히 이루어질 경우. 관심 있는 클러스터링 영역을 설정하여 해당 과제 군이 어떤 연구 분야로 구성되었는지 파악 가능

- 해당 군집에 토픽클러스터링을 수행하여 빈도가 높은 키워드를 도출함



[그림 4-32] 사업 내 특정 과제군의 연구분야 탐색 예시

□ 생명보건의료분야 대표사업인 바이오의료기술개발사업의 2011년-2017년 사이의 사업 속성 변화를 분석하였음

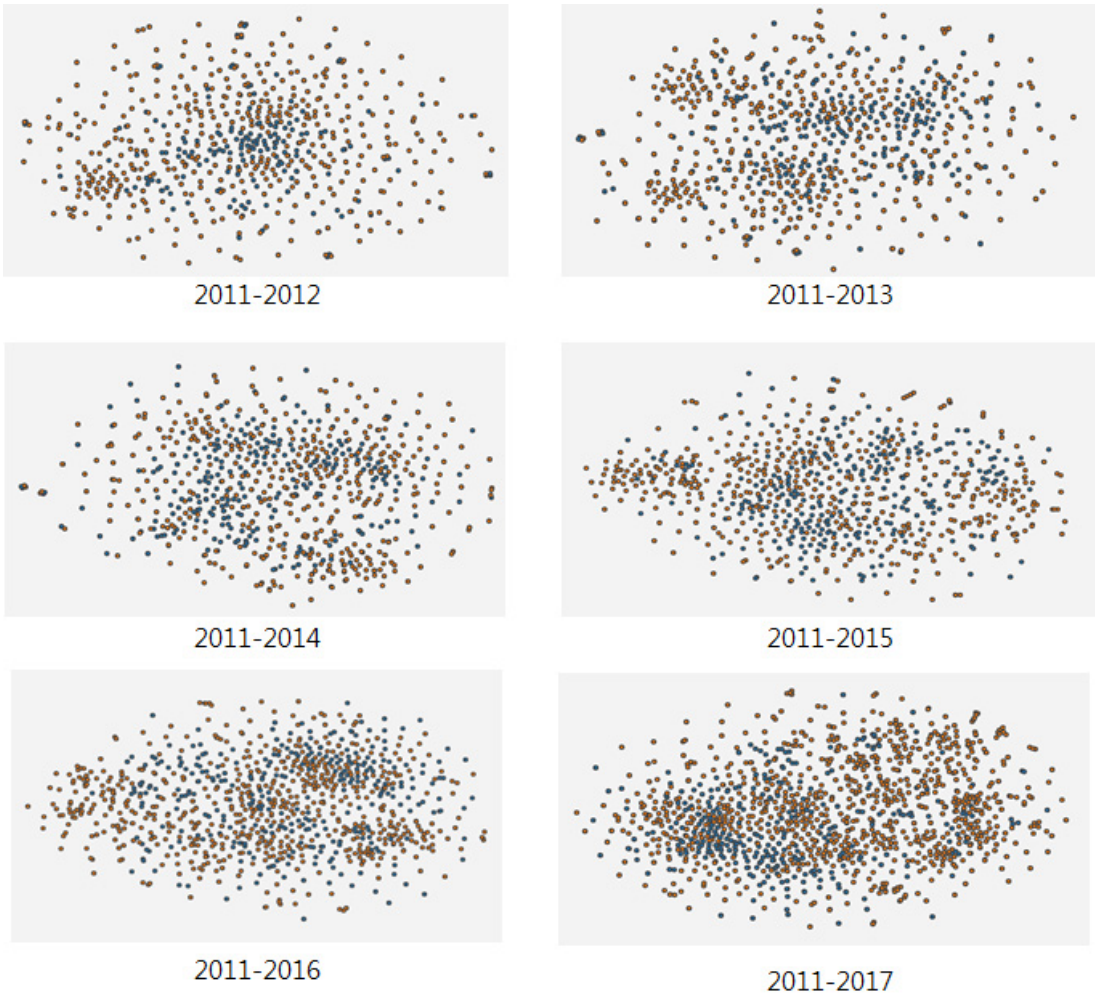
o 바이오의료기술개발사업은 16년 이전까지는 내역사업 개편이 크게 이루어지지 않았으나, 17년도에 신규 내역사업이 대폭 추가되었음

〈표 4-4〉 바이오의료기술개발사업 내역사업 개편 결과

16년 내역사업	17년 내역사업
신약개발분야	신약개발분야
차세대의료기술개발분야	차세대의료기술개발분야
줄기세포/조직재생분야	줄기세포/조직재생분야
차세대 바이오 분야	차세대 바이오 분야
바이오 인프라분야(연구소재포함)	바이오 인프라 분야(연구소재포함)
신약후보 물질발굴및최적화사업(16년 종료)	-
국가마우스표현형분석기반구축사업	국가마우스표현형분석기반구축사업
전통천연물기반유전자-동의보감사업	전통천연물기반유전자-동의보감사업
연구소재지원사업	연구소재지원사업
-	신시장창조차세대의료기기 개발사업
-	첨단바이오의약품글로벌진출사업
-	미래감염병기술개발
-	바이오융복합기술개발
-	미래의료혁신대응기술개발
-	첨단GW바이오

ㅇ 바이오의료기술개발사업의 2011년 사업을 기준으로 2017년까지 과제의 공간 상 배치를 분석해보면 2016년까지는 큰 차이를 보이지 않았으나 2017년에는 시각적인 차이를 보이고 있음

- 2011년 사업들은 좌하단 부에 주로 배치되고 2017년 사업은 그 외의 공간에 넓게 분포함을 확인할 수 있었음



[그림 4-33] 바이오의료기술개발사업 속성 변화 분석 결과

ㅇ 실제 연도별 과제의 좌표 평균값을 비교해본결과 2011년-2017년 사이의 좌표평균값의 이동이 가장 크게 나타남

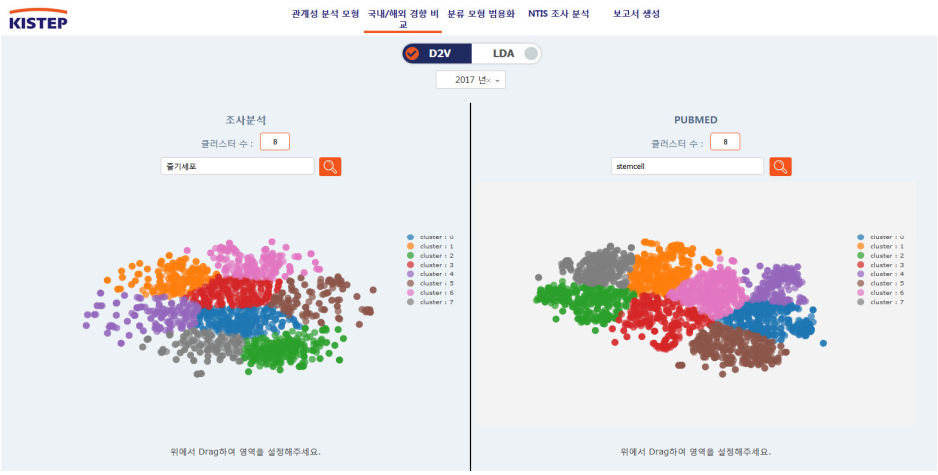
	X_Coord	Y_Coord	Euclidean Distance
2011	2.76	-0.17	5.190
2012	-2.43	-0.21	
2011	4.99	-0.27	7.889
2013	-2.85	0.61	
2011	-3.1	0.97	5.461
2014	0.86	-2.79	
2011	0.78	0.86	1.324
2015	-0.29	0.08	
2011	3.04	1.42	6.615
2016	-3.12	-0.99	
2011	-7.09	-4.54	13.224
2017	3.69	3.12	

[그림 4-34] 바이오의료기술개발사업 속성 변화 분석 결과

나. 딥러닝 기반 토픽 클러스터링 분석 방법론 연구

□ 딥러닝 기반 토픽 클러스터링은 다음과 같은 환경에서 수행됨

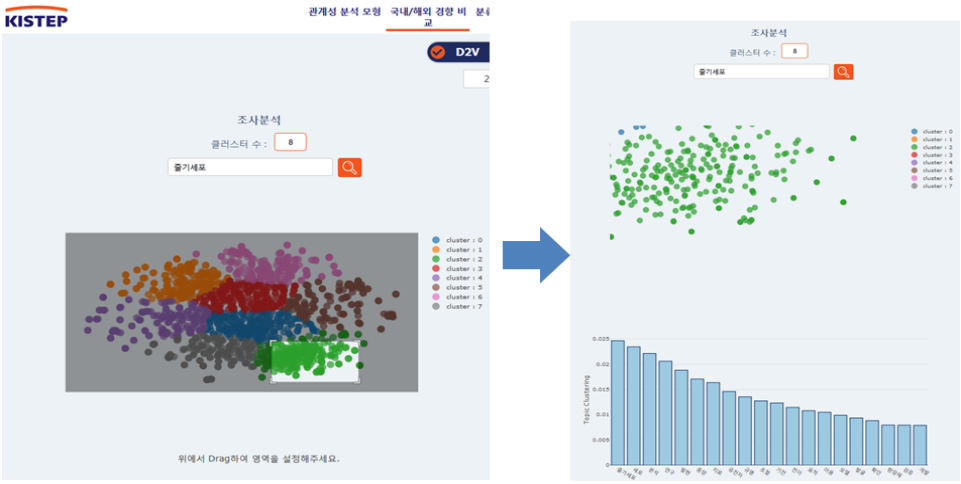
○ [그림 4-35]와 같이 국내/해외 경향비교 모듈에서 분석 연도를 선택 후 조사 분석 데이터(좌), Pubmed 문헌정보(우)에 공통의 키워드를 입력하고 구분하려는 클러스터링 개수를 입력하면 클러스터링 결과가 도출됨



[그림 4-35] 딥러닝 기반 클러스터링 수행 예시

○ 관심 있는 클러스터를 마우스로 드래그 시 토픽(키워드) 분석을 수행할 수 있음

- 또한, 특정 노드를 선택 시 해당 연구과제나 논문 초록정도를 확인할 수 있음



[그림 4-36] 딥러닝 기반 클러스터링 수행 후 토픽(키워드) 발굴 예시

□ 이와 같은 분석모형의 기능을 바탕으로 Pubmed 문헌정보(논문, 서적 등)와 국내 조사 분석 과제의 토픽 클러스터링(Topic clustering)을 수행하여 연구트렌드의 변화 및 국내·외 차이를 분석함

ㅇ 동 연구에서는 일반적인 토픽 클러스터링 알고리즘인 LDA(Latent Dirichlet allocation)가 아닌 딥러닝 알고리즘 doc2vec을 사용하였음

- 텍스트 임베딩 알고리즘인 doc2vec은 글로 이루어진 과제 및 문헌정보를 숫자로 이루어진 다차원 벡터로 변환할 수 있음

ㅇ Doc2vec 알고리즘을 이용하여 조사 분석 과제 또는 Pubmed 문헌정보 각각을 300차원 벡터로 변환한 후, t-SNE(t-distributed stochastic neighbor embedding) 알고리즘을 사용하여 사람이 공간적으로 해석 가능한 2차원 벡터로 축소함

- t-SNE 역시 기계학습 기반 임베딩 알고리즘으로, 데이터의 시각화에 주로 사용됨

ㅇ 2차원으로 축소된 데이터를 k-means 알고리즘을 사용하여 그룹화하고, 그룹 내 자주 등장하는 키워드 및 과제(논문)명을 연구자가 읽고 그룹의 주제를 판단함

- k-means 알고리즘을 통하여 2차원 공간상에서 서로 가까이 있는 벡터들을 그룹화(clustering)했으며, 그룹의 수는 6~10 사이에서 최적을 선택함

□ 본 연구에서는 doc2vec을 토픽 클러스터링에 적용하는 방법론을 제안하며, 방법론의 유효성을 판단하기 위해 '유전체' 분야의 국내과제 및 Pubmed 문헌을 대상으로 분석을 실시함

ㅇ 전체 분야를 대상으로 할 경우 광범위함에 따라 지나치게 많은 수의 그룹이 필요할 수 있고 연도별 차이가 드러나지 않을 수 있어 최근 활발히 연구되고 있는 '유전체' 분야에 한정하였음

- 국내 과제의 경우 과제명, 연구목적, 연구내용, 기대효과에 "유전체"라는 단어를 포함하는 과제를 입력으로 사용

- Pubmed 논문의 경우 제목(article title), 초록(abstract text), MeSH 키워드(MeSH Heading), 키워드(keyword)에 "genome", "genomics"를 포함하는 문헌을 입력으로 사용

□ 2017년 국내 과제와 Pubmed 문헌에 해당 방법론을 적용하기 위해, 클러스터(그룹)의 수를 최적화하는 과정을 거쳤으며, 최적 그룹 수는 8개였음

ㅇ 그룹의 수가 지나치게 적을 경우, 각 그룹에 포함된 내용이 광범위하게 되어 주제(토픽)를 특정하기 어렵게 됨

ㅇ 반면에 그룹의 수가 너무 많으면 그룹이 불필요하게 세분화되어 내용적으로 구분되지 않는 그룹들이 발생하게 됨

ㅇ 동 연구에서는 2017년 데이터에 대하여 그룹의 수를 6, 8, 10개로 설정하여 시험하였고, 8개로 설정하였을 때 가장 적절한 수준에서 클러스터링이 이루어짐이 확인되어 이 조건을 2012년 데이터에도 적용함

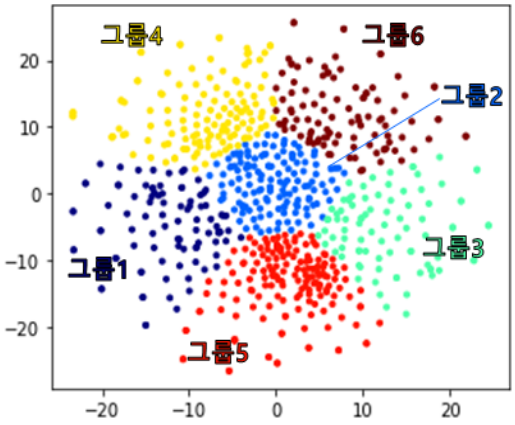
- 그룹의 수가 너무 많을 경우에도 오히려 주제를 특정하기 어려운 그룹이 발생하여, 토픽 클러스터링에서 그룹 수 설정은 매우 중요함이 드러남

ㅇ (국내 과제 그룹 수 최적화 결과) 8개 그룹으로 나누었을 때 가장 그룹별 특징이 뚜렷한 것으로 나타났음

- 6개 그룹으로 나눌 경우 그룹 내 과제의 내용이 너무 다양하거나(그룹1), 그룹의 특징이 서로 유사하게 나타나(그룹2, 3) 적절히 분류되지 못함

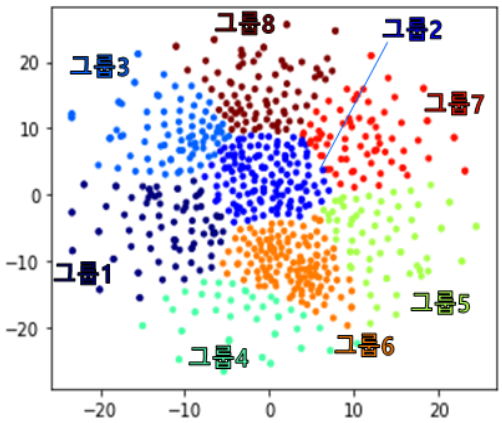
- 10개 그룹으로 나누었을 때는 유사한 과제들을 포함하는 그룹들이 증가하여(그룹3, 9 및 그룹4, 6 등) 최적으로 볼 수 없음

- 유사한 그룹들의 경우 공간상에서 서로 붙어있는 것이 드러남



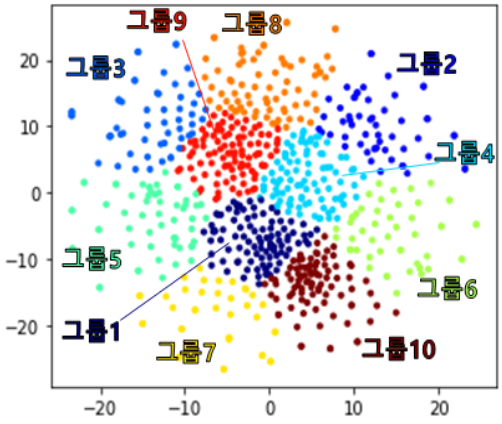
그룹	주제
그룹1	분야특정 불가능
그룹2	유전자교정 활용 육종/치료기술
그룹3	유전자교정 활용 육종/치료기술
그룹4	유전정보 분석기술
그룹5	생물자원 확보/보존관리
그룹6	기초연구(유전자마커 발굴 등)

[그림 4-37] 2017년 유전체분야 국내 조사·분석 과제 클러스터링 결과(그룹 수: 6개)



그룹	주제
그룹1	유전자변형 생물체 개발 및 위해성평가
그룹2	유전자교정 활용 육종/치료기술
그룹3	유전정보 분석기술
그룹4	생물자원 보존 및 다양성 연구
그룹5	유전자치료제 기술
그룹6	생물자원 확보/수집
그룹7	유전학 및 기초기전연구
그룹8	질병치료 응용기술(약리유전체학 등)

[그림 4-38] 2017년 유전체분야 국내 조사·분석 과제 클러스터링 결과(그룹 수: 8개)



그룹	주제
그룹1	생물자원 확보/보존관리 및 유전자변형 생물체 위해성평가
그룹2	유전자치료 기반기술(후생유전학 등)
그룹3	유전정보 분석기술 및 기초기전연구
그룹4	유전자교정 활용 육종/치료기술
그룹5	분야특정 불가능
그룹6	유전자교정 활용 육종/치료기술
그룹7	생물자원 확보/보존관리 및 기초기전연구
그룹8	유전정보 분석기술
그룹9	유전정보 분석기술 및 기초기전연구
그룹10	생물자원 확보/보존관리

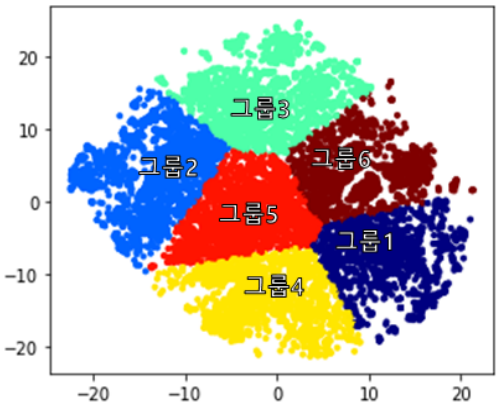
[그림 4-39] 2017년 유전체분야 국내 조사·분석 과제 클러스터링 결과(그룹 수: 10개)

○ (Pubmed 문헌 그룹 수 최적화) 국내 과제와 동일하게 8개의 그룹으로 분류하는 것이 최적으로 나타남

- 6개 그룹으로 분류하는 경우 그룹 하나의 범위가 광범위해짐에 따라 차별성이 떨어지는 결과를 초래함(그룹3, 5)

