인공지능(기계학습) 기반 R&D정보데이터 분석시스템 개발

2022.2.21.

예비타당성조사2센터 유거송 부연구위원

예비타당성조사3센터 김한해 연구위원

목 차

Ⅰ. 추진배경 및 필요성

연구의 필요성

- 정부R&D 예산 급증에 따른 KISTEP의 업무량, 복잡도가 증가하고 있음
  - 소부장, 코로나19 대응 등을 위한 정부R&D 예산 급증 → 세부사업의 다변화, 개수 증가에 따른 관리 업무량 증가

(단위: 조원, %)

연 도	’17	’18	’19	’20	’21	’22(안)
R&D 예산(조원)	19.5	19.7	20.5	24.2	27.4	29.8
R&D예산 증가율(%)	1.9	1.1	4.4	18.0	13.1	8.8

- R&D 트렌드 변화, 글로벌 이슈가 지속 발생함에 따라 수시적으로 신속한 분석 수행이 필요
  - 다양한 분야에 대한 신규 R&D 투자영역 발굴, 기존 투자 효율화 등
    ※ (예시) 코로나19 대유행으로 인한 감염병 관련 R&D 현황 분석
  - 소수의 담당 직원이 모든 이슈에 효과적으로 대응하기에는 한계점 존재
- **의사결정 지원을 위한 시스템 구축 필요성 제기**
  - **분석 자동화· 효율화 & 주관적 판단 최소화**

인공지능 언어모델의 발전

- 인공지능 기술의 발전으로 언어모델 등 자연어처리 기술이 실용화
  - (언어모델) 영어, 한국어와 같은 일반적인 자연 언어의 단어 배열(sequence)의 확률을 계산하는 수학적 모델
    - ※ 문장(=단어의 나열)이 입력값으로 주어졌을 때, 그 다음에 나올 단어가 무엇인지 예측
    - ※ (예시)

> 정부 중앙방역대책본부(방대본)가 어제(25일) 국내 신규 코로나 확진자가 583명(국내 발생 553명)이라고 밝혔다. 지난 2~3월 대구 유행이 한창이던 3월 2일 600명 확진 이후 268일만에 하루 사이 가장 많은 …

가장 많은… 확진자? 도시락? 발표자? 자동차?

최근 기사들이 언어모델에 잘 학습되어 있다면, "확진자"의 확률이 가장 높게 나올 것

  - (텍스트임베딩) 자연언어를 학습하여 비정형 데이터인 단어들을 숫자로 된 벡터로 치환시켜 다양한 어플리케이션*에 활용할 수 있게 하는 기술
    - ※ 해당 단어가 훈련 데이터 내에서 가지는 맥락 정보를 담도록 학습하여, 서로 연관성이 높은 단어들은 그들의 벡터도 공간 상에서 비슷한 위치를 가짐
    - * 검색, 챗봇, 요약봇, 스팸분류, etc.

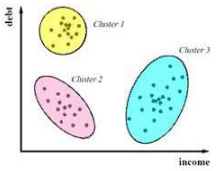
기계학습(인공지능) 도입의 이점

- 기존의 정책연구 방법론(전문가 자문 중심)은 전문가풀 유지관리와 균일한 결과를 얻는 것이 어려움
- 기계학습(인공지능)은 이를 보완하여 업무의 효율화 및 신뢰성 제고가 가능
  - 데이터 기반, 자동화로 신속하고 일관된 결과

기계학습 (Machine Learning)

– Unsupervised Learning (비지도학습, 간단예시: Clustering, 집단)

○ 입력용 데이터만 제공하고 라벨링 없이 데이터에 내재하는 구조 파악

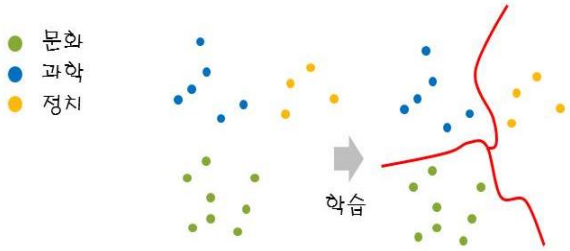


사업, 과제간 관계성 분석,

토픽 클러스터링 분석 등

– Supervised Learning (지도학습, 간단예시: Classification, 분류)