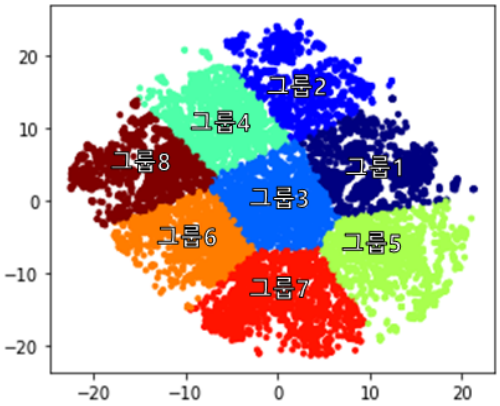
- 10개 그룹으로 분류하면 보다 세부적인 연구 분야로 분류되는 것이 확인되나, 오히려 특징을 파악할 수 없는 그룹이 발생함(그룹8)

※ 그룹8은 8개 그룹으로 분류하였을 때의 그룹 3, 5, 7이 서로 만나는 부분에 위치하고 있어, 임베딩 벡터 좌표 상으로는 가깝지만 실제 내용은 혼재되어 있는 것으로 사료됨



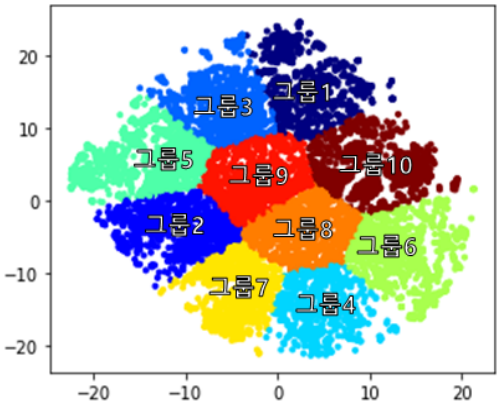
그룹	주제
그룹1	암 발생기전 및 연관성 연구
그룹2	계통학 및 유전자다양성 연구
그룹3	미생물 유전체 해독
그룹4	약리유전체학 및 연관성 연구
그룹5	미생물 유전체 해독
그룹6	유전자교정 관련연구

[그림 4-40] 2017년 유전체분야 Pubmed 문헌 클러스터링 결과(그룹 수: 6개)



그룹	주제
그룹1	유전자교정(CRISPR) 연구
그룹2	유전체 해독(시퀀싱)
그룹3	유전자-형질 연관성 연구
그룹4	계통유전체학
그룹5	암 관련 연관성 연구(후생유전학)
그룹6	유전 및 진화 관련 연구
그룹7	정밀의료 기반연구
그룹8	기초기전 및 유전자다양성 연구

[그림 4-41] 2017년 유전체분야 Pubmed 문헌 클러스터링 결과(그룹 수: 8개)



그룹	주제
그룹1	미생물 유전체 해독
그룹2	유전학 및 진화 관련 연구
그룹3	군집유전체학
그룹4	유전자-형질 연관성 연구
그룹5	작물 관련 연관성 연구
그룹6	암 발생기전 및 치료법 연구
그룹7	정밀의료 기반연구
그룹8	분야특정 불가능
그룹9	계통학 및 유전학
그룹10	유전자편집(CRISPR) 연구

[그림 4-42] 2017년 유전체분야 Pubmed 문헌 클러스터링 결과(그룹 수: 10개)

□ (국내과제와 Pubmed 문헌 특성 비교) 동일한 조건으로 8개 그룹으로 분류할 시, 국내과제는 주로 연구목적을 중심으로 분류된 반면 Pubmed 문헌은 학문분야 위주로 분류되는 경향을 보여 차이가 드러남

- o 국내과제는 학습 내용에 연구목적 및 기대효과가 포함되므로 “기반 구축”, “자원 확보”, “치료제 개발”과 같은 표현이 많이 사용된 것으로 보이며, doc2vec 토픽 클러스터링 결과로 목적지향적인 그룹핑이 나타남
- o Pubmed 문헌의 경우 대부분이 논문이라 제목과 초록(abstract)에 학문적인 표현이 많이 포함되었기 때문에 “sequence”(시퀀싱), “association”(연관성 연구)과 같은 연구방법론 관련 키워드가 상위에 등장함
- o 각 그룹에 주제를 부여할 시, 상위 키워드만을 가지고 주제를 부여하기에는 공통적으로 등장하는 일반적 단어가 많아 무리가 있으며, 실제로 포함된 과제(논문)의 제목을 근거로 판단하였음
- o 국내과제는 유전자원 기탁등록(자원은행) 과제라거나 유전자변형 생물 위해성평가와 같은 목적성이 뚜렷한 과제들을 중심으로 그룹화된 사례가 많은 반면, Pubmed 문헌은 계통유전학(phylogenomics), 진화(evolution) 같은 학문(연구) 테마를 중심으로 그룹화 됨
- o 이는 현재의 클러스터링 방법론만으로는 국내과제와 Pubmed 문헌을 1:1로 비교하기에는 한계가 있음을 시사하지만, 연구과제와 논문의 성질을 기계학습 방법론이 잘 반영하고 있다고 볼 수 있음

〈표 4-5〉 2017년 유전체분야 국내 연구과제 그룹별 상위 키워드 및 출현 횟수(20개)

그룹1 유전자변형 생물체 개발 및 해성평가	그룹2 유전자교정 활용 육종/ 치료기술	그룹3 유전정보 분석기술	그룹4 생물자원 보존 및 다양성 연구	그룹5 유전자치료제 기술	그룹6 생물자원 확보/수집	그룹7 유전학 및 기초기전연구	그룹8 질병치료 응용기술(약리 유전체학 등)
분석(466)	분석(581)	분석(753)	분석(214)	치료(381)	보존(295)	조절(340)	분석(517)
개발(377)	개발(460)	개발(447)	개발(169)	개발(322)	자원(272)	분석(290)	개발(381)
기술(327)	기술(313)	기술(284)	자원(142)	기술(230)	개발(272)	개발(267)	치료(224)
마우스(234)	이용(282)	데이터(241)	활용(108)	분석(222)	특성(262)	기술(185)	약물(191)
변형(196)	정보(217)	기반(224)	확보(108)	세포(210)	수집(250)	세포(180)	환자(190)
이용(170)	기반(175)	구축(210)	생물(105)	기반(144)	분석(243)	유전(167)	기반(189)

그룹1 유전자변형 생물체 개발 및 해성평가	그룹2 유전자교정 활용 육종/치료기술	그룹3 유전정보 분석기술	그룹4 생물자원 보존 및 다양성 연구	그룹5 유전자치료제 기술	그룹6 생물자원 확보/수집	그룹7 유전학 및 기초기전연구	그룹8 질병치료 응용기술(약리 유전체학 등)
대한(170)	발현(172)	정보(204)	기술(93)	이용(142)	관리(236)	발현(160)	데이터(174)
식품(168)	치료(168)	활용(179)	비료(90)	대한(128)	평가(206)	기능(153)	발굴(165)
표현형(158)	통한(165)	시스템(164)	치료(90)	전달(113)	확보(198)	모델(148)	이용(153)
질환(154)	활용(159)	통한(123)	유전(85)	활용(113)	활용(183)	단백질(147)	정보(144)
통한(142)	세포(155)	질환(121)	정보(83)	평가(112)	증식(162)	기전(144)	기술(137)
치료(128)	단백질(151)	진단(116)	평가(82)	통한(110)	이용(159)	대한(140)	모델(132)
특성(117)	기능(148)	통합(110)	기능(81)	발현(110)	구축(156)	규명(137)	검증(127)
유전(117)	구축(146)	제공(108)	이용(80)	질환(108)	통한(154)	이용(129)	대한(127)
발현(114)	대한(141)	관련(107)	관리(78)	면역(105)	기관(151)	분화(125)	예측(125)
기반(112)	시스템(138)	치료(105)	유방암(78)	종양(100)	기술(144)	활용(124)	구축(125)
확보(112)	방법(135)	대한(104)	식물(72)	유전(98)	대한(128)	통한(122)	세포(122)
활용(111)	확보(132)	이용(102)	구축(71)	효과(98)	치료(126)	발굴(122)	활용(119)
정보(109)	모델(126)	표준(92)	미생물(70)	검증(97)	다양성(123)	변화(115)	질환(104)
가능(107)	질환(126)	확보(91)	통한(65)	단백질(97)	조사(112)	후성(113)	진단(100)

※ “유전자”, “유전체”, “통해”, “연구” 네 단어는 의미가 낮은 단어로 제외함

〈표 4-6〉 2017년 유전체분야 Pubmed 문헌 그룹별 상위 키워드 및 출현 횟수(20개)

그룹1 유전자교정 (CRISPR) 연구	그룹2 유전체 해독 (시퀀싱)	그룹3 유전자-형질 연관성 연구	그룹4 계통 유전체학	그룹5 암 관련 연관성 연구(후생유전학)	그룹6 유전 및 진화 관련 연구	그룹7 정밀의료 기반연구	그룹8 기초기전 및 유전자다양성 연구
cell (2809)	sequence (1929)	cell (1790)	analysis (1548)	cancer (3060)	data (2125)	disease (1784)	analysis (1603)
dna (2605)	virus (1472)	dna (1627)	plant (1404)	cell (2385)	analysis (1580)	association (1630)	sequence (1464)
protein (1269)	strain (1338)	sequence (1593)	sequence (1380)	tumor (1422)	population (1305)	analysis (1292)	plant (1455)
replication (935)	analysis (1259)	expression (1526)	expression (917)	expression (1343)	sequence (1241)	nan (1083)	species (1328)
crispr (908)	disease (752)	analysis (1292)	protein (799)	analysis (1212)	model (961)	data (943)	dna (1109)
repair (868)	infection (683)	protein (1280)	strain (656)	mutation (1178)	method (955)	risk (937)	evolution (956)
model (710)	resistance (676)	rna (1271)	family (630)	disease (1035)	dna (882)	genomics (897)	chromosome (903)

그룹1 유전자교정 (CRISPR) 연구	그룹2 유전체 해독 (시퀀싱)	그룹3 유전자-형질 연관성 연구	그룹4 계통 유전체학	그룹5 암 관련 연관성 연구(후생유 전학)	그룹6 유전 및 진화 관련 연구	그룹7 정밀의료 기반연구	그룹8 기초기전 및 유전자다양성 연구
cas9 (697)	dna (514)	study (1094)	genomics (623)	patient (1034)	evolution (855)	research (827)	region (855)
trait (664)	protein (508)	methylation (1025)	stress (512)	association (989)	genomics (844)	genetics (764)	trait (708)
system (662)	genomics (424)	data (804)	species (507)	dna (949)	selection (832)	expression (740)	marker (701)
expression (647)	isolates (419)	cancer (788)	cell (485)	treatment (758)	variation (687)	disorder (737)	protein (689)
sequence (641)	host (412)	genomics (761)	community (458)	risk (740)	species (665)	cancer (705)	expression (617)
chromatin (627)	pathogen (361)	disease (715)	evolution (455)	carcinoma (658)	approach (634)	variant (691)	genomics (611)
damage (603)	phage (338)	regulation (687)	host (450)	methylation (637)	association (610)	patient (687)	data (563)
transcription (594)	type (319)	biology (666)	data (446)	factor (609)	biology (573)	population (623)	family (539)
analysis (590)	species (313)	transcription (653)	role (418)	poly- morphism (607)	result (498)	health (619)	variation (518)
virus (583)	virulence (307)	development (652)	dna (414)	drug (586)	expression (497)	locus (606)	population (514)
mechanism (577)	evolution (304)	function (645)	response (411)	protein (585)	region (477)	dna (590)	locus (499)
rna (573)	region (295)	role (588)	bacteria (398)	breast (582)	chromosome (456)	variation (578)	association (468)
region (570)	announce ment (285)	method (546)	production (362)	response (568)	protein (456)	mutation (567)	relationship (442)

※ “gene”, “genome”, “study”, “NaN” 네 단어는 의미가 낮은 단어로 제외함

□ 연구과제와 논문의 차이점으로 인하여 직접적 비교에 한계가 있으나, 국내과제와 Pubmed 문헌(2017년 유전체 분야) 토픽 클러스터링 결과에 따르면 국내는 해외에 비해 우수 품종개발 또는 치료제 개발을 목적으로 하는 연구의 비중이 유전학 등의 기초연구에 비해 높은 편이었음

- 국내 유전체분야 연구과제의 큰 비중을 차지하는 분야는(연구비 기준) '유전자교정 활용 육종/치료기술', '유전정보 분석기술', '유전자변형 생물체 개발 및 위해성평가' 등이 있음
- 최근 CRISPR 등 유전자교정(편집) 기술이 크게 발전함에 따라 이를 이용한 우수 품종개발 또는 유전자치료제 개발 등이 활발하며, 유전체에서 유의미한 부분을 알아내기 위한 유전정보 분석기술의 연구 비중이 높은 것으로 파악됨

〈표 4-7〉 2017년 유전체분야 국내 과제 클러스터링 결과

번호	그룹(클러스터)	과제 수	정부연구비 (백만원)
1	유전자변형 생물체 개발 및 위해성평가	84(11.6%)	14,881(13.5%)
2	유전자교정 활용 육종/치료기술	133(18.4%)	21,738(19.6%)
3	유전정보 분석기술	82(11.3%)	16,772(15.2%)
4	생물자원 보존 및 다양성 연구	53(7.3%)	14,070(12.7%)
5	유전자치료제 기술	68(9.4%)	10,490(9.5%)
6	생물자원 확보/수집	140(19.4%)	12,933(11.7%)
7	유전학 및 기초기전연구	79(10.9%)	8,966(8.1%)
8	질병치료 응용기술(약리유전체학 등)	84(11.6%)	10,782(9.7%)
합 계		723	110,632

- Pubmed 문헌 분석 결과, 유전자-형질 연관성 연구(association study)의 비중이 특히 높았으며, 그 외 후생유전학(epigenetics), 계통유전체학(phylogenomics), 유전자교정, 진화유전체학, 기초기전 연구 등이 비슷한 비중을 차지하였음
- 문헌 비중이 높은 유전자교정 연구, 연관성 연구 및 후생유전학은 유전자 기반 치료법 개발 및 우수 품종 개발과 관련되므로, 국내와 세계에서 공통적으로 이에 대한 연구 비중이 높다고 볼 수 있음
- Pubmed 문헌 중에는 기초기전, 유전학, 계통유전체학 등 순수 기초연구에 해당하는 문헌의 비중이 약 38.7%이었으나, 국내의 경우 기초분야 연구비 비중은 20.8% 수준(생물자원 보존 및 다양성 연구, 유전학 및 기초기전연구)으로 다소 낮게 나타남

〈표 4-8〉 2017년 유전체분야 Pubmed 문헌 클러스터링 결과

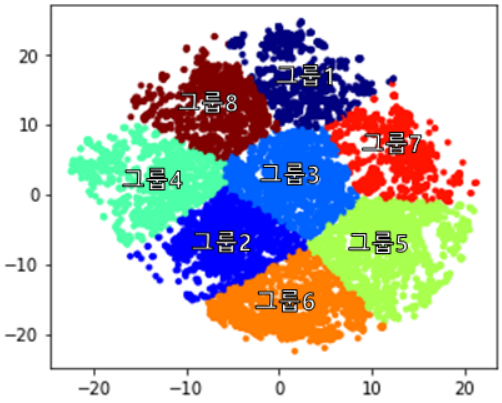
번호	그룹(클러스터)	문헌 수
1	유전자교정(CRISPR) 연구	1,632(12.9%)
2	유전체 해독(시퀀싱)	1,444(11.4%)
3	유전자-형질 연관성 연구	2,328(18.4%)
4	계통유전체학	1,698(13.4%)
5	암 관련 연관성 연구(후생유전학)	1,753(13.8%)
6	유전 및 진화 관련 연구	1,601(12.6%)
7	정밀의료 기반연구	598(4.7%)
8	기초기전 및 유전자다양성 연구	1,606(12.7%)
합 계		12,660

□ 2012년과 2017년 Pubmed 문헌 데이터(유전체 분야)를 각각 토픽 클러스터링하여 비교한 결과, 최근 학문·기술 발전에 따른 트렌드 변화가 확연히 드러남

○ 시간 변화에 따른 토픽 변화 분석 가능성을 시험하기 위하여, 2012년 Pubmed 문헌(유전체 분야)을 동일한 조건(8개 그룹)으로 클러스터링하였음

○ 2012년은 CRISPR와 같은 최신 유전자편집 기술이 등장하기 전으로 해당 분야의 비중이 적게 나타났으며, 유전체 해독(시퀀싱) 그룹이 2개(1, 8)로 게놈 해독의 수요가 높았던 시기를 반영함

- 그룹1과 그룹8은 서로 인접해있어 합칠 수 있음



그룹	주제
그룹1	유전체 해독(시퀀싱)
그룹2	유전체 분석기술 연구
그룹3	유전자 해석 기초연구(연관성 연구 등)
그룹4	진화유전체학
그룹5	기초기전 연구
그룹6	약리/영양유전체학 등
그룹7	연관성 연구 및 유전체 편집 연구
그룹8	유전체 해독(시퀀싱)

[그림 4-43] 2012년 유전체분야 Pubmed 문헌 클러스터링 결과(그룹 수: 8개)

〈표 4-9〉 2012년 유전체분야 Pubmed 문헌 클러스터링 결과

번호	그룹(클러스터)	문헌 수
1	유전체 해독(시퀀싱)	1,131(10.5%)
2	유전체 분석기술 연구	1,230(11.4%)
3	유전자 해석 기초연구(연관성 연구 등)	1,572(14.6%)
4	진화유전체학	1,178(10.9%)
5	기초기전 연구	1,772(16.5%)
6	약리/영양유전체학 등	1,537(14.3%)
7	연관성 연구 및 유전체 편집 연구	1,067(9.9%)
8	유전체 해독(시퀀싱)	1,277(11.9%)
합 계		10,764

ㅇ 2012년과 2017년의 Pubmed 문헌 클러스터링 결과를 비교한 결과, 기술·학문 발전에 따른 상당한 트렌드 차이가 발견됨

- 단순 유전체 해독 결과를 보고하는 논문의 비중은 절반 수준(22.4% → 11.4%)으로 감소함

- 반면에 CRISPR와 같은 유전체편집 기술의 등장에 따라 유전자와 형질의 연관성에 대한 연구 및 유전자교정 연구의 비중이 크게 증가하였음(9.9% → 31.3%, 붉은색 표시)

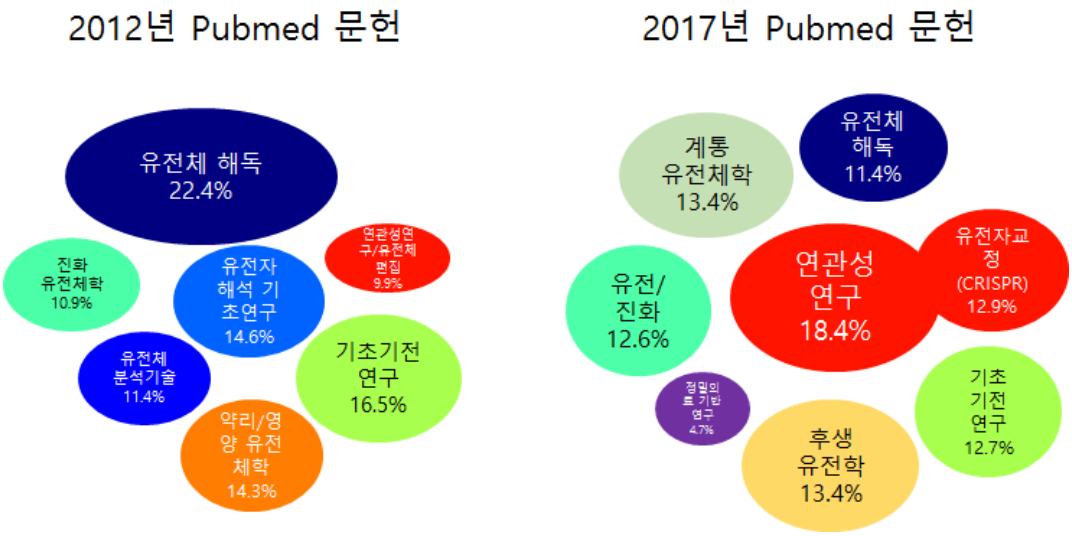
- 암 등 질병에 대한 기초기전연구의 비중은 다소 감소(16.5% → 12.7%)했으나, 진화 및 유전에 관한 기초연구 비중은 증가하였고(10.9% → 12.6%), 계통 유전체학 분야가 새롭게 등장하였음

※ 2017년에 유전체 분야 문헌 전체 수가 2012년에 비하여 17.6%(10,764건 → 12,660건) 증가했음을 감안하면, 2012년 대비 비중이 줄어든 일부 분야의 경우 문헌의 절대 수는 크게 감소하지 않음

- 약리유전체학(pharmacogenomics), 영양유전체학(nutrigenomics)으로 대변되던 그룹이 없어진 반면*, 후생유전학(epigenetics)을 중심으로 하는 그룹이 등장하였음

* 비중이 낮아져 다른 그룹에 섞여 들어갔다고 볼 수 있음

- 유전체 분석 기법에 대한 그룹은 없어지고 정밀의료 기반연구에 해당하는 그룹이 생성된 것은 이제 단순 분석법은 충분히 이용 가능한 수준이고 이를 정밀의료에 응용하고자 하는 연구로 전환되었음을 시사함



[그림 4-44] 2012년, 2017년 유전체분야 Pubmed 문헌 클러스터링 결과 비교

□ 따라서, 본 연구에서 제안하는 딥러닝(doc2vec) 기반의 토픽 클러스터링 기법은 유전체 분야 국내 연구과제 및 Pubmed 문헌정보에 적용한 결과 유의미한 분석결과가 도출되었다고 판단됨

□ 추가적으로, 위에 2017년도 Pub med 데이터에 같은 조건으로 일반적인 토픽클러스터링(LDA)을 적용한 결과는 〈표 4-10〉, 〈표 4-11〉과 같음

○ 딥러닝(doc2vec) 기반 토픽 클러스터링 결과와 〈표 4-10〉, 〈표 4-11〉를 비교 시 '암 관련 연관성 연구' 등 일부 유사 그룹이 생성되는 것을 확인할 수 있었음

- 딥러닝 기반 토픽클러스터링 시 암 관련 연관성 연구 그룹에서 cancer 키워드 순위가 상위 포진(1위)하였고, 관련된 'mutation', 'carcinoma'등의 키워드도 도출되거나 순위가 상위권이었던 점을 감안하면 보다 특징적인 클러스터링이 이루어진 것으로 사료됨

※ 〈표 4-10〉, 〈표 4-11〉 에서는 cancer 연구를 특정 지을 수 있는 'carcinoma' 키워드는 도출되지 않았음

○ 클러스터링의 결과는 방법론 및 데이터의 특성에 다르게 나타나기 때문에 사용자 스스로 사용하는 데이터를 바탕으로 활용 목적 가장 부합하는 방법론을 선택하는 것이 바람직하다고 판단됨

〈표 4-10〉 2017년 유전체분야 Pubmed 문헌 LDA 토픽클러스터링결과(가중치 기준)

그룹1	그룹2	그룹3	그룹4	그룹5	그룹6	그룹7	그룹8
protein (0.007866)	dna (0.008385)	analysis (0.010549)	sequence (0.00814)	dna (0.009624)	dna (0.008675)	cell (0.006744)	sequence (0.007557)
dna (0.00711)	genetic (0.008031)	genetic (0.008634)	genetic (0.007379)	cell (0.008778)	human (0.006589)	dna (0.006422)	analysis (0.006976)
analysis (0.007026)	genomic (0.005987)	dna (0.0071)	cell (0.007284)	genetic (0.007154)	analysis (0.006506)	disease (0.006422)	dna (0.006578)
genetic (0.006271)	sequence (0.005915)	human (0.006304)	analysis (0.0067)	protein (0.006078)	genomic (0.006002)	protein (0.006352)	genomic (0.006116)
sequence (0.006178)	analysis (0.005764)	association (0.006041)	human (0.00568)	analysis (0.005769)	cell (0.005901)	rna (0.006153)	protein (0.004905)
genomic (0.006091)	model (0.005127)	genomic (0.005932)	expression (0.005439)	sequence (0.004835)	cancer (0.005199)	sequencing (0.006016)	expression (0.004737)
cell (0.006022)	data (0.004879)	sequence (0.005741)	sequencing (0.005371)	genomic (0.004734)	genetic (0.005134)	analysis (0.005951)	cell (0.004482)
human (0.005584)	cell (0.004456)	disease (0.00505)	protein (0.005006)	disease (0.004409)	sequence (0.004468)	expression (0.005696)	genetic (0.004428)
sequencing (0.004757)	using (0.004348)	cell (0.004914)	data (0.00477)	rna (0.004228)	expression (0.004242)	genetic (0.005524)	population (0.004333)
wide (0.004146)	human (0.004321)	protein (0.004747)	molecular (0.004348)	human (0.004035)	genomics (0.003959)	genomic (0.005268)	data (0.00428)
data (0.004121)	wide (0.004182)	sequencing (0.004118)	disease (0.004304)	expression (0.003941)	protein (0.003832)	human (0.004689)	sequencing (0.004203)
species (0.003611)	association (0.004169)	wide (0.004043)	genomics (0.004151)	cancer (0.003614)	molecular (0.003613)	sequence (0.003986)	human (0.003928)
expression (0.003375)	sequencing (0.004149)	data (0.003895)	genomic (0.00404)	mutation (0.003525)	wide (0.003475)	molecular (0.003967)	species (0.003591)
high (0.003309)	protein (0.004126)	population (0.003799)	wide (0.003921)	model (0.003269)	sequencing (0.003469)	species (0.003866)	cancer (0.003566)
region (0.003213)	rna (0.004108)	variant (0.003437)	dna (0.003772)	associated (0.003256)	disease (0.003421)	plant (0.003689)	genomics (0.00339)
association (0.003206)	cancer (0.004021)	using (0.003235)	cancer (0.003485)	wide (0.003237)	data (0.003349)	mutation (0.003291)	plant (0.003139)
molecular (0.003105)	single (0.003739)	rna (0.003211)	using (0.003301)	association (0.003214)	associated (0.003193)	data (0.003075)	based (0.003047)
associated (0.003023)	locus (0.003663)	genomics (0.003195)	species (0.003169)	molecular (0.003156)	strain (0.003057)	wide (0.003036)	molecular (0.002994)
model (0.002847)	trait (0.003492)	expression (0.003091)	based (0.00287)	using (0.003128)	rna (0.002904)	genomics (0.002972)	disease (0.002953)
genomics (0.002681)	species (0.003462)	cancer (0.003021)	associated (0.002835)	patient (0.003002)	high (0.002821)	patient (0.002708)	chromosome (0.002884)

※ "gene", "genome", "study", "NaN" 네 단어는 의미가 낮은 단어로 제외함. 괄호 안은 가중치(weight) 값 의미

〈표 4-11〉 2017년 유전체분야 Pubmed 문헌 LDA 토픽클러스터링결과(빈도 기준)

그룹1	그룹2	그룹3	그룹4	그룹5	그룹6	그룹7	그룹8
protein	dna	dna	analysis	sequence	cell	dna	sequence
dna	genetic	cell	genetic	genetic	dna	human	analysis
analysis	genomic	genetic	dna	cell	disease	analysis	dna
genetic	sequence	protein	human	analysis	protein	genomic	genomic
sequence	analysis	analysis	association	human	rna	cell	protein
genomic	model	sequence	genomic	expression	sequencing	cancer	expression
cell	data	genomic	sequence	sequencing	analysis	genetic	cell
human	cell	disease	disease	protein	expression	sequence	genetic
sequencing	using	rna	cell	data	genetic	expression	population
wide	human	human	protein	molecular	genomic	genomics	data
data	wide	expression	sequencing	disease	human	protein	sequencing
species	association	cancer	wide	genomics	sequence	molecular	human
expression	sequencing	mutation	data	genomic	molecular	wide	species
high	protein	model	population	wide	species	sequencing	cancer
region	rna	associated	variant	dna	plant	disease	genomics
association	cancer	wide	using	cancer	mutation	data	plant
molecular	single	association	rna	using	data	associated	based
associated	locus	molecular	genomics	species	wide	strain	molecular
model	trait	using	expression	based	genomics	rna	disease
genomics	species	patient	cancer	associated	patient	high	chromosome

□ 향후 본 연구결과를 보완하고 발전시키기 위해 추가 작업으로 다음을 제안함

- 2차원으로 차원을 축소하는 과정에서 정보의 손실이 크게 일어나므로, 3차원으로 입체 공간상에 축소하여 클러스터링 품질 향상을 도모
- 본 연구에서 사용한 t-SNE 알고리즘은 벡터들을 균일하게 배치하는 경향이 있으므로, 원래 데이터의 특성을 보다 그대로 반영 가능한 대체 알고리즘 탐색

다. 17년도 정부 신약개발 R&D과제 DB구축

1) 범용화 된 분류모형 작동 방식

□ 의약과제 분류모형은 범용화 모델로 재구성되며 새로운 6개 기계학습 방법론* 적용이 가능하도록 개선되었음

* Deep Learning, Linear Regression, Random Forest, Naive Bayes, GBM, SVM

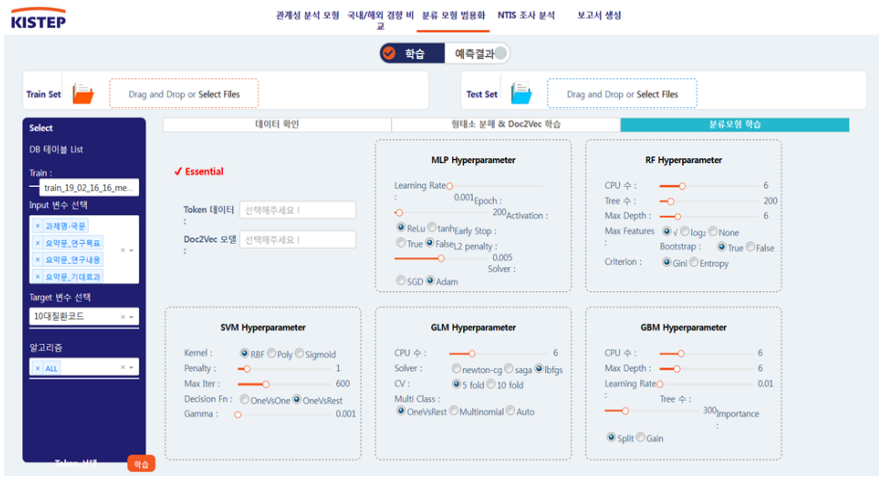
□ 분류 모형의 학습 컬럼을 선택하여 학습할 과제정보를 입력(Train: 항목)하고 과제명, 요약문 등의 과제 구분을 위한 특성을 선택하고(Input변수), 학습대상항목(Target 변수)을 지정

[그림 4-45] 분류 모형 학습과정(데이터 입력)

□ 다음으로는 형태소 분석 및 doc2vec 학습을 진행하게 되며, 이때 윈도우 및 벡터 사이즈, 학습 횟수(Epoch) 등을 설정할 수 있음

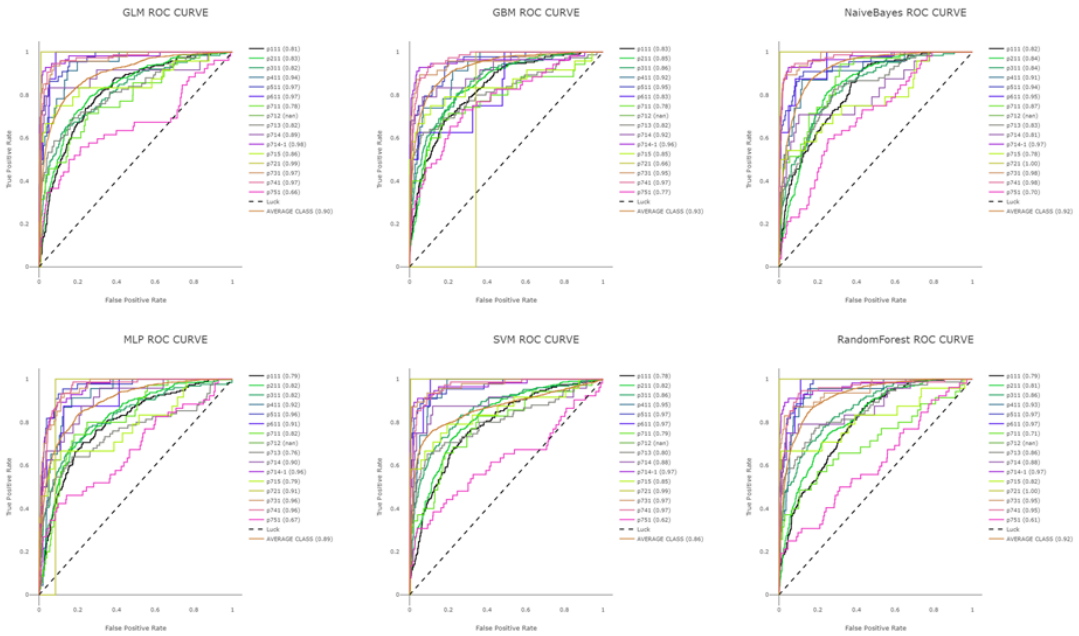
[그림 4-46] 분류 모형 학습과정(형태소 분석 및 doc2vec 학습)

□ 이 후 주어진 기계학습 방법론별로 학습파라미터를 사용자가 직접조정하고 분류모형 학습 후 분류를 수행



[그림 4-47] 분류 모형 학습과정(분류모형 학습)

□ 사용자는 분류 모형 학습 후 어떠한 방법론이 분류정보에 적합한지를 ROC 커브 분석을 통해 사전에 파악할 수 있음

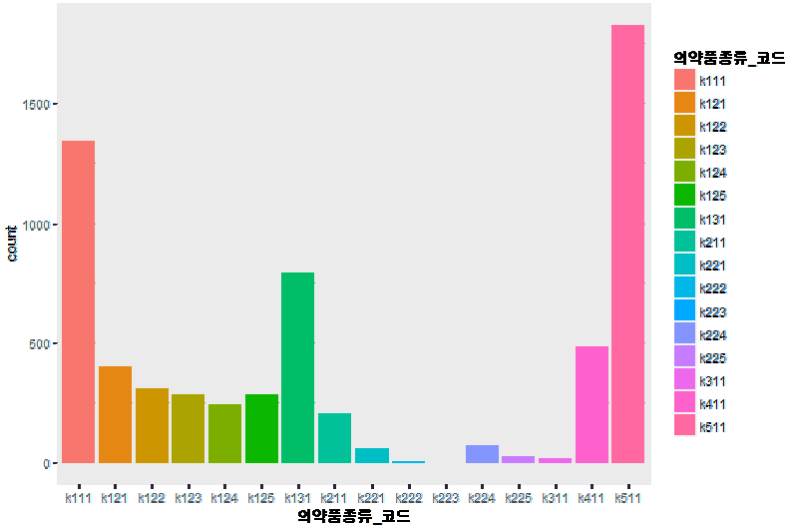


[그림 4-48] 제공되는 방법론별 ROC 커브 분석을 통한 예측성능 평가 결과 예시

2) 17년도 신약개발연구과제DB 분류 결과

□ 기본적인 연구방법은 전년과 동일하게 진행하였으며 16년도 신약개발연구과제DB 967건을 포함하여 총 7,318 건의 과제정보를 분류모형에 훈련정보로 사용하였음

- o 그럼에도 훈련정보의 비대칭성(코드별 연구과제수의 차이)로 인해 의약품종류/신약개발단계/대상질환별 분류 시 분류 퍼포먼스에서 차이가 발생함
  - 가령, 의약품 종류 코드의 경우 k511(공통기반기술 및 기타, 공통기반기술), k111(신약, 합성신약) 코드에는 많은 양의 연구과제가 존재하나 K222(개량신약, 유전자치료제), K223(개량신약, 세포치료제) 기존의 훈련정보가 매우 부족하거나 부재하였음



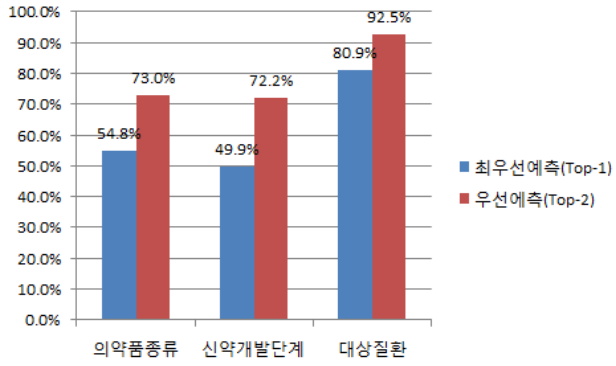
[그림 4-49] 의약품종류코드별 연구과제수

※ 출처: 김한해 외 (2017), 「기계학습 기반 바이오의료분야 과학기술정보데이터 분석·활용 모형 개발」, 한국과학기술기획평가원.

□ 17년도 신약개발연구과제DB 분류결과[20]와 16년도 분류결과를 비교한 결과 의약과제분류모형의 예측성능이 상당히 향상된 것을 확인할 수 있었음

○ 16년 DB 분류결과 예측 시 16년도 과제는 제외하고 훈련시켰으며, 16년 계속과제 역시 분류대상에서 제외*하여 총 467건의 신규과제에 대해서만 모형에 의한 분류를 수행하였음

* 16년도 계속과제의 경우 기 훈련정보에 포함되어 제외



[그림 4-50] 16년 신약개발연구과제 DB 분류모형 예측결과

※ 출처: 김한해 외 (2017), 「기계학습 기반 바이오의료분야 과학기술정보데이터 분석·활용 모형 개발」, 한국과학기술기획평가원.