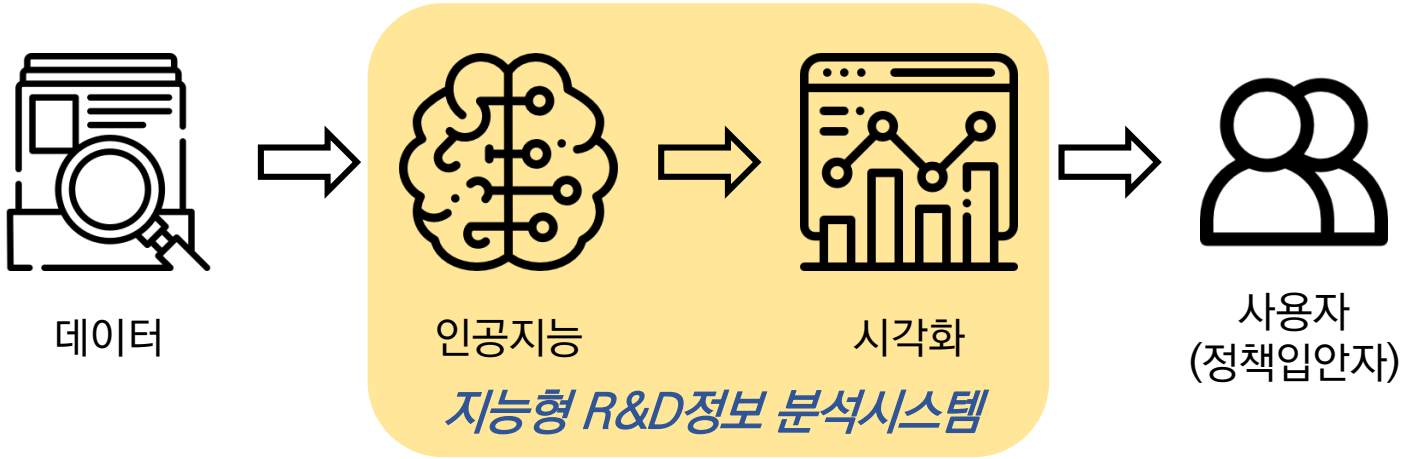
○ 입력과 올바른 출력(분류결과)이 세트가 된 훈련데이터(골드스탠다드)를 미리 준비하여 학습시키고 어떤 입력이 주어졌을 때 올바른 출력(분류)이 가능토록 함