20) 전문가에 의한 분류결과는 '제2장 신약개발 정부 R&D 투자포트폴리오 분석'에 제시

ㅇ 17년 DB의 경우 총 1169개의 대상 과제 중 649개의 계속과제를 제외*한 520개의 신규과제에 대해서만 분류 예측을 수행하였음

* 17년 계속과제 역시 7,318건의 훈련정보에 포함되어 제외

ㅇ 의약품종류/신약개발단계/대상 질환별 분류 모두 16년 대비 17년 신규과제에 대한 모형의 예측률이 모두 상당히 상승하여 동 분류모형을 구축 시 수립한 가설에 부합*하는 결과가 도출됨

* 학습정보가 축적될수록 모형의 분류성능이 향상됨

- 16년 신규의약과제 분류 시 Top1(최우선) 예측 기준 의약품종류 54.8%, 신약개발단계 49.9%, 대상질환 80.9%의 성능을 보여준 반면, 17년도의 경우 Top1 예측 기준 의약품종류 60.57%, 신약개발단계 52.89%, 대상질환 81.55%의 향상된 예측 성능을 보여줌

※ 신약개발단계의 경우 Top2(우선) 예측은 전년도 대비 소폭 낮게 예측됨

- 이에 따라, 전년도와 마찬가지로 딥러닝 기반의 분류가 타 방법론 기반 분류 대비 월등히 우수한 성능을 보임을 확인할 수 있었음

의약품종류코드		
모형	Top1	Top2
RandomForest	32.12%	45.00%
SVM	50.38%	67.12%
GLM	50.58%	71.92%
DeepLearning	60.57%	77.12%
NaiveBayes	49.62%	66.92%
GBM	48.27%	71.92%

신약개발단계코드		
모형	Top1	Top2
RandomForest	33.65%	55.96%
SVM	42.65%	65.00%
GLM	39.42%	59.62%
DeepLearning	52.89%	69.42%
NaiveBayes	34.23%	54.42%
GBM	41.15%	62.88%

10대질환코드		
모형	Top1	Top2
RandomForest	49.61%	69.81%
SVM	75.19%	88.27%
GLM	72.88%	92.50%
DeepLearning	81.55%	92.50%
NaiveBayes	72.31%	87.69%
GBM	63.65%	82.50%

[그림 4-51] 17년 신약개발연구과제 DB 분류모형 예측결과

□ 신약개발단계 분류모형의 Top1(최우선) 예측성능이 가장 낮은 것은 연구과제 상 복수개발단계를 수행 경우가 많은 점에 기인하는 것으로 판단됨

ㅇ 코드 분류를 살펴보면 신약개발연구과제DB의 '소분류' 기준으로 신약개발단계는 15개 코드인 반면 의약품종류와 질환은 각각 11개 코드로 구성됨

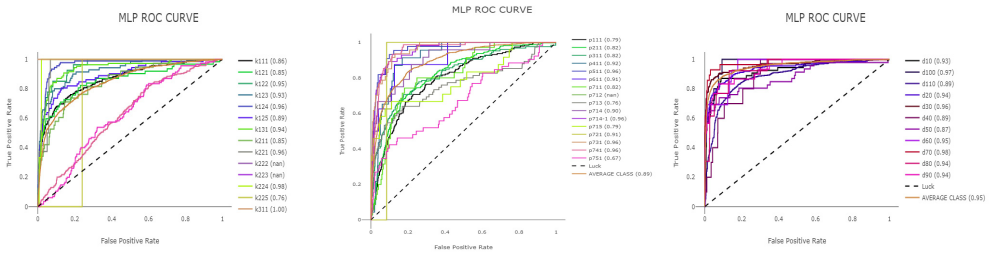
- 분류 코드가 많을 경우 훈련데이터(연구과제수)가 상대적으로 적게 배분되거나 고르지 못한 경우가 많아 훈련 시 예측성능에 영향을 줄 수 있음

- o 따라서, 한 개의 과제 내 에서 복수의 신약개발단계를 수행하는 과제들의 분류가 적절히 되기 어려워 이러한 결과가 야기되는 것으로 사료됨
- o 가령, 기 추진된 과제('16년 신약개발연구과제DB 구축)에서 500개 계속과제의 대 전문가 재검토결과 일부 계속과제들은 분류가 용이하지 않다*는 의견[21]을 받음

  * 의약품종류 19.2%, 신약개발단계 23.4%, 대상 질환 6.0%

□ 분류모형의 예측성능을 테스트하기 위한 ROC 분석에서 의약품종류/신약개발단계/대상 질환 모두 준수한 AUC 수치를 보여주었음

- 딥러닝 기반 분류모형의 경우 의약품종류/신약개발단계/대상 질환의 AUC 값은 0.85~0.95 수준으로 나타남



[그림 4-52] 딥러닝 기반 분류모형의 ROC 분석 결과 (의약품종류/신약개발단계/대상질환 순)

21) 분류기준별 전문가 개인의 시각차 존재하고 연구과제가 복수의 내용으로 구성될 수 있음

□ 모형의 분류결과의 타당성 검토를 위해 2017년도 기준 분류기준별 모형에 의한 예측결과와 전문가 분류에 따른 신약개발과제DB 분석결과(전문가분류)[22]를 비교함

- ㅇ 신약개발단계별분류는 데이터의 특성 상 예측률이 가장 낮았음에서도 전문가 분류결과와 유사한 투자 포트폴리오 분석 결과를 보여주었다고 사료됨
- ㅇ 비중을 기준으로 타겟 발굴 및 검증 단계에서 약 4%의 차이를 보였으나 이외의 단계에서는 전문가 분류결과 대비 약 2% 이내 수준의 차이를 보였음

〈표 4-12〉 신약개발분야 정부 R&D 신약개발단계별 투자 현황

구분			2017년 (모형예측)		2017년 (전문가분류)	
			연구비 (백만원)	비중 (%)	연구비 (백만원)	비중 (%)
타켓발굴 및 검증	타겟발굴 및 검증		25,577	7.4	38,419	11.1
후보물질 도출 및 최적화	후보물질도출 및 최적화		118,988	34.3	113,837	32.8
비임상	비임상		42,603	12.3	34,708	10.0
임상	임상1상		5,395	1.6	14,820	4.3
	임상2상		24,798	7.1	28,423	8.2
	임상3상		18,596	5.4	5,535	1.6
인프라	신약 플랫폼 기술	타겟발굴 플랫폼	7,682	2.2	8,349	2.4
		후보물질 발굴 플랫폼	38,312	11.0	34,199	9.9
		비임상 플랫폼	9,161	2.6	8,221	2.4
		질환동물 플랫폼	17,001	4.9	16,718	4.8
		임상 플랫폼	6,333	1.8	6,895	2.0
	인력양성		80	0.0	60	0.0
	제도·정책		5,814	1.7	7,557	2.2
	인·허가		13,529	3.9	13,904	4.0
기타	기타		13,223	3.8	15,447	4.5
합계			347,092	100.0	347,092	100.0

22) 전문가 분류에 의한 상세 포트폴리오

○ 의약품종류별 모형예측과 전문가분류에 의한 투자 포트폴리오 분석 결과는 항목별 투자비중 차가 대부분 2% 이내로 나타나며 유사한 경향을 보임을 확인할 수 있었음

〈표 4-13〉 신약개발분야 정부 R&D 의약품 종류별 투자 현황

구분		2017년 (모형예측)		2017년 (전문가분류)	
		연구비 (백만원)	비중 (%)	연구비 (백만원)	비중 (%)
합성신약		91,619	26.4	99,204	28.6
바이오 신약	단백질 치료제	22,432	6.5	18,888	5.4
	유전자 치료제	12,448	3.6	10,451	3.0
	세포 치료제	20,577	5.9	19,714	5.7
	백신	28,612	8.2	24,171	7.0
	항체	23,004	6.6	22,930	6.6
천연물신약		26,818	7.7	24,942	7.2
개량신약(합성)		11,880	3.4	10,756	3.1
바이오 베터	단백질 치료제	2,598	0.7	3,008	0.9
	유전자 치료제	-	-	-	-
	세포 치료제	-	-	-	-
	백신	200	0.1	575	0.2
	항체	-	-	890	0.3
바이오시밀러		1,200	0.3	1,200	0.3
공통기반기술		87,660	25.3	86,924	25.0
기타		18,044	5.2	23,441	6.8
총 합계		347,092	100.0%	347,092	100.0

o 가장 분류 성능이 좋게 나타난 대상 질환별 투자현황의 경우 비중을 기준으로 대부분 1%이내의 차이를 보이며 전문가 분류에 의한 투자 포트폴리오 결과와 거의 일치하는 경향을 나타냄

〈표 4-14〉 신약개발분야 정부 R&D 대상 질환별 투자 현황

구분	2017년 (모형예측)		2017년 (전문가분류)	
	연구비 (백만원)	비중 (%)	연구비 (백만원)	비중 (%)
감염증	43,158	12.4	45,204	13.0
골다공증	1,928	0.6	1,988	0.6
관절염	11,938	3.4	10,298	3.0
당뇨	8,630	2.5	7,905	2.3
비만	2,011	0.6	2,591	0.7
정신질환	1,172	0.3	1,772	0.5
종양	89,190	25.7	86,842	25.0
천식	3,790	1.1	4,410	1.3
퇴행성 뇌질환	12,517	3.6	13,356	3.8
혈관질환	14,352	4.1	16,875	4.9
기타	158,405	45.6	155,851	44.9
합계	347,092	100.0%	347,092	100.0

3) 분류 결과 문서정보화

□ 분류 결과는 모형을 통해 보고서 형식의 리포트 출력(PDF파일)이 가능함

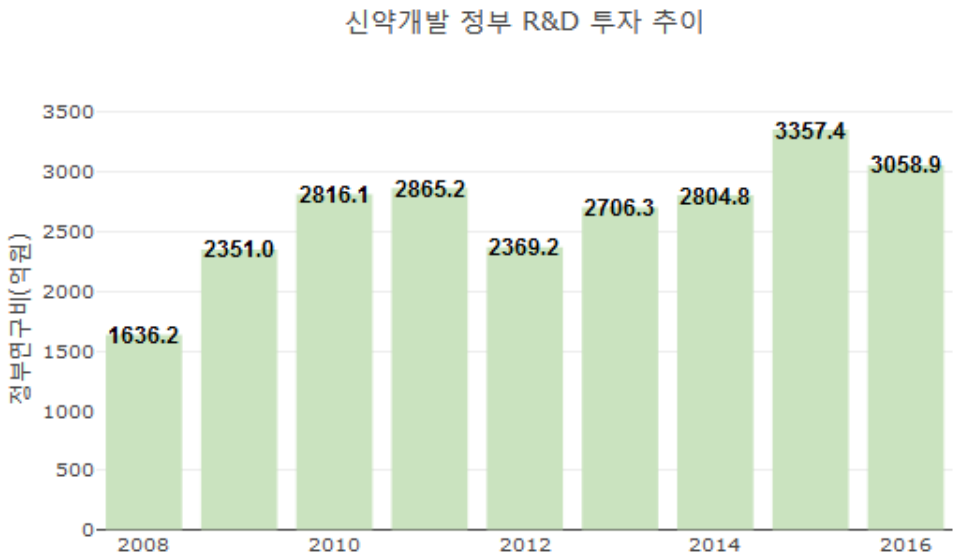
○ 그래프 표현과 함께 기본적인 분석결과 내용이 자동 작성됨

신약개발 정부 R&D 총 투자

◎ 최근 9년(2008년-2016년) 동안 총 '23,965'억원 투자(연간 평균 '2,663')

정부 R&D 총 투자 금액의 연평균 증가율은 '9.6%'

신약개발 정부 R&D 총 투자



[그림 4-53] 분석 결과 보고서화 예시(9년간 총 투자현황)

'◎' 2016년 3,059억원 중 인프라 단계에 가장 많은 1,126 억원(36.8%)이 투자되었고, 후보물질도출 및 최적화 임상 순

- 후보물질도출 및 최적화 674억원(22.0%), 임상 428억원(14.0%), 타겟발굴 및 검증 373억원(12.2%) 순으로 투자

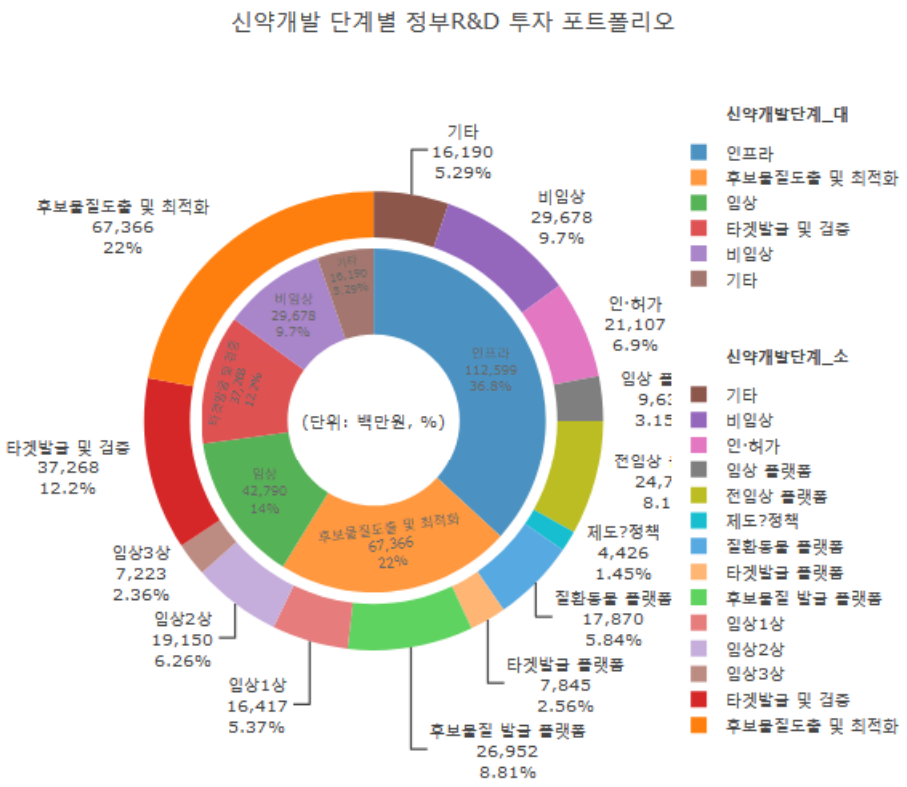
- 인프라 단계는 후보물질 발굴 플랫폼 270억원(23.9%)으로 가장 많이 투자되었고,

전임상 플랫폼 248억원(22.0%), 인·허가 211억원(18.7%) 순으로 투자

- 임상 단계는 임상2상 192억원(44.8%)으로 가장 많이 투자되었고,

임상1상 164억원(38.4%), 임상3상 72억원(16.9%) 순으로 투자

신약개발 단계별 포트폴리오



[그림 4-54] 분석 결과 보고서화 예시(신약개발단계별 투자현황)

제5장

신약개발 정부 R&D 투자포트폴리오 분석

제5장 신약개발 정부 R&D 투자포트폴리오 분석[23)]

제1절 신약개발 투자포트폴리오 분류기준

□ 생명보건의료분야 예산심의대상 정부 연구 개발사업 중 신약개발분야에 해당하는 사업('17년 기준 30개)을 중심으로, 국가과학기술정보서비스(NTIS)에서 제공하는 국가연구 개발사업 조사 분석데이터('17)의 과제정보를 활용

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

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

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

23) 신약분야 전문가를 통해 구축한 17년도 정부 신약개발 R&D과제 DB 결과 분석

〈표 5-1〉 신약개발분야 정부 R&D 투자포트폴리오 분류기준

구분	대분류	중분류	소분류
신약개발 단계	타겟 발굴 및 검증	타겟 발굴 및 검증	타겟 발굴 및 검증
	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
			비임상 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
			유전자 치료제
			세포 치료제
			백신
			항체
		천연물신약	천연물신약
	개량신약	개량신약(합성)	개량신약
		바이오 베터	단백질 치료제
			유전자 치료제
			세포 치료제
			백신
			항체
	복제약	바이오 시밀러	바이오 시밀러
	공통기반기술 및 기타	공통기반기술	공통기반기술
		기타	기타
질환	혈관질환, 천식, 종양, 감염증, 정신질환, 퇴행성뇌질환, 골다공증, 당뇨, 비만, 관절염, 기타		

제2절 2017년도 신약개발 R&D 투자포트폴리오 분석

가. 신약개발 분야 정부 R&D 투자 현황

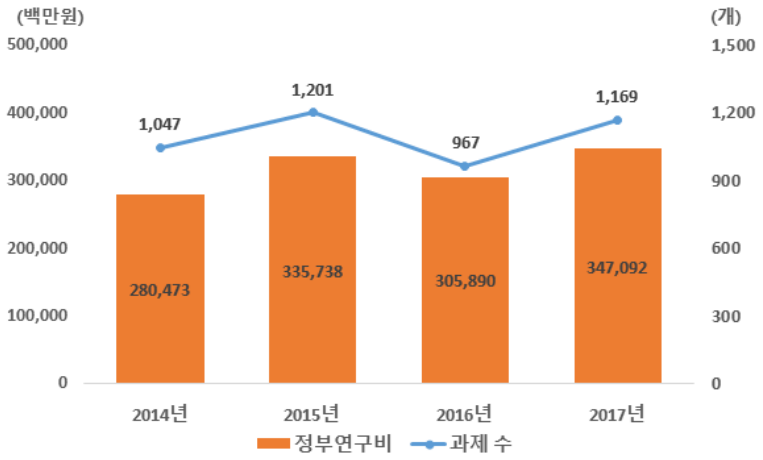
□ 신약개발분야 정부 R&D 투자 현황

○ 2017년 신약개발분야 정부 R&D 투자 규모는 3,471억 원으로, 연구비 기준 4년간 ('14~'17) 연평균 약 7.4% 증가

- 신약개발과제 수는 '14년 1,047건에서 '17년 1,169건으로 연평균 약 3.7% 증가

〈표 5-2〉 신약개발 분야 정부 R&D 투자 규모

구분	2014년	2015년	2016년	2017년	연평균 증가율
과제 수(개)	1,047	1,201	967	1,169	3.7%
정부연구비 (백만원)	280,473	335,738	305,890	347,092	7.4%



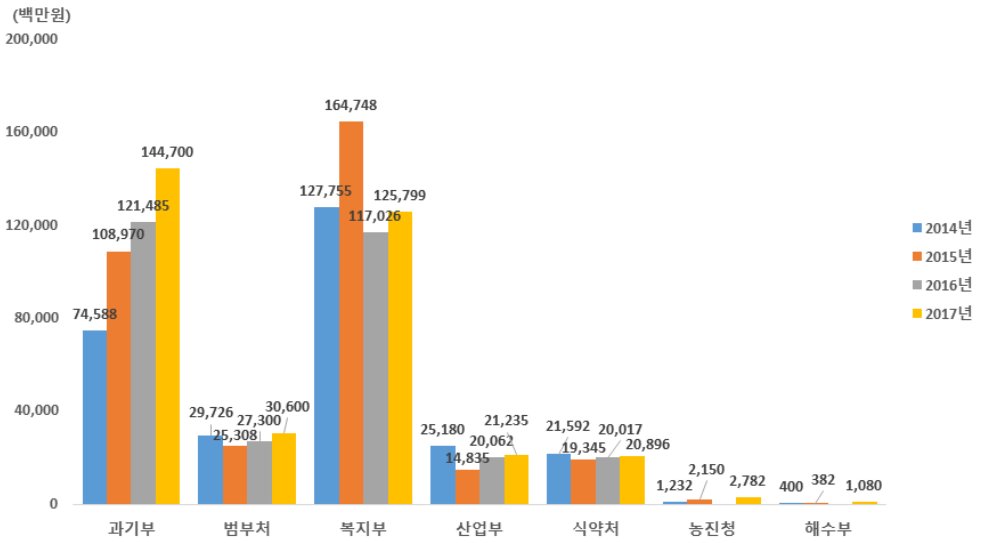
[그림 5-1] 신약개발분야 정부 R&D 투자 현황

□ 부처별 투자 현황

○ 2017년 신약개발 분야에 정부연구비 총액 3,471억 원 중 과기정통부의 투자가 1,447억 원으로 가장 큰 비중(41.7%)을 차지(연평균 24.7% 증가)