과제 자동분류 등

연구개발 정보 빅데이터 + 인공지능

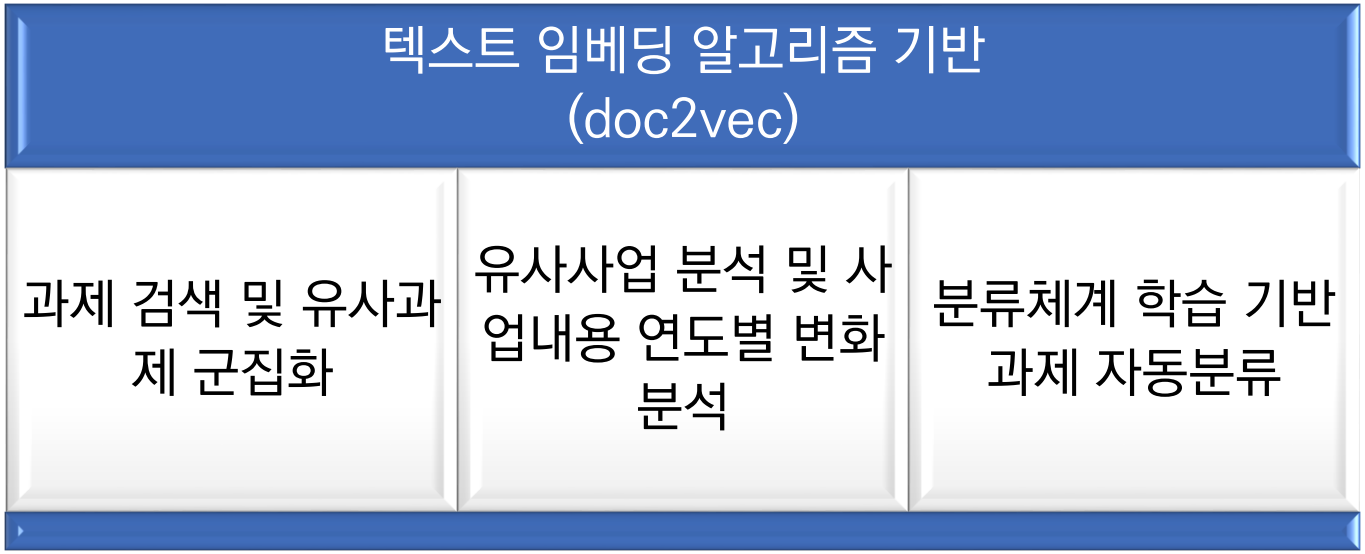
- 그간 여러 기관에서 연구개발과제, 논문 등 방대한 연구개발 정보 데이터를 축적해왔으며, 자연어처리 기술은 이러한 데이터에 충분히 적용이 가능함
  - (조사분석) NTIS의 우리나라 정부R&D 과제정보
    ※ 과제명, 사업명, 연구비, 연구목표, 연구내용, 과학기술표준분류 등
  - (논문DB) PubMed, Web of Science, Scopus 등 논문 DB
    ※ 초록, 저자, 국가 등
  - (해외과제) NIH, NSF 등 해외 부처에서 발주한 연구과제 정보



II. 지능형 연구개발정보데이터 분석시스템 소개

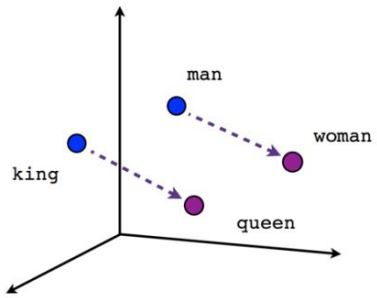
지능형 R&D정보 분석시스템

- R&D사업 관리(예산, 평가, 기획)를 위한 인사이트를 제공하는 SW
  - 연구과제들의 과학계량학적 분석을 자동화
  - 사업 단위의 분석 결과 제공
- 그래픽 인터페이스(손쉽게 사용) 기반으로 개발
  - 원내 직원 공동사용을 위한 온라인화 개발 완료, 서비스는 미개시

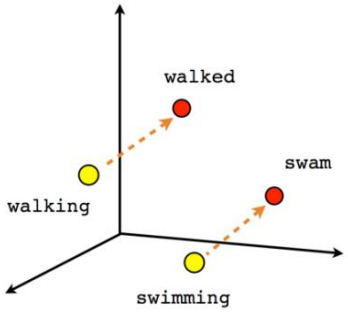


Word2vec 알고리즘

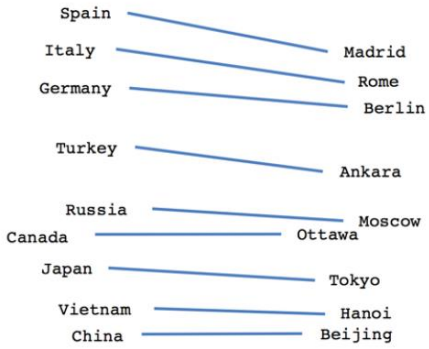
- 구글에서 개발한(2014) 텍스트 임베딩 알고리즘
  - 인공신경망(neural network)을 자연어처리에 활용
  - 텍스트 임베딩(text embedding) : 단어나 문서를 숫자로 된 벡터로 수치화
  - 비정형 데이터를 숫자로 바꿈으로써 다양한 전산 처리가 가능해짐
- Word2vec은 아래 그림과 같이 단어의 의미적 관계성이 해당 벡터 간(공간상)의 관계성에도 드러나도록 임베딩함
  - 단어의 벡터값 계산으로 관계성 유추 가능



Male-Female



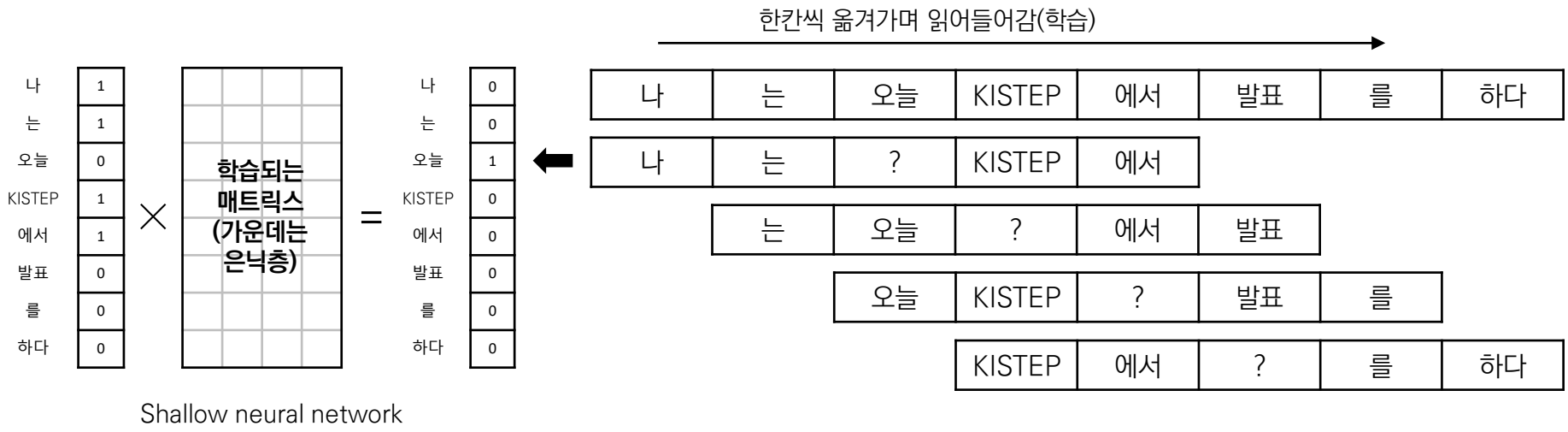
Verb tense



Country-Capital

Word2vec 알고리즘

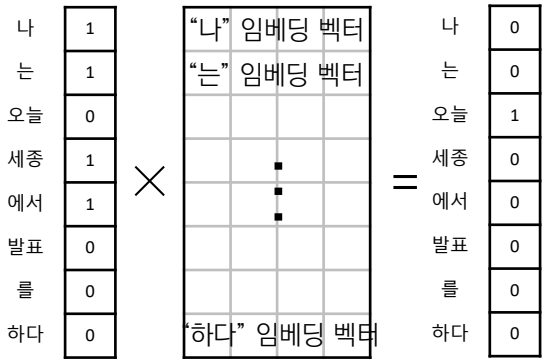
- 인공신경망 훈련 : input과 output 사이의 행렬을 튜닝하는 것
  - 뇌의 뉴런들 사이에서의 신호전달은 수학적으로 벡터와 행렬의 곱으로 표현됨
  - 행렬의 각 원소들은 각 신호들의 가중치에 해당하며, 학습 과정에서 조정됨
- Word2vec은 훈련 데이터(=글)를 읽어들이면서, 각 단어 주변의 단어를 보고 그 단어가 무엇인지 맞히도록 인공신경망을 학습