○ 다음으로 복지부(1258억 원, 36.2%), 범부처(306억 원, 8.8%) 순으로 신약개발과제 지원

- 3개 부처가 전체 예산의 86.7%를 차지



[그림 5-2] 신약개발분야 정부 R&D 부처별 투자 현황

〈표 5-3〉 신약개발분야 정부 R&D 부처별 투자 현황

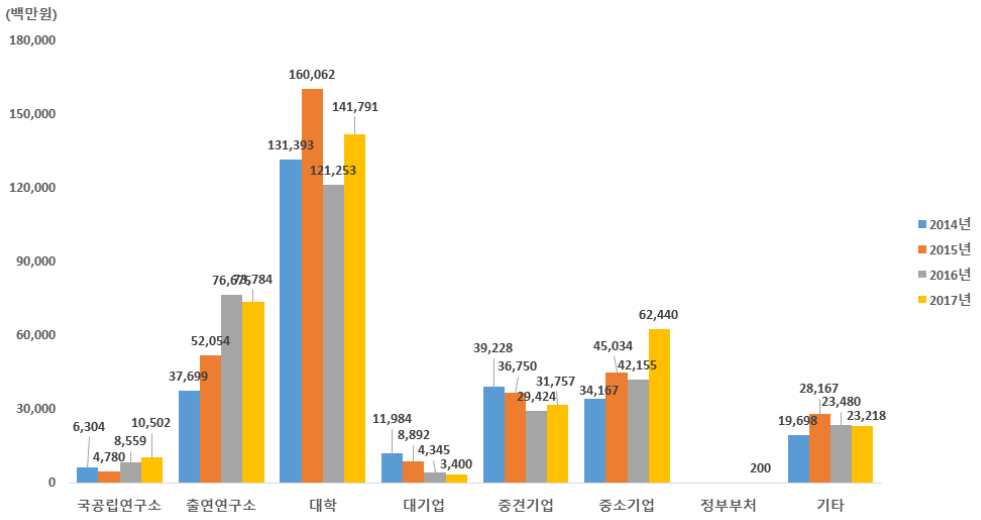
구분	2014년		2015년		2016년		2017년		연평균 증가율 (%)
	연구비 (백만원)	비중 (%)	연구비 (백만원)	비중 (%)	연구비 (백만원)	비중 (%)	연구비 (백만원)	비중 (%)	
과학기술 정보통신부	74,588	26.6	108,970	32.5	121,485	39.7	144,700	41.7	24.7
범부처 사업	29,726	10.6	25,308	7.5	27,300	8.9	30,600	8.8	1.0
보건복지부	127,755	45.5	164,748	49.1	117,026	38.3	125,799	36.2	-0.5
산업통상 자원부	25,180	9.0	14,835	4.4	20,062	6.6	21,235	6.1	-5.5
식품의약품 안전처	21,592	7.7	19,345	5.8	20,017	6.5	20,896	6.0	-1.1
농촌진흥청	1,232	0.4	2,150	0.6	-	-	2,782	0.8	31.2
해양수산부	400	0.1	382	0.1	-	-	1,080	0.3	39.2
합계	280,473	100.0	335,738	100.0	305,891	100.0	347,092	100.0	7.4

□ 연구수행주체별 투자 현황

o '17년 기준 대학에서 1,418억 원(40.9%) 규모의 가장 높은 투자 비중을 보임

- 다음으로 출연연(738억 원, 21.3%), 중소기업(624억 원, 18.0%), 중견기업(318억 원, 9.1%) 순으로 투자

o 대기업 및 중견기업의 신약개발분야 정부 R&D 투자는 축소된 반면(연평균 -34.3%, -6.8%), 출연연은(연구비 기준) '14년 377억 원에서 '17년 738억 원으로(연평균 약 25.1%)로 가장 많이 증가



[그림 5-3] 신약개발분야 정부 R&D 연구수행주체별 투자 현황

〈표 5-4〉 신약개발분야 정부 R&D 연구수행주체별 투자 현황

구분	2014년 ·		2015년 ·		2016년 ·		2017년		연평균 증가율 (%)
	연구비 (백만원)	비중 (%)	연구비 (백만원)	비중 (%)	연구비 (백만원)	비중 (%)	연구비 (백만원)	비중 (%)	
국공립연구소	6,304	2.2	4,780	1.4	8,559	2.8	10,502	3.0	18.5
출연연구소	37,699	13.4	52,054	15.5	76,675	25.1	73,784	21.3	25.1
대학	131,393	46.8	160,062	47.7	121,253	39.6	141,791	40.9	2.6
대기업	11,984	4.3	8,892	2.6	4,345	1.4	3,400	1.0	-34.3
중견기업	39,228	14.0	36,750	10.9	29,424	9.6	31,757	9.1	-6.8
중소기업	34,167	12.2	45,034	13.4	42,155	13.8	62,440	18.0	22.3
정부부처	-	-	-	-	-	-	200	0.1	-
기타	19,698	7.0	28,167	8.4	23,480	7.7	23,218	6.7	5.6
합계	280,473	100.0	335,738	100.0	305,891	100.0	347,092	100.0	7.4

□ 주요 대상사업별 투자 현황

o 신약개발분야를 지원하는 주요 사업으로는 과기정통부의 바이오·의료기술개발이 857억 원(24.7%)을 지원하고 있었으며, 동 사업 내 신약개발분야 투자 비중은 33.5% 수준

o 다음으로 복지부의 첨단의료기술개발(379억 원, 10.9%), 범부처전주기신약개발(297억 원, 8.6%), 산업부의 바이오산업핵심기술개발(212억 원, 6.1%) 순으로 신약개발분야 주요 사업으로 나타남

〈표 5-5〉 신약개발분야 정부 R&D 주요 사업(2017)

사업명	총 사업비 (백만원) (A)	의약분야 투자액 (백만원)		비중(%) (B/A)
		(B)	비중(%)	
바이오.의료기술개발	255,643	85,651	24.7	33.5
첨단의료기술개발	73,751	37,888	10.9	51.4
범부처전주기신약개발	33,000	29,700	8.6	90.0
바이오산업핵심기술개발	66,443	21,235	6.1	32.0
질환극복기술개발	65,443	18,866	5.4	28.8
감염병위기대응기술개발	27,287	17,726	5.1	65.0
한국생명공학연구원연구운영비지원	86,495	17,137	4.9	19.8
의약품등안전관리	24,032	15,693	4.5	65.3
한국화학연구원연구운영비지원	69,867	12,052	3.5	17.3
연구중심병원육성	24,375	9,883	2.8	40.5
한국한의학연구원연구운영비지원	48,692	9,863	2.8	20.3
글로벌프론티어지원	87,000	7,772	2.2	8.9
임상연구인프라조성	47,327	7,487	2.2	15.8
첨단바이오의약품글로벌진출사업	11,250	7,485	2.2	66.5
암연구소및국가암관리사업본부연구운영비지원	54,265	7,447	2.1	13.7
국가항암신약개발사업	7,619	6,719	1.9	88.2
선도형특성화연구사업	10,500	5,783	1.7	55.1
감염병관리기술개발연구	22,661	4,753	1.4	21.0
첨단의료복합단지기반기술구축	7,304	4,137	1.2	56.6
안전성평가연구소연구운영비지원	26,988	4,059	1.2	15.0
한의약선도기술개발	16,306	3,679	1.1	22.6
안전성평가기술개발연구	12,525	3,153	0.9	25.2
차세대바이오그린21	53,406	2,782	0.8	5.2
안전기술선진화	3,100	2,050	0.6	66.1
해양수산생명공학기술개발	30,558	1,080	0.3	3.5
포스트게놈신산업육성을위한다부처유전체사업	48,797	900	0.3	1.8
국가보건의료연구인프라구축	11,959	720	0.2	6.0
뇌과학원천기술개발	41,750	680	0.2	1.6
양·한방융합기반기술개발	6,092	412	0.1	6.8
100세사회대응고령친화제품연구개발	3,353	300	0.1	8.9
합계	1,277,787	347,092	100.0	27.2

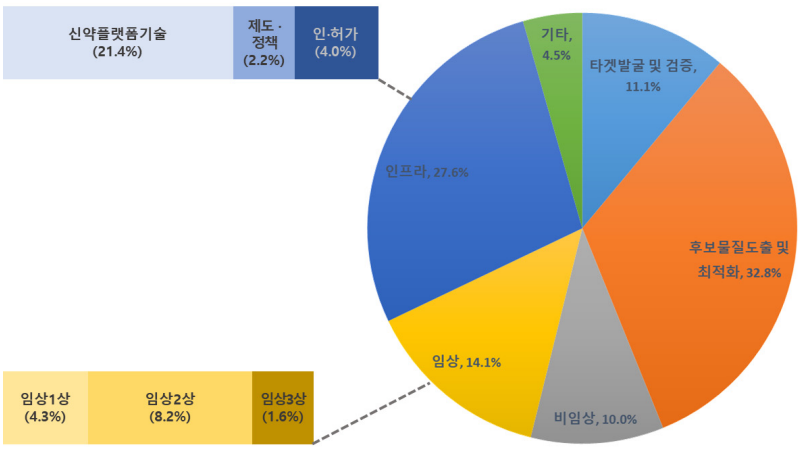
나. 신약개발단계별 정부 R&D 투자 현황

□ 신약개발단계별 정부 R&D 투자 현황

○ 중분류 기준 인프라 단계 중 제도·정책부분은 '14년 15억 원에서 '17년 76억 원으로 연평균 71.4%, 질환동물 플랫폼 단계는 연평균 38.1%(('14) 66억 원 → ('17) 167억 원)으로 증가한 반면 비임상 플랫폼 단계는 연평균 -40.9%(('14) 399억 원 → ('17) 82억 원)로 감소

○ '17년 대분류 기준 후보물질 도출 및 최적화(1,138억 원, 32.8%), 인프라(959억 원, 27.6%), 임상(488억 원, 14.1%) 순으로 투자가 이루어지고 있음

- 정부 투자는 신약개발단계 후반부로 갈수록 감소하나 인프라에 대한 투자는 확대



[그림 5-4] 신약개발분야 정부 R&D 신약개발단계별 투자 현황(2017)

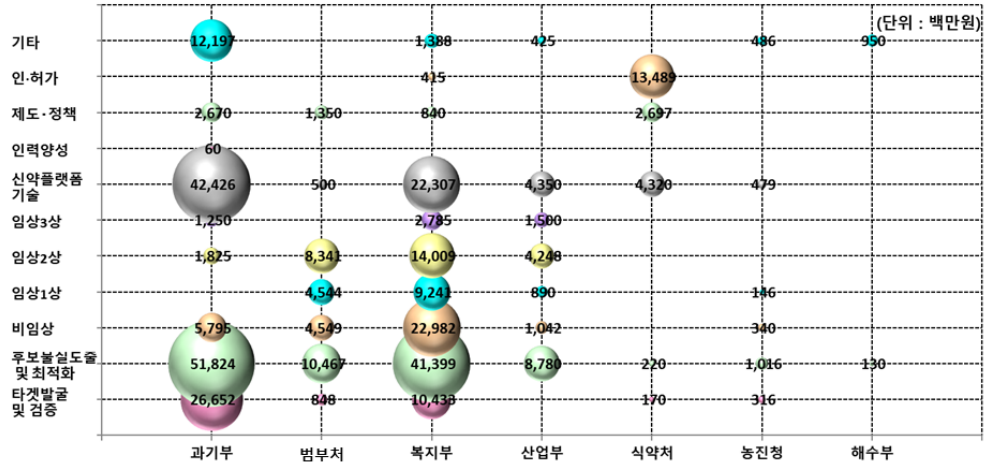
〈표 5-6〉 신약개발분야 정부 R&D 신약개발단계별 투자 현황

구분			2014년		2015년		2016년		2017년		연평균 증가율 (%)
			연구비 (백만원)	비중 (%)	연구비 (백만원)	비중 (%)	연구비 (백만원)	비중 (%)	연구비 (백만원)	비중 (%)	
타켓 발굴 및 검증	타겟 발굴 및 검증		16,637	5.9	15,519	4.6	37,268	12.2	38,419	11.1	32.2
후보물질 도출 및 최적화	후보물질도출 및 최적화		47,422	16.9	42,563	12.7	67,366	22.0	113,837	32.8	33.9
비임상	비임상		49,887	17.8	63,828	19.0	29,678	9.7	34,708	10.0	-11.4
임상	임상1상		19,315	6.9	27,122	8.1	16,417	5.4	14,820	4.3	-8.5
	임상2상		14,619	5.2	16,474	4.9	19,150	6.3	28,423	8.2	24.8
	임상3상		5,597	2.0	7,356	2.2	7,223	2.4	5,535	1.6	-0.4
인프라	신약 플랫폼 기술	타겟발굴 플랫폼	12335	4.4	11942	3.6	7,845	2.6	8,349	2.4	-12.2
		후보물질 발굴 플랫폼	23002	8.2	28271	8.4	26,952	8.8	34,199	9.9	14.1
		비임상 플랫폼	39853	14.2	42978	12.8	24,767	8.1	8,221	2.4	-40.9
		질환동물 플랫폼	6351	2.3	6485	1.9	17,870	5.8	16,718	4.8	38.1
		임상 플랫폼	15245	5.4	18735	5.6	9,632	3.1	6,895	2.0	-23.2
	인력양성		200	0.1	1,150	0.3	-	-	60	0.0	-33.1
	제도·정책		1,500	0.5	6,308	1.9	4,426	1.4	7,557	2.2	71.4
	인·허가		13,149	4.7	14,407	4.3	21,107	6.9	13,904	4.0	1.9
기타	기타		15,360	5.5	32,600	9.7	16,190	5.3	15,447	4.5	0.2
합계			280,473	100.0	335,738	100.0	305,891	100.0	347,092	100.0	7.4

□ 부처별 단계별 투자 현황

ㅇ '17년 신약개발분야에 가장 많이 지원한 과기정통부는 임상 이전 단계에 투자를 집중하였으며, 복지부 역시 후보물질 도출 및 최적화 단계(414억 원)에 투자를 주력하였음

- 과기정통부는 후보물질 도출 및 최적화(518억 원), 타겟 발굴 및 검증(267억 원), 신약플랫폼기술(424억 원) 순으로 투자
- 복지부는 후보물질 도출 및 최적화 단계(414억 원), 비임상(230억 원), 신약플랫폼 기술(223억 원) 순으로 투자
- 범부처 및 산업부의 경우 임상단계에의 투자 비중이 타 부처에 비해 높게 나타남



[그림 5-5] 신약개발분야 정부 R&D 부처별 단계별 투자 현황(2017)

〈표 5-7〉 신약개발분야 정부 R&D 부처별 단계별 투자 현황(2017)

(단위: 백만원)

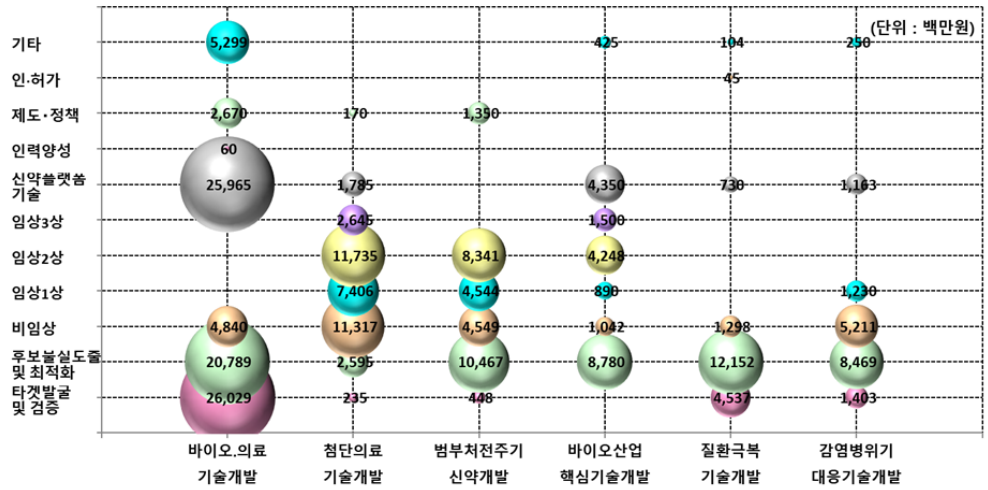
구분		과기정통부	범부처	복지부	산업부	식약처	농진청	해수부	합계
타겟발굴 및 검증		26,652	848	10,433	-	170	316	-	38,419
후보물질도출 및 최적화		51,824	10,467	41,399	8,780	220	1,016	130	113,837
비임상		5,795	4,549	22,982	1,042	-	340	-	34,708
임상1상		-	4,544	9,241	890	-	146	-	14,820
임상2상		1,825	8,341	14,009	4,248	-	-	-	28,423
임상3상		1,250	-	2,785	1,500	-	-	-	5,535
신약 플랫폼 기술	타겟발굴 플랫폼	6,014	200	756	-	900	479	-	8,349
	후보물질 발굴 플랫폼	21,252	300	9,747	1,650	1,250	-	-	34,199
	전임상 플랫폼	3,877	-	2,785	1,500	60	-	-	8,221
	질환동물 플랫폼	10,230	-	3,478	1,200	1,810	-	-	16,718
	임상 플랫폼	1,054	-	5,541	-	300	-	-	6,895
인력양성		60	-	-	-	-	-	-	60
제도·정책		2,670	1,350	840	-	2,697	-	-	7,557
인·허가		-	-	415	-	13,489	-	-	13,904
기타		12,197	-	1,388	425	-	486	950	15,447
합계		144,700	30,600	125,799	21,235	20,896	2,782	1,080	347,092

□ 주요사업별 단계별 투자 현황

o 신약개발분야 주요 사업 중 가장 비중이 큰 과기정통부의 바이오·의료기술개발사업은 타겟 발굴 및 검증(260억 원)에 가장 많이 투자하였으며, 신약플랫폼기술(260억 원), 후보물질 도출 및 최적화(208억 원) 순으로 지원

- 복지부의 첨단의료기술개발은 비임상 및 임상단계를 중점적으로 투자하였으며, 임상2상(117억 원), 비임상(113억 원), 임상1상(74억 원) 순

- 범부처전주기신약개발은 후보물질도출 및 최적화(105억 원)에 가장 많이 투자하고 있으며, 다음으로 임상2상(834억 원), 비임상(46억 원), 임상1상(45억 원) 순으로 지원



[그림 5-6] 신약개발분야 정부 R&D 주요사업별 단계별 투자 현황(2017)

〈표 5-8〉 신약개발분야 정부 R&D 주요사업별 단계별 투자 현황(2017)