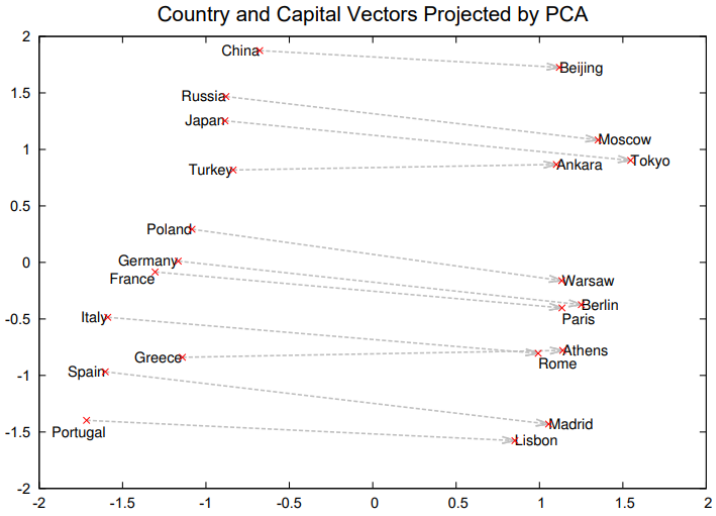


Shallow neural network

Word2vec 알고리즘

- 인공신경망의 예측치와 실제 답(맞춰야 하는 단어) 간의 차이가 최소화 되도록 가중치를 조정
  - 이를 학습 데이터 전체에 대해 반복
- 결과로, 은닉층 행렬의 각 행들이 단어들의 임베딩 벡터가 됨



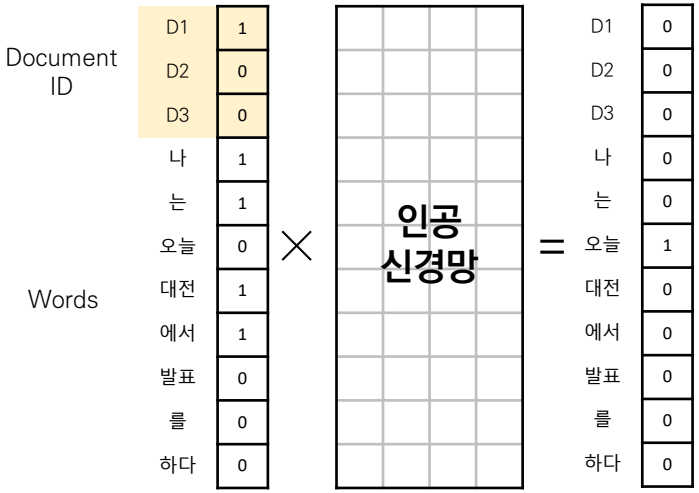


Mikolov et al. (2013) Advances in neural information processing systems, pp. 3111-3119

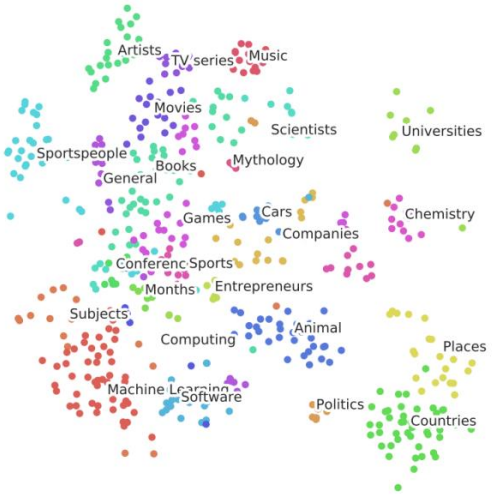
(Word2vec 오리지널 논문)

Doc2vec 알고리즘

- 동 분석시스템의 기반이 되는 알고리즘
- Word2vec에서 개량된 알고리즘으로, 단어뿐만 아니라 문서도 숫자 벡터로 임베딩
  - 학습 결과로 각 단어 및 문서의 임베딩 벡터를 얻을 수 있고, 서로 연관성이 높은 단어&문서의 벡터 값이 유사하게 됨