(단위: 백만원)

구분		바이오의료 기술개발	첨단의료기술개발	범부처 전주기 신약개발	바이오 산업핵심기술개발	질환극복기술개발	감염병 위기대응기술개발
타겟발굴 및 검증		26,029	235	448	-	4,537	1,403
후보물질도출 및 최적화		20,789	2,595	10,467	8,780	12,152	8,469
비임상		4,840	11,317	4,549	1,042	1,298	5,211
임상1상		-	7,406	4,544	890	-	1,230
임상2상		-	11,735	8,341	4,248	-	-
임상3상		-	2,645	-	1,500	-	-
신약 플랫폼 기술	타겟발굴 플랫폼	5,654	-	-	-	-	-
	후보물질 발굴 플랫폼	8,572	380	-	1,650	650	633
	전임상 플랫폼	885	-	-	1,500	-	-
	질환동물 플랫폼	9,800	530	-	1,200	80	530
	임상 플랫폼	1,054	875	-	-	-	-
인력양성		60	-	-	-	-	-
제도·정책		2,670	170	1,350	-	-	-
인·허가		-	-	-	-	45	-
기타		5,299	-	-	425	104	250
합계		85,651	37,888	29,700	21,235	18,866	17,726

다. 의약품 종류별 정부 R&D 투자 현황

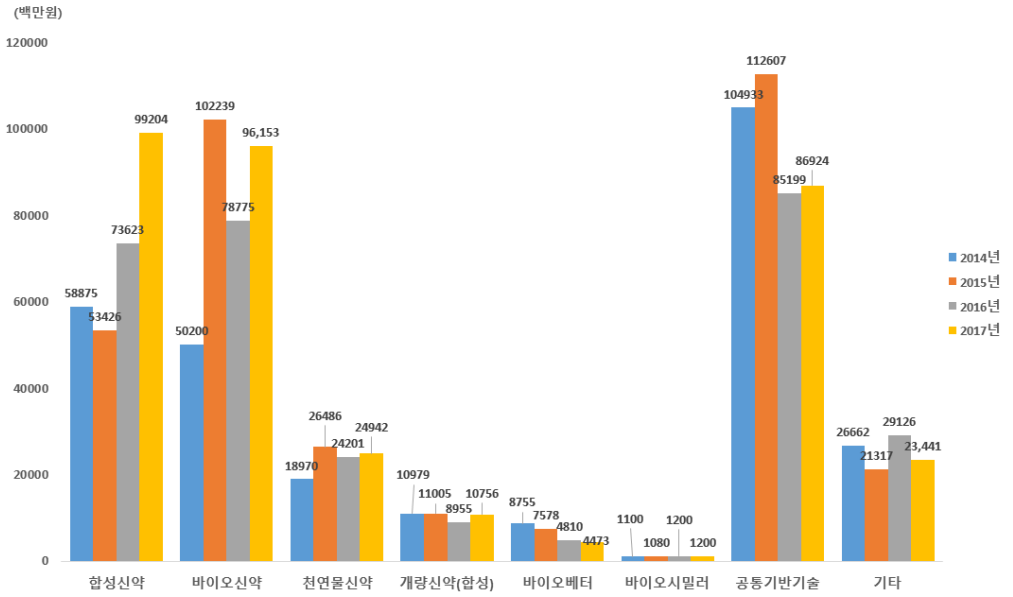
□ 의약품 종류별 투자 현황

o 바이오신약에 대한 투자는 연평균 약 24.2%의 수준으로 증가한 반면 바이오 베터에 대한 투자는 연평균 약 -20.1% 수준으로 감소

- 바이오신약 중 백신에 대한 투자는 '14년 79억 원에서 '17년 242억 원으로 가장 많이 확대됨(연평균 약 45.0%)

o '17년 기준 투자 규모의 절반 이상이 신약개발(63.5%)*에 집중되어 있으며, 합성신약(992억 원, 28.6%), 바이오신약(962억 원, 27.7%), 공통기반기술(869억 원, 25.0%) 순으로 투자

* 개량신약 포함 시 66.6%



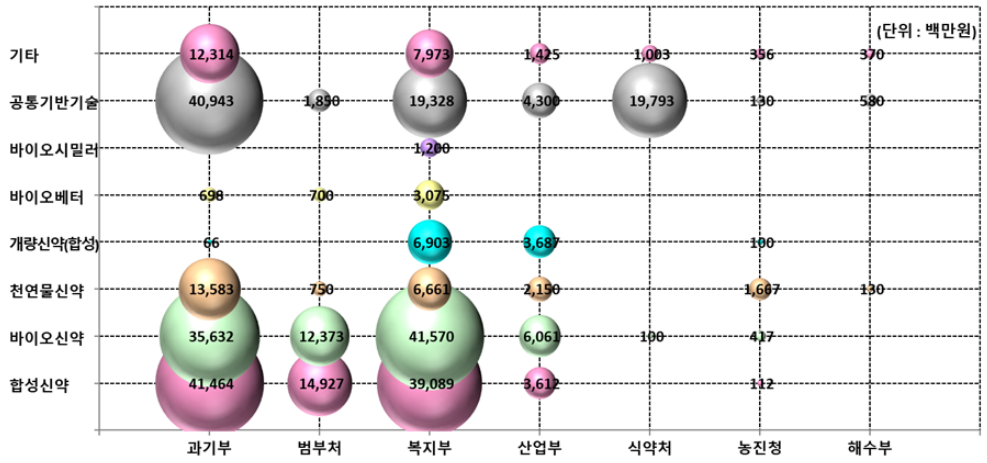
[그림 5-7] 신약개발분야 정부 R&D 의약품 종류별 투자 현황

〈표 5-9〉 신약개발분야 정부 R&D 의약품 종류별 투자 현황

구분		2014년		2015년		2016년		2017년		연평균 증가율 (%)
		연구비 (백만원)	비중 (%)	연구비 (백만원)	비중 (%)	연구비 (백만원)	비중 (%)	연구비 (백만원)	비중 (%)	
합성신약		58,875	21.0	53,426	15.9	73,623	24.1	99,204	28.6	19.0
바이오신약	단백질 치료제	12,810	4.6	18,180	5.4	14,384	4.7	18,888	5.4	13.8
	유전자 치료제	9,362	3.3	11,426	3.4	7,632	2.5	10,451	3.0	3.7
	세포 치료제	7,107	2.5	45,793	13.6	17,313	5.7	19,714	5.7	40.5
	백신	7,923	2.8	13,049	3.9	25,878	8.5	24,171	7.0	45.0
	항체	12,998	4.6	13,791	4.1	13,568	4.4	22,930	6.6	20.8
천연물신약		18,970	6.8	26,486	7.9	24,201	7.9	24,942	7.2	9.6
개량신약(합성)		10,979	3.9	11,005	3.3	8,955	2.9	10,756	3.1	-0.7
바이오베터	단백질 치료제	2,665	1.0	2,050	0.6	1,573	0.5	3,008	0.9	4.1
	유전자 치료제	200	0.1	250	0.1	-	-	-	-	-
	세포 치료제	75	0.0	130	0.0	-	-	-	-	-
	백신	3,980	1.4	3,883	1.2	1,523	0.5	575	0.2	-47.5
	항체	1,835	0.7	1,265	0.4	1,714	0.6	890	0.3	-21.4
바이오 시밀러		1,100	0.4	1,080	0.3	1,200	0.4	1,200	0.3	2.9
공통기반기술		104,933	37.4	112,607	33.5	85,199	27.9	86,924	25.0	-6.1
기타		26,662	9.5	21,317	6.3	29,126	9.5	23,441	6.8	-4.2
총 합계		280,473	100.0	335,738	100.0	305,891	100.0	347,092	100.0	7.4

□ 부처별 의약품종류별 투자 현황

- ㅇ 과기정통부는 합성신약(415억 원), 공통기반기술(409억 원), 바이오신약(356억 원) 순으로 투자
- ㅇ 복지부는 바이오신약(416억 원)의 투자가 높고 합성신약(391억 원), 공통기반기술(193억 원), 개량신약(합성)(69억 원) 순으로 지원
- ㅇ 범부처는 합성신약(149억 원), 산업부는 바이오신약(60억 원), 식약처는 공통기반기술(198억 원)에 가장 많이 투자하는 것으로 나타남



[그림 5-8] 신약개발분야 정부 R&D 부처별 의약품종류별 투자 현황(2017)

〈표 5-10〉 신약개발분야 정부 R&D 부처별 의약품종류별 투자 현황(2017)

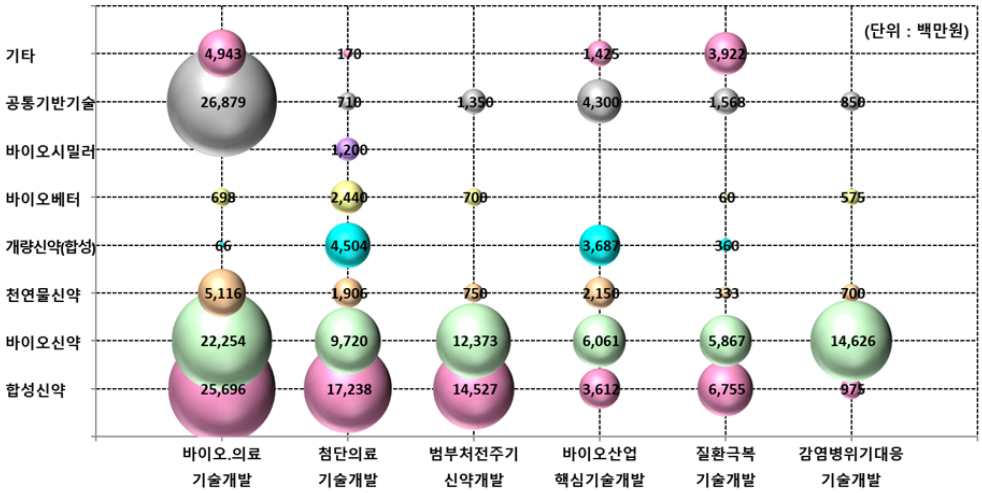
(단위: 백만원)

구분		과기정통부	범부처	복지부	산업부	식약처	농진청	해수부	합계
합성신약		41,464	14,927	39,089	3,612	-	112	-	99,204
바이오신약	단백질 치료제	5,619	4,820	5,222	2,810	-	417	-	18,888
	유전자 치료제	5,005	3,000	2,446	-	-	-	-	10,451
	세포 치료제	9,845	-	9,188	681	-	-	-	19,714
	백신	3,985	594	18,593	900	100	-	-	24,171
	항체	11,179	3,960	6,121	1,670	-	-	-	22,930
천연물신약		13,583	750	6,661	2,150	-	1,667	130	24,942
개량신약(합성)		66	-	6,903	3,687	-	100	-	10,756
바이오베터	단백질 치료제	698	700	1,610	-	-	-	-	3,008
	유전자 치료제	-	-	-	-	-	-	-	-
	세포 치료제	-	-	-	-	-	-	-	-
	백신	-	-	575	-	-	-	-	575
	항체	-	-	890	-	-	-	-	890
바이오시밀러		-	-	1,200	-	-	-	-	1,200
공통기반기술		40,943	1,850	19,328	4,300	19,793	130	580	86,924
기타		12,314	-	7,973	1,425	1,003	356	370	23,441
합계		144,700	30,600	125,799	21,235	20,896	2,782	1,080	347,092

□ 주요사업별 의약품종류별 투자 현황

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

- 과기정통부의 바이오·의료기술개발은 공통기반기술(269억 원)이 가장 높은 투자 비중을 보였으며, 복지부의 첨단의료기술개발은 합성신약(173억 원)의 투자가 높았음



[그림 5-9] 신약개발분야 정부 R&D 주요사업별 의약품종류별 투자 현황(2017)

〈표 5-11〉 신약개발분야 정부 R&D 주요사업별 의약품종류별 투자 현황(2017)

(단위: 백만원)

구분		바이오.의료 기술개발	첨단의료기술 개발	범부처 전주기 신약개발	바이오산업핵심 기술개발	질환극복기술 개발	감염병 위기대응기술 개발
합성신약		25,696	17,238	14,527	3,612	6,755	975
바이오 신약	단백질 치료제	3,306	1,559	4,820	2,810	2,157	225
	유전자 치료제	2,540	430	3,000	-	947	400
	세포 치료제	7,425	5,631	-	681	1,366	125
	백신	1,370	375	594	900	588	13,876
	항체	7,613	1,725	3,960	1,670	808	-
천연물신약		5,116	1,906	750	2,150	333	700
개량신약		66	4,504	-	3,687	360	-
바이오 베터	단백질 치료제	698	1,550	700	-	60	-
	유전자 치료제	-	-	-	-	-	-
	세포 치료제	-	-	-	-	-	-
	백신	-	-	-	-	-	575
	항체	-	890	-	-	-	-
바이오시밀러		-	1,200	-	-	-	-
공통기반기술		26,879	710	1,350	4,300	1,568	850
기타		4,943	170	-	1,425	3,922	-
합계		85,651	37,888	29,700	21,235	18,866	17,726

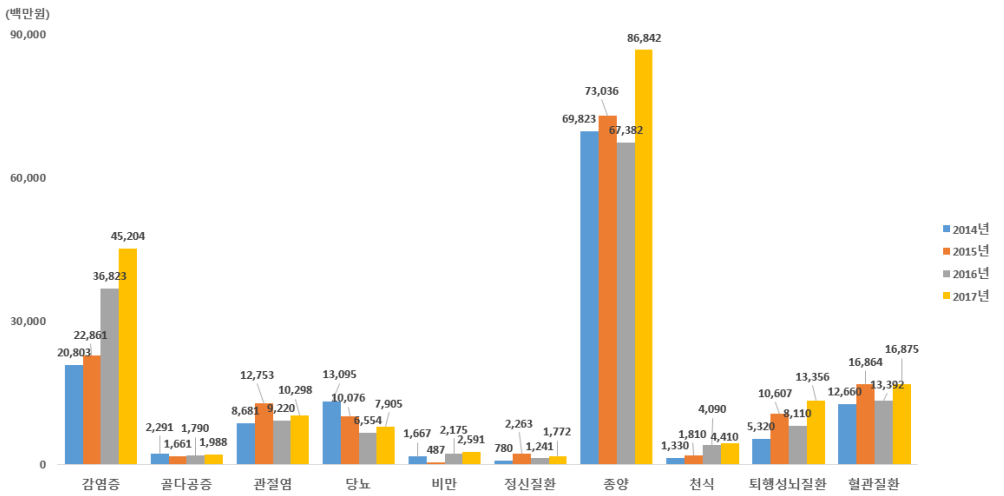
라. 질환별 정부 R&D 투자 현황

□ 질환별 투자 현황

ㅇ '17년 신약개발분야 정부 R&D는 종양(868억 원, 25.0%) 과제에 투자가 높았음

- 천식(연평균 49.1%), 퇴행성뇌질환(연평균 35.9%), 정신질환(연평균 31.5%), 감염증(연평균 29.5%)에 대한 투자도 확대되었음

- 반면, 당뇨는 연평균 -15.5%, 골다공증은 -4.6% 수준으로 투자가 감소되고 있음



[그림 5-10] 신약개발분야 정부 R&D 질환별 투자 현황

〈표 5-12〉 신약개발분야 정부 R&D 질환별 투자 현황

구분	2014년 ·		2015년 ·		2016년 ·		2017년		연평균 증가율 (%)
	연구비 (백만원)	비중 (%)	연구비 (백만원)	비중 (%)	연구비 (백만원)	비중 (%)	연구비 (백만원)	비중 (%)	
감염증	20,803	7.4	22,861	6.8	36,823	12.0	45,204	13.0	29.5
골다공증	2,291	0.8	1,661	0.5	1,790	0.6	1,988	0.6	-4.6
관절염	8,681	3.1	12,753	3.8	9,220	3.0	10,298	3.0	5.9
당뇨	13,095	4.7	10,076	3.0	6,554	2.1	7,905	2.3	-15.5
비만	1,667	0.6	487	0.1	2,175	0.7	2,591	0.7	15.8
정신질환	780	0.3	2,263	0.7	1,241	0.4	1,772	0.5	31.5
종양	69,823	24.9	73,036	21.8	67,382	22.0	86,842	25.0	7.5

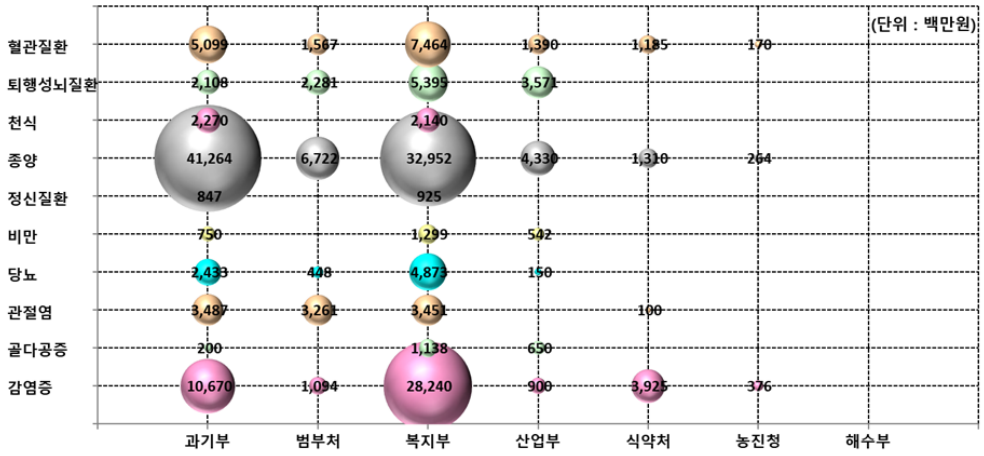
구분	2014년		2015년		2016년		2017년		연평균 증가율 (%)
	연구비 (백만원)	비중 (%)	연구비 (백만원)	비중 (%)	연구비 (백만원)	비중 (%)	연구비 (백만원)	비중 (%)	
천식	1,330	0.5	1,810	0.5	4,090	1.3	4,410	1.3	49.1
퇴행성 뇌질환	5,320	1.9	10,607	3.2	8,110	2.7	13,356	3.8	35.9
혈관질환	12,660	4.5	16,864	5.0	13,392	4.4	16,875	4.9	10.1
기타	144,025	51.4	183,320	54.6	155,113	50.7	155,851	44.9	2.7
합계	280,473	100.0	335,738	100.0	305,891	100.0	347,092	100.0	7.4

□ 부처별 질환별 투자 현황

o 부처별로 질환별 투자 현황을 분석한 결과 과기정통부·범부처·복지부·산업부 모두 종양(413억 원, 67억 원, 330억 원, 43억 원)에 가장 많이 투자함

- 과기정통부와 복지부는 종양 다음으로 감염증(107억 원, 282억 원)에 대한 투자가 높았음

- 식약처와 농진청은 감염증(39억 원, 4억 원)에 대한 투자 비중이 가장 높았음



[그림 5-11] 신약개발분야 정부 R&D 부처별 질환별 투자 현황(2017)