Shallow neural network

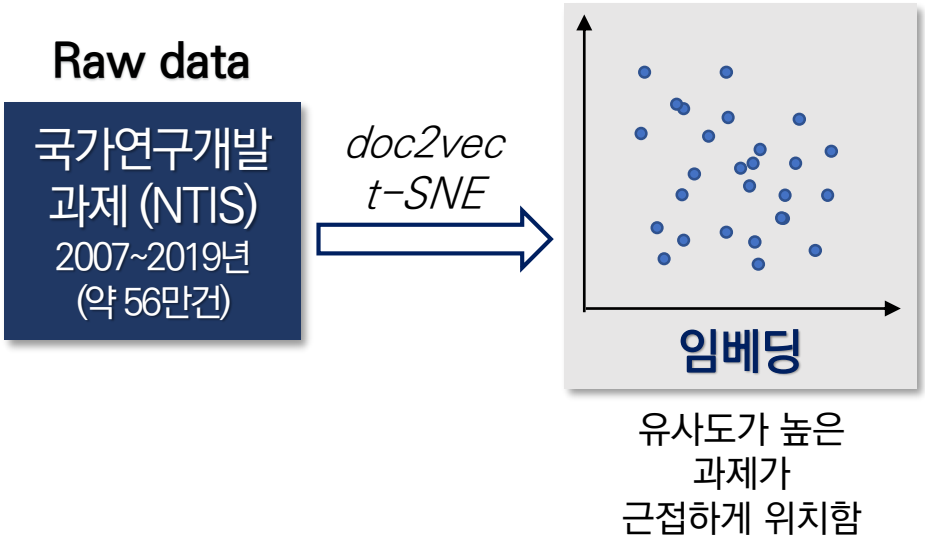


Fisher et al. (2017), "Exploring Optimizations to Paragraph Vectors"
(doc2vec representation of Wikipedia articles)

시스템 동작 방식

- NTIS 과제를 doc2vec 으로 임베딩하여 각 과제를 숫자 벡터로 대응시킴
  - 각 과제의 과제명, 연구목적, 연구내용, 국문키워드, 영문키워드를 하나의 긴 텍스트로 이어서 그 과제를 대변하는 텍스트로 활용

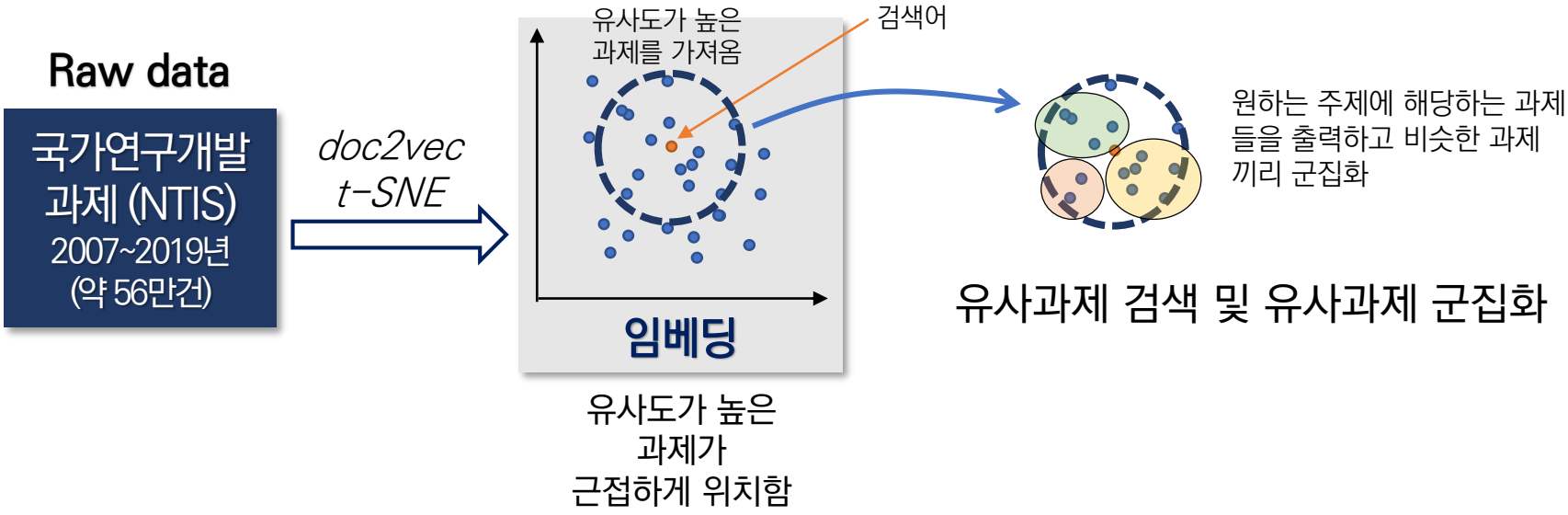
※ doc2vec 임베딩 전 형태소 분석 실행



시스템 동작 방식

- NTIS 과제를 doc2vec 으로 임베딩하여 각 과제를 숫자 벡터로 대응시킴
  - 과제명, 연구목적, 연구내용, 국문키워드, 영문키워드를 하나의 긴 텍스트로 이어서 그 과제를 대변하는 텍스트로 활용