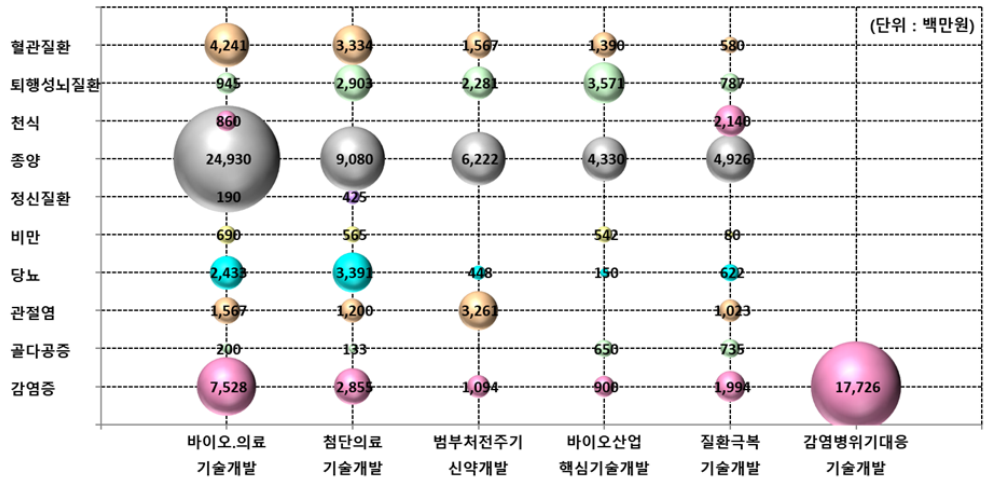
〈표 5-13〉 신약개발분야 정부 R&D 부처별 질환별 투자 현황(2017)

(단위: 백만원)

구분	과기정통부	범부처	복지부	산업부	식약처	농진청	해수부	합계
감염증	10,670	1,094	28,240	900	3,925	376	-	45,204
골다공증	200	-	1,138	650	-	-	-	1,988
관절염	3,487	3,261	3,451	-	100	-	-	10,298
당뇨	2,433	448	4,873	150	-	-	-	7,905
비만	750	-	1,299	542	-	-	-	2,591
정신질환	847	-	925	-	-	-	-	1,772
종양	41,264	6,722	32,952	4,330	1,310	264	-	86,842
천식	2,270	-	2,140	-	-	-	-	4,410
퇴행성뇌질환	2,108	2,281	5,395	3,571	-	-	-	13,356
혈관질환	5,099	1,567	7,464	1,390	1,185	170	-	16,875
기타	75,572	15,227	37,922	9,702	14,376	1,972	1,080	155,851
합계	144,700	30,600	125,799	21,235	20,896	2,782	1,080	347,092

□ 주요사업별 질환별 투자 현황

ㅇ 대부분의 주요사업에서도 종양에 가장 많이 투자하고 있는 것으로 나타남



[그림 5-12] 신약개발분야 정부 R&D 부처별 질환별 투자 현황(2017)

〈표 5-14〉 신약개발분야 정부 R&D 부처별 질환별 투자 현황(2017)

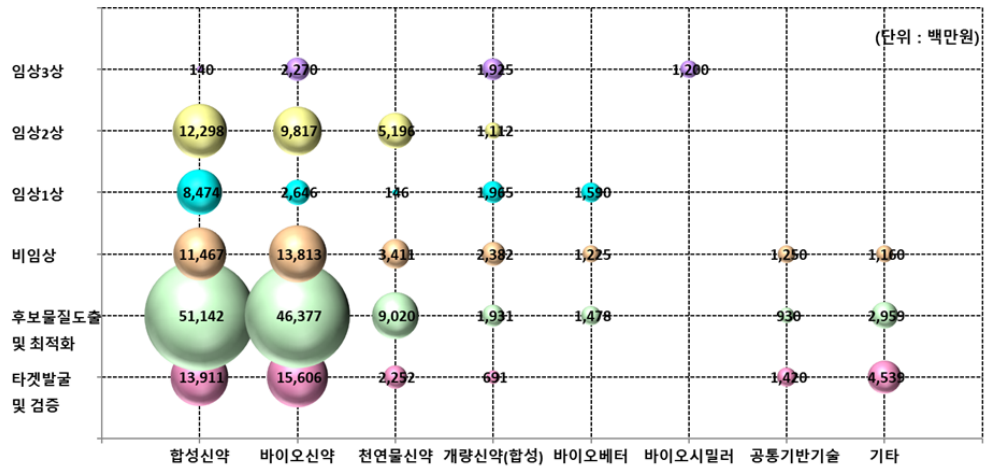
(단위: 백만원)

구분	바이오. 의료기술개발	첨단의료기술 개발	범부처 전주기 신약개발	바이오산업핵심 기술개발	질환극복기술개발	감염병 위기대응 기술개발
감염증	7,528	2,855	1,094	900	1,994	17,726
골다공증	200	133	-	650	735	-
관절염	1,567	1,200	3,261	-	1,023	-
당뇨	2,433	3,391	448	150	622	-
비만	690	565	-	542	80	-
정신질환	190	425	-	-	-	-
종양	24,930	9,080	6,222	4,330	4,926	-
천식	860	-	-	-	2,140	-
퇴행성 뇌질환	945	2,903	2,281	3,571	787	-
혈관질환	4,241	3,334	1,567	1,390	580	-
기타	42,067	14,002	14,827	9,702	5,978	-
합계	85,651	37,888	29,700	21,235	18,866	17,726

마. 교차분석

□ 신약개발단계별 의약품종류별 투자 현황

- o 합성·바이오·천연물신약은 후보물질도출 및 최적화 단계에 투자가 집중
- o 합성신약은 후보물질도출 및 최적화(511억 원), 타겟발굴 및 검증(139억 원), 임상2상(123억 원) 순으로 투자하고 있으며, 바이오신약은 후보물질도출 및 최적화(464억 원), 타겟 발굴 및 검증(156억 원), 비임상(138억 원) 순으로 임상 전 단계에 집중 투자
- o 천연물신약은 후보물질도출 및 최적화(90억 원), 임상2상(52억 원), 비임상(34억 원) 단계 순으로 투자
- o 개량신약(합성)은 비임상(24억 원), 임상1상(20억 원), 임상3상(19억 원) 순으로 지원
- o 바이오베터는 임상1상(16억 원) 단계에 가장 많이 투자되었으며, 바이오시밀러는 임상3상(12억 원)에만 투자되었음



[그림 5-13] 신약개발분야 정부 R&D 신약개발단계별 의약품종류별 투자 현황(2017)

〈표 5-15〉 신약개발분야 정부 R&D 신약개발단계별 의약품종류별 투자 현황(2017)

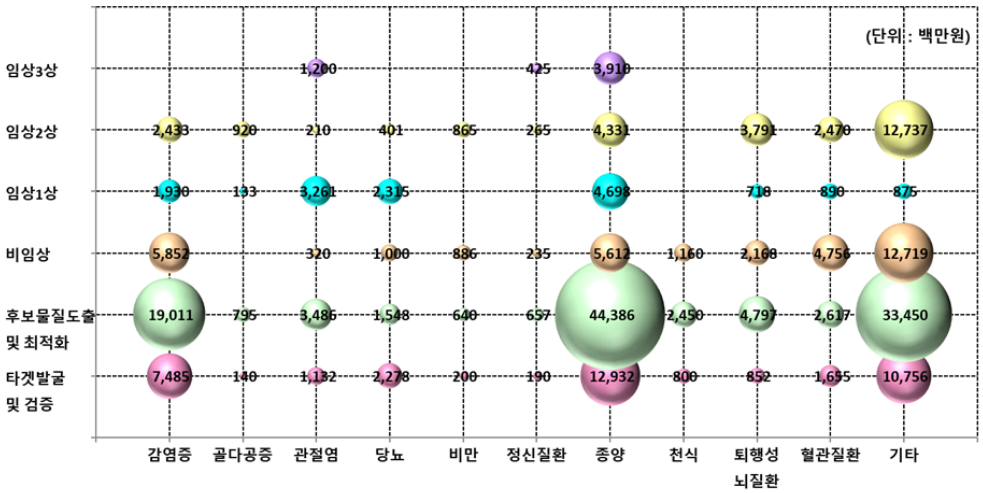
(단위: 백만원)

구분		합성신약	바이오신약					천연물신약	개량신약	바이오베터					바이오시밀러	공통기반기술	기타	합계
			단백질치료제	유전자치료제	세포치료제	백신	항체			단백질치료제	유전자치료제	세포치료제	백신	항체				
타겟발 굴 및 검증		13,911	2,358	2,320	2,740	1,740	6,448	2,252	691	-	-	-	-	-	-	1,420	4,539	38,419
후보물질도출 및 최적화		51,142	11,296	3,156	8,470	11,653	11,802	9,020	1,931	-	-	1,278	200	-	-	930	2,959	113,837
비임상		11,467	1,215	2,870	2,000	5,387	2,340	3,411	2,382	-	-	850	375	-	-	1,250	1,160	34,708
임상1상		8,474	-	-	133	1,230	1,283	146	1,965	-	-	700	-	890	-	-	-	14,820
임상2상		12,298	3,523	1,825	3,781	653	36	5,196	1,112	-	-	-	-	-	-	-	-	28,423
임상3상		140	-	-	1,020	1,250	-	-	1,925	-	-	-	-	-	1,200	-	-	5,535
신약 플랫폼 기술	타겟 발굴 플랫폼	750	150	-	-	-	327	479	-	-	-	-	-	-	-	6,643	-	8,349
	후보물질 발굴 플랫폼	-	-	280	721	1,133	333	3,423	250	-	-	180	-	-	-	27,472	407	34,199
	전임상 플랫폼	212	166	-	-	-	90	316	-	-	-	-	-	-	-	7,437	-	8,221
	질환동물 플랫폼	67	-	-	-	280	-	-	-	-	-	-	-	-	-	16,372	-	16,718
	임상 플랫폼	-	180	-	600	475	187	-	500	-	-	-	-	-	-	4,853	100	6,895
인력양성		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	60	60
제도·정책		-	-	-	-	-	-	-	-	-	-	-	-	-	-	5,552	2,005	7,557
인·허가		-	-	-	-	370	-	-	-	-	-	-	-	-	-	12,831	703	13,904
기타		743	-	-	249	-	83	699	-	-	-	-	-	-	-	2,165	11,508	15,447
합계		99,204	18,888	10,451	19,714	24,171	22,930	24,942	10,756	-	-	3,008	575	890	1,200	86,924	23,441	347,092

□ 신약개발단계별 질환별 투자 현황

○ 가장 높은 투자를 보인 종양 및 감염증의 경우 후보물질도출 및 최적화 단계(444억 원, 190억 원)와 타겟 발굴 및 검증 단계(129억 원, 75억 원)에 가장 많이 투자되었음

- 타 질환의 경우 비교적 고른 양상의 투자를 나타내었음
- 혈관질환은 비임상(48억 원), 관절염 및 퇴행성뇌질환은 후보물질 도출 및 최적화(35억 원, 48억 원)단계 투자 비중이 높았음



[그림 5-14] 신약개발분야 정부 R&D 신약개발단계별 질환별 투자 현황(2017)

〈표 5-16〉 신약개발분야 정부 R&D 신약개발단계별 질환별 투자 현황(2017)

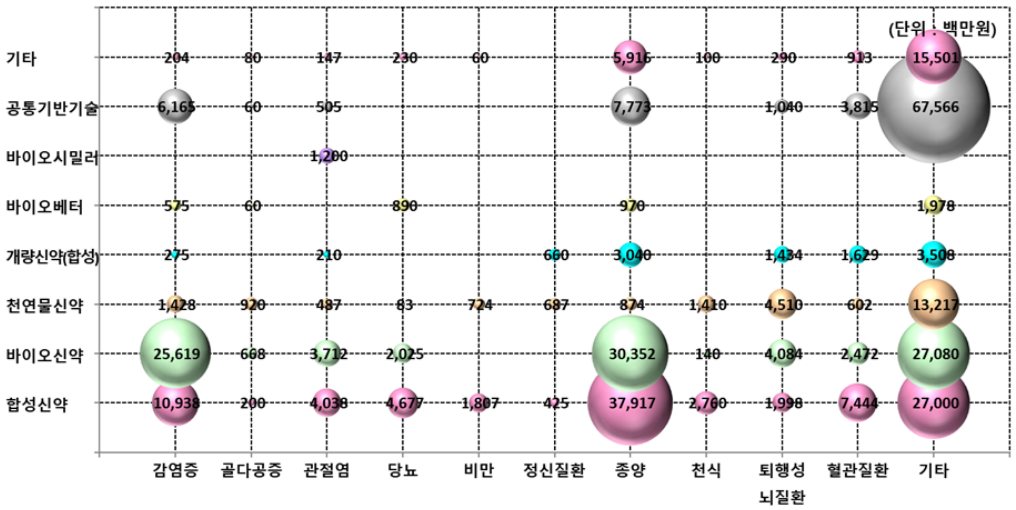
(단위: 백만원)

구분		감염증	골다공증	관절염	당뇨	비만	정신질환	종양	천식	퇴행성 뇌질환	혈관질환	기타	합계
타겟 발굴 및 검증		7,485	140	1,132	2,278	200	190	12,932	800	852	1,655	10,756	38,419
후보물질도출 및 최적화		19,011	795	3,486	1,548	640	657	44,386	2,450	4,797	2,617	33,450	113,837
비임상		5,852	-	320	1,000	886	235	5,612	1,160	2,168	4,756	12,719	34,708
임상1상		1,930	133	3,261	2,315	-	-	4,698	-	718	890	875	14,820
임상2상		2,433	920	210	401	865	265	4,331	-	3,791	2,470	12,737	28,423
임상3상		-	-	1,200	-	-	425	3,910	-	-	-	-	5,535
신약 플랫폼 기술	타겟 발굴 플랫폼	130	-	327	-	-	-	1,489	-	-	-	6,403	8,349
	후보물질 발굴 플랫폼	2,583	-	-	-	-	-	1,766	-	110	1,660	28,080	34,199
	전임상 플랫폼	-	-	-	212	-	-	743	-	150	820	6,297	8,221
	질환동물 플랫폼	790	-	-	-	-	-	3,096	-	650	290	11,893	16,718
	임상 플랫폼	662	-	180	-	-	-	460	-	-	320	5,273	6,895
인력양성		-	-	-	-	-	-	-	-	-	-	60	60
제도·정책		125	-	100	-	-	-	450	-	-	-	6,882	7,557
인·허가		3,760	-	-	-	-	-	500	-	-	1,185	8,459	13,904
기타		444	-	83	150	-	-	2,469	-	120	212	11,968	15,447
합계		45,204	1,988	10,298	7,905	2,591	1,772	86,842	4,410	13,356	16,875	155,851	347,092

□ 의약품종류별 질환별 투자 현황

○ 투자가 가장 높은 종양은 합성신약(379억 원), 감염증은 바이오신약(256억 원)분야의 투자가 가장 높게 나타남

- 전반적으로 개량신약분야보다는 보다는 새로운 신약을 개발하는 분야에 투자 비중이 높았음



[그림 5-15] 신약개발분야 정부 R&D 의약품종류별 질환별 투자 현황(2016)

〈표 5-17〉 신약개발분야 정부 R&D 의약품종류별 질환별 투자 현황(2017)

(단위: 백만원)

구분		감염증	골다공증	관절염	당뇨	비만	정신질환	종양	천식	퇴행성 뇌질환	혈관질환	기타	합계
합성신약		10,938	200	4,038	4,677	1,807	425	37,917	2,760	1,998	7,444	27,000	99,204
바이오신약	단백질 치료제	1,037	535	315	-	-	-	2,240	-	500	900	13,361	18,888
	유전자 치료제	1,000	-	120	-	-	-	5,604	-	100	280	3,347	10,451
	세포 치료제	181	133	2,120	1,577	-	-	7,858	140	1,506	1,292	4,906	19,714
	백신	21,013	-	-	-	-	-	2,650	-	-	-	508	24,171
	항체	2,389	-	1,157	448	-	-	11,999	-	1,979	-	4,957	22,930
천연물신약		1,428	920	487	83	724	687	874	1,410	4,510	602	13,217	24,942
개량신약(합성)		275	-	210	-	-	660	3,040	-	1,434	1,629	3,508	10,756
바이오베터	단백질 치료제	-	60	-	-	-	-	970	-	-	-	1,978	3,008
	유전자 치료제	-	-	-	-	-	-	-	-	-	-	-	-
	세포 치료제	-	-	-	-	-	-	-	-	-	-	-	-
	백신	575	-	-	-	-	-	-	-	-	-	-	575
	항체	-	-	-	890	-	-	-	-	-	-	-	890
바이오시밀러		-	-	1,200	-	-	-	-	-	-	-	-	1,200
공통기반기술		6,165	60	505	-	-	-	7,773	-	1,040	3,815	67,566	86,924
기타		204	80	147	230	60	-	5,916	100	290	913	15,501	23,441
합계		45,204	1,988	10,298	7,905	2,591	1,772	86,842	4,410	13,356	16,875	155,851	347,092

바. 분석결과 및 시사점

□ 신약개발단계별 투자의 경우 인력양성, 인허가 등의 인프라 부분에 대한 투자를 보다 강화 할 필요

ㅇ 후보물질 도출 및 최적화 단계(약 32.8%)의 비중이 높은 반면, 인프라 중 인력양성, 제도·정책, 인·허가 부분의 투자 비중은 4% 이하 수준

ㅇ 신약개발 촉진을 통한 동 분야의 성장을 모색하기 위해서는 제도·정책 및 인프라적인 부분에 대한 정부 차원의 지원이 필요[24)]

- 제도·정책에 대한 지원은 높은 증가율(연평균 71.4%)을 보이고 있으나, 인허가 및 인력 양성을 위한 지원은 여전히 미비한 수준

- 신약개발 각 단계에 대한 인력양성은 물론, 정부/민간 영역을 고려하여 임상 후 각 단계의 리스크 관리 전문인력에 대한 지원 필요성 검토

□ 의약품종류별 투자의 경우 바이오 신약 및 합성신약에 대한 투자가 연평균 각 24.2%, 19.0% 수준으로 증가하며 글로벌 신약개발 트렌드 변화에 부합[25)]

□ 질환분야 중 가장 높은 투자가 이루어지는 종양질환(25.0%)은 연평균 7.5% 증가율을 보이고 있으며, 천식, 퇴행성뇌질환, 정신질환에 대한 투자가 급증(연평균 증가율 각각 49.1%, 35.9%, 31.5%)

ㅇ 환경적 변화 및 치매, 신종 감염병 등에 대한 대응을 위한 공공적 보건의료 R&D 지원을 강화한 것으로 보임

□ 부처별 투자의 경우 「바이오 중기('16~'18) 육성전략(안)(2016)」에서는 신약개발 관련 부처의 역할분담(안) 에 따라 적절한 투자가 이루어지도록 노력할 필요

※ (과기정통부) 기초연구~후보물질최적화 단계, (복지부) 비임상 및 임상단계, (산업부) IP사업화, (식약처) 허가 및 컨설팅

24) 신영기 (2010), "우리나라 신약개발의 주요 현안 및 대응방안", 「과학기술정책 」, 178 : 39-42.
25) 생명공학정책연구센터 (2018), "의약품 유형별(합성, 바이오) 개발 특성", 「BioINwatch(BioIN+ Issue+Watch) 」: 18-23

제6장

결론

제6장 결론

가. 개선 및 활용방안

1) 분석·활용 모형 고도화

□ 분석·활용 모형 고도화 결과 사업 간의 관계성(유사정도) 분석, 특정 사업 속성변화, 분류 모형의 범용화 등에서 유의미한 결과를 도출함

- ㅇ 연구개발사업의 분석하는 과정에서 어떠한 사업이 서로 간에 유사한지는 사업의 기획, 평가 등에 고려되는 사항으로 동 분석·활용 모형이 시각적·정량적으로 분석 결과를 도출한다는 점에서 의미가 있다고 사료됨

□ 의약과제 분류모형의 경우 범용화를 통해 사용자 접근성을 제고함은 물론, 딥러닝 방법론에 국한된 기 모형을 고도화하여 6개의 방법론을 적용하여 사용자 선택의 폭을 넓힘

- ㅇ 모형 고도화 과정에서 데이터 축적에 의해 전년대비 성능이 향상되는 것을 확인하였고, 이를 통해 딥러닝 방법론의 우수성 및 강점 또한 확인할 수 있었음

□ 분류모형을 활용하여 신규과제를 선분류하고 이를 전문가에게 제공함으로써 의약과제 분류연구의 추진효율성 제고가 가능할 것으로 판단됨

- ㅇ 동 연구 과제를 통해 모형의 과제분류결과와 전문가 분류결과 간 유사한 투자포트폴리오가 구성됨을 확인할 수 있었음
- ㅇ 따라서, 의약과제 분류과정에서 동 분류모형을 의사결정지원모형으로 활용할 경우 연구 추진효율성 개선이 가능할 것으로 보임
- ㅇ 추가적으로, 동 과제 수행 과정에서 모형의 과제분류 결과를 검토하고 직접 의약과제 분류를 수행한 전문가로부터 제안된 개선사항은 다음과 같음
  - (의약품종류) 줄기세포 관련 기술, 의약품 개발이 아닌 경우(IVD, 동반진단 등)를 사전에 필터링 할 경우 예측 성능이 개선될 여지가 있음
  - (개발단계) 특수한 상황을 제외하고 대체적으로 양호한 예측을 보였으나 단계가 중첩된 임상시험(1,2상, 비임상-임상1상)을 연구내용으로 포함하고 있는 경우는 과제 추진시점을 고려하여 분류를 수행할 필요
  - (대상질환) 보다 범용적인 질환분류 체계 활용이 요구됨

□ 그럼에도, 동 과제모형이 사용자의 의사결정을 지원하는데 적합한 모형(전문가시스템)임을 숙지할 필요

ㅇ 모든 잠긴 문을 열 수 있는 만능열쇠가 존재하지 않듯이, 동 과제 모형도 최신 기계학습 방법을 적용한 모형일지라도 일부 한계점이 존재하는 것을 확인함

- 사전에 등재되지 않은 단어의 경우 유사성, 연관성 분석에 개선이 요구되는 것으로 판단됨
- 사전에 과학기술용어를 적극 추가하여 이러한 단점을 극복할 필요

ㅇ 관련성 있는 과제를 판단 시 일반적으로 사용되는 유사단어 출현빈도에 근거하지 않기 때문에 다량의 문장을 활용 시 오히려 결과가 좋지 않는 경우도 발생 가능함

- 내재된 공간상에서 벡터 값을 기반으로 유사도를 측정하는 과정에서 불필요한 문맥의 이입은 해당 입력 내용이 적정한 값을 나타내는 공간에 배치되는 것을 방해할 수 있음

ㅇ 이는 실질적으로 최신 기계학습 방법론을 적용한 모형이 인간과 같이 문맥이나 핵심키워드, 행간에서 정보를 이해한다기보다는 사용자가 입력한 질문에 대한 답을 과거보다 정확성 높게 찾아준다는 의미로 해석될 수 있음

ㅇ 관련 과제를 신속히 도출하기 위한 연구과제간 관계성 분석 기능의 경우 현재 활발히 사용되는 키워드 중심의 접근 방법을 접목하는 등의 개선방안을 마련하여 기능을 보다 고도화 할 필요가 있음

□ 임베딩 기술을 적용하여 비정형데이터인 텍스트데이터를 벡터화하여 정형화된 데이터로 변환하고 이를 바탕으로 클러스터링, 분류 연구를 시도한 점은 동 연구과제의 특이점으로 볼 수 있음

□ 향후 추가적인 연구를 통해 동 연구과제에서 제시된 접근법의 유용성이 검증될 경우 텍스트 기반 정보서비스 제공분야에 활용이 가능할 것으로 판단됨

ㅇ 2018년 10월 구글에서 발표한 BERT[26]라는 새로운 인공지능 언어모델은 당초 딥러닝 기반 텍스트마이닝 방법론을 포함하여 현존하는 대부분의 방법론을 뛰어넘는 성능을 보여주고 있음

26) J. Devlin et al., (2018), BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv:1810.04805.

- BERT 발표 이후 카카오 등 국내 대표 IT 기업에서도 BERT를 활용한 연구를 지속 수행하며 우수성을 확인하고 있음

ㅇ BERT의 활용이 보편화되고, 이를 과학기술지식정보에 접목할 경우 동 연구과제에서 수행한 클러스터링 및 분류방법론의 정확도가 한층 제고되어 실제 업무 활용의 개연성 또한 한층 더 높아질 것으로 사료됨

2) 딥러닝 기반 토픽 클러스터링 분석 방법론 연구

□ 연구과제와 일반적 문헌정보(논문, 특허 등)의 근본적 차이점을 고려한 접근 필요

ㅇ 일반적 문헌정보는 연구과제의 결과물에 해당하는 것으로, 연구과제가 논문이나 특허로 이어지는데는 보통 수년의 시간이 소요된다는 점을 고려할 필요

- 국내 과제정보와 해외 문헌정보를 직접적으로 비교 시, 이 시차로 인한 착시현상을 고려하여 해석해야 함
- 또한 이 시차로 인하여 문헌정보는 연구개발 트렌드(현황) 분석에 참고하는 수준으로만 활용해야 하며, 연구개발 투자에 대한 제언을 도출하기에는 제한적임

ㅇ 논문이나 특허는 연구결과 또는 지식재산으로 보호받고자 하는 범위가 명확하게 설정되어 연구단계 또는 연구 분야가 대개 특정되는 반면, 연구과제는 과제의 성격에 따라 여러 연구단계 또는 분야를 포괄하는 경우가 있음

- 특히 사업단 과제와 같은 내역사업 수준의 내용을 포괄하는 대형과제의 경우 기초연구에서 개발연구 까지 포괄하는 등 워드 임베딩 과정에서 한 개의 점(node)으로 간주하기 어려운 성격을 가짐
- 조사·분석 DB에 포함된 연구내용(목적, 내용, 기대효과)은 연구제안서를 기반으로 작성된 경우가 많으므로, 아이디어 수준의 다양한 아이템을 담는 경우가 있어 연구분야도 다양할 수 있음

ㅇ 조사·분석 DB 상의 연구과제 중 일부는 인프라 또는 기관운영비(기평비) 등 특정 연구 분야로 분류할 수 없는 성격으로, 기계학습 시 외란(disturbance)으로 작용할 수 있음

- 특히 출연연의 경우 기관 전체의 장비구입비가 한 개의 세부과제로 편성되어 있는 경우가 있어, 데이터 정제 과정에서 제외시키거나 하는 고려가 필요함

- ㅇ 동 과제에서는 데이터의 양 및 품질이 높은 조사·분석 데이터를 보다 고도화된 방법론으로 분석하기 위해 기계학습을 적용하였으며, 향후에는 일반 문헌정보와는 차별화되는 연구과제의 특성을 고려한 방법론 개선이 필요함

□ 연구비 규모에 따른 가중치 설정 및 텍스트 표준화를 통한 학습 품질 제고 필요

- ㅇ 동 과제에서는 과제의 연구비에 따른 차이를 두지 않고 각 과제에 동일한 가중치를 두었는데, 이는 분석결과의 신뢰도에 영향을 미침
  - 사업 간의 유사도를 판단하거나, 토픽 클러스터링(그룹화) 수행 시 연간 50억 원이 투입되는 과제와 5천만원이 투입되는 과제가 동일하게 고려되는 것은 정확하지 않음
- ㅇ Pubmed 서지분석에서도 각 논문의 가중치를 동일하게 볼 것인가 여부의 이슈가 존재하지만, 기본적으로 논문의 가중치의 기준이 되는 수치의 선정이 곤란함
  - 예를 들어, 임팩트 팩터는 분야에 따른 차이가 존재하므로 가중치 설정의 기준이 될 경우 오히려 편향(bias)된 결과를 야기할 수 있음
- ㅇ 학습 입력 데이터인 연구목적, 연구내용, 기대효과 텍스트의 길이도 과제마다 상이한데, 과제들의 텍스트를 하나로 합쳐서 입력으로 사용한 내역사업 클러스터링 분석의 경우 텍스트가 긴 과제의 비중이 높게 설정되는 문제가 있음
  - 과제들의 연구목적의 길이는 짧게는 20~30단어에서 길게는 수백단어로 작성됨
  - Pubmed 초록의 경우 일반적인 논문 초록의 길이가 비슷한 편(보통 200~300단어)이므로 해당 이슈에 대한 우려는 낮음
- ㅇ 따라서 국내 연구과제의 분석 품질의 제고를 위해서는 연구비 규모를 반영할 수 있도록 알고리즘 개선이 필요하며, 데이터 정제 시 텍스트 길이 편차를 줄일 수 있는 방안이 요구됨

□ Pubmed의 MeSH Subject Heading을 활용한 세계 바이오의료 과학기술정보의 심도 있는 분석이 가능

- ㅇ Pubmed를 운영하는 미국 NLM(National Library of Medicine)에서 개발한 MeSH Subject Heading은 매주 업데이트되는 일종의 바이오분야 키워드 사전임
  - 초록이 확보된 논문에 대해서 NLM 측에서 MeSH 키워드(headings)를 부여하여 연구자가 직접 입력한 키워드뿐만 아니라 보다 표준화된 키워드 체계를 제공함

- MeSH 키워드는 상하관계(hierarchy)를 이루고 있어 대분류·중분류·소분류와 같은 개념을 제공하지만, 분류가 상호 배타적이지 않기 때문에 특정 키워드가 두 개 이상의 상위 키워드를 가질 수 있음

- 각 논문들은 여러 개의 MeSH 키워드가 부여될 수 있으므로 논문 간의 상하관계를 나타내지는 않지만, 이 키워드가 부여된 논문들은 쉽게 분야를 파악할 수 있음

ㅇ MeSH Subject Heading은 크게 세 종류의 키워드로 나눠지며, 지속적으로 확대되고 있기 때문에 최근 논문일수록 더욱 다양한 MeSH 키워드 데이터를 제공함

- (Descriptor) 가장 중심이 되는 키워드 체계로, 16가지 대(大) 키워드 하에 매우 세부적인 키워드 체계가 짜여져 있음

(예시)

Diseases(질환)
└ Eye Diseases(안과 질환)
└ Corneal Diseases(각막 질환)
└ Corneal Opacity(각막 혼탁)
└ Arcus Senilis(노인환)

- (Qualifier) 주로 해부학적 명칭, 약품명, 질환명 등으로 분류된 descriptor를 보완하기 위해 학문 분야 또는 치료법 등의 응용분야를 중심으로 수립된 보조 키워드 체계로, subheadings 라고도 칭함

※ "Liver(간)"이라는 descriptor와 "Drug effects(약물 효과)"라는 qualifier를 조합하여 보다 문헌의 내용을 명확히 설명 가능

- (Supplementary Records) Descriptor를 보완하기 위해 도입된 추가 키워드 체계로, 화합물, 치료법(프로토콜), 질병, 생물 총 4가지 클래스의 키워드가 마련되어 있으며, 상하관계를 이루지 않는 대신 각각 해당하는 descriptor 키워드가 매칭되어 있음

ㅇ MeSH 키워드는 매우 정교하게 설계되어 문헌들의 성격을 잘 표현해주고 있으므로 기계학습의 입력 값으로 적절하다고 사료되나, 반면에 단순 통계를 내기에는 중복 키워드가 많아 적합하지 않음

- 반대로 조사·분석에서 활용하는 분류체계는 중복 분류 없이 상호 배타적으로 설계되어 있어 정부투입연구비 산출 등 통계적 목적으로는 훌륭하나, 연구과제의 내용을 특정하기에 다소 한계가 존재

ㅇ Pubmed의 데이터에 기계학습 방법론을 적용하면 보다 심도 있는 연구동향 분석이 가능할 것으로 예상됨

- 현재의 연구동향 분석은 단순히 키워드 검색을 통해 얻어진 해당 분야의 문헌 수를 중심으로 하고 있어 정밀성에 한계
- Pubmed의 키워드 검색은 단어의 포함여부 중심이므로 doc2vec 같은 딥러닝 방법론을 적용하면 더욱 의미 있는 결과를 도출할 수 있을 것으로 예상
- 또한 Pubmed의 문헌정보는 수백만 건에 달하는 빅데이터에 해당하므로 기계학습 방법론의 장점이 발휘될 수 있음

ㅇ Pubmed에는 우리나라 연구자가 출판한 논문도 등록되어 있으므로 연구자의 소속기관 정보 등을 활용하여 국내 연구동향도 파악 가능

- 향후 바이오의료 과학기술정보 분석에 있어 Pubmed의 활용 확대 필요

□ NTIS에 수집되는 논문 성과 데이터와 Pubmed를 연동한 과학기술 정보 분석을 제안

ㅇ 국가 연구 개발사업 과제 수행을 통하여 생산된 논문은 NTIS 성과정보로 입력되므로, 해당 논문이 Pubmed에 등재된 경우 국내 연구과제와 Pubmed 문헌의 직접적 연결이 가능

ㅇ 연구과제 수행과 논문 출판의 시차 문제를 해결하기 위해, 과제에서 생산된 논문을 분석하고 역으로 과제 및 사업의 연구 분야 분류·분석에 활용하는 접근이 가능

- 연구과제의 내용은 연구계획서를 기반으로 작성되기 때문에 실제 수행 내용 및 결과와 상이할 수 있어 이러한 접근을 통하여 정확성을 제고할 수 있음
- 과제가 아닌 사업 단위로 분석을 수행할 경우 데이터가 충분하므로 논문, 특허 개수뿐만 아니라 결과물의 내용적 분석을 통해 연구 성과가 사업의 목적에 부합했는지 여부를 분석하는 추적평가 차원의 활용이 가능

참고 문헌

참고 문헌

생명공학정책연구센터 (2018). “의약품 유형별(합성, 바이오) 개발 특성”, 「BioINwatch(BioIN+ Issue+Watch) 」: 18-23

정지연 (2018) 2016년 신약개발 정부 R&D 투자 포트폴리오 분석, 한국과학기술기획평가원.

가마타 마사히로 (2017), 처음 만나는 파이썬, 제이펍.

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

홍미영 외 (2016), 「신약개발 분야 정부/민간 R&D의 역할조정을 통한 효율화 방안 연구」, 한국과학기술기획평가원.

제이슨벨 (2016), 머신 러닝 워크북, 길벗.

김의중 (2016), 알고리즘으로 배우는 인공지능, 머신러닝, 딥러닝 입문, 위키북스.

마쓰오 유타카 (2015), 인공지능과 딥러닝 - 인공지능이 불러올 산업구조의 변화와 혁신, 동아엠엔비.

홍세호 (2013), 「국가연구개발사업 유사·중복 검색 시스템 개발을 위한 실증연구」, 한국과학기술기획평가원.

신영기 (2010), “우리나라 신약개발의 주요 현안 및 대응방안”, 「과학기술정책 」178 : 39-42.

J. Devlin, M.-W. Chang, K. Lee, K. Toutanova (2018), BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv:1810.04805.

I. Goodfellow, Y. Bengio, and A. Courville (2016), Deep Learning, The MIT Press.

J. Hirschberg, C. D. Manning (2015), Advances in natural language processing, Science, 349(6245), p. 261-266

G. James, D. Witten, T. Hastie, and R. Tibshirani (2013), An Introduction to Statistical Learning, Springer.

K. P. Murphy (2012), Machine Learning: A Probabilistic Perspective, The MIT Press.

D. A. Ferrucci (2012), "Introduction to "This is Watson"," IBM Journal of Research and Development, Vol. 56, pp. 1:1-1:15.

C. M. Bishop (2006), Pattern Recognition and Machine Learning, Springer.

M. Campbell, A. J. Hoane Jr., and F.-h. Hsu (2002), "Deep Blue," Artificial Intelligence, Vol. 134, pp. 57-83.

T. Hastie, R. Tibshirani, and J. Friedman (2001), The Elements of Statistical Learning, Springer.

R. O. Duda, P. E. Hart, and D. G. Stork (2001), Pattern Classification, John Wiley &Sons, Inc.,

T. M. Mitchell (1997), Machine Learning, McGraw-Hill.

P. E. Hart and R. O. Duda (1977), "PROSPECTOR – a computer-based consultation system for mineral exploration," Artificial Intelligence Center, SRI International, Technical Note, No. 155,

[웹사이트]

국가과학기술지식정보 (https://www.ntis.go.kr)

위키백과 (https://www.wikipedia.org)

ratsgo's blog for textmining (https://ratsgo.github.io/)

임베딩:저차원 공간으로 변환 (https://developers.google.com/machine-learning/crash-course/embeddings/translating-to-a-lower-dimensional-space?hl=ko)

쉽게 씌어진 word2vec (https://dreamgonfly.github.io/machine/learning,/natural/language/processing/2017/08/16/word2vec_explained.html)

용언 – 동사, 형용사 (https://brunch.co.kr/@adipoman/190)

배재경 (2017) 신경망 번역 모델의 진화 과정, 카카오AI리포트 (https://brunch.co.kr/@kakao-it/155)

U.S. National Library of Medicine (https://www.nlm.nih.gov/databases/download/pubmed_medline.html)

Mathworks (https://kr.mathworks.com/examples/text-analytics/mw/textanalytics-ex62579343-visualize-word-embeddings-using-text-scatter-plots)

Google AI Blog (2016) (https://ai.googleblog.com/2016/09/a-neural-network-for-machine.html)

Andrej Karpathy (2015) The Unreasonable Effectiveness of Recurrent Neural Networks (http://karpathy.github.io/2015/05/21/rnn-effectiveness/)

주 의

1. 이 보고서는 한국과학기술기획평가원에서 수행한 연구(또는 사업)보고서입니다.
2. 이 보고서 내용을 발표할 때에는 반드시 한국과학기술기획평가원에서 수행한 연구결과임을 밝혀야 합니다.
3. 국가과학기술 기밀유지에 필요한 내용은 대외적으로 발표 또는 공개하여서는 아니됩니다.

KISTEP 한국과학기술기획평가원
Korea Institute of S&T Evaluation and Planning
서울시 서초구 마방로 68 동원산업빌딩 9층~12층 (06775)
TEL : (02)589-2200 FAX : (02)589-2222
서울시 서초구 마방로 60 동원F&B빌딩 4~6층 (06775)
TEL : (02)589-2220
http://www.kistep.re.kr