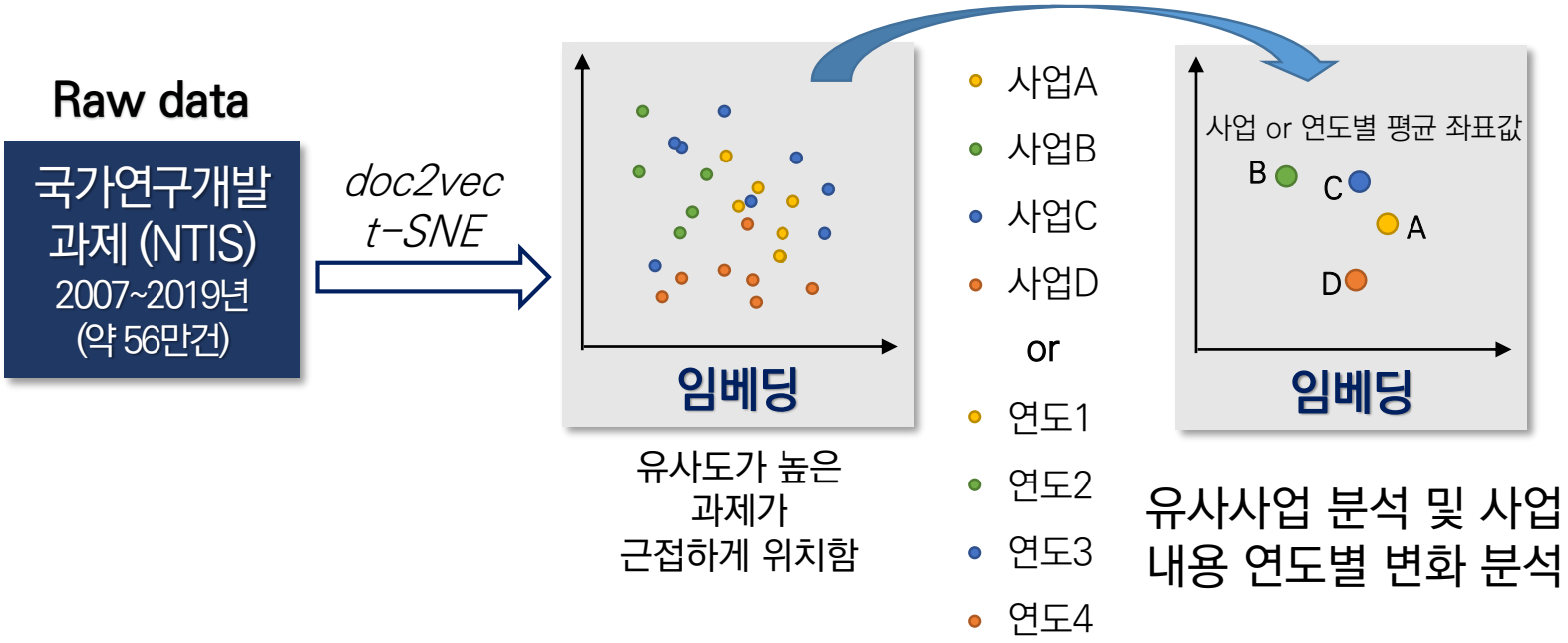
※ doc2vec 임베딩 전 형태소 분석 실행



시스템 동작 방식

- NTIS 과제를 doc2vec 으로 임베딩하여 각 과제를 숫자 벡터로 대응시킴
  - 과제명, 연구목적, 연구내용, 국문키워드, 영문키워드를 하나의 긴 텍스트로 이어서 그 과제를 대변하는 텍스트로 활용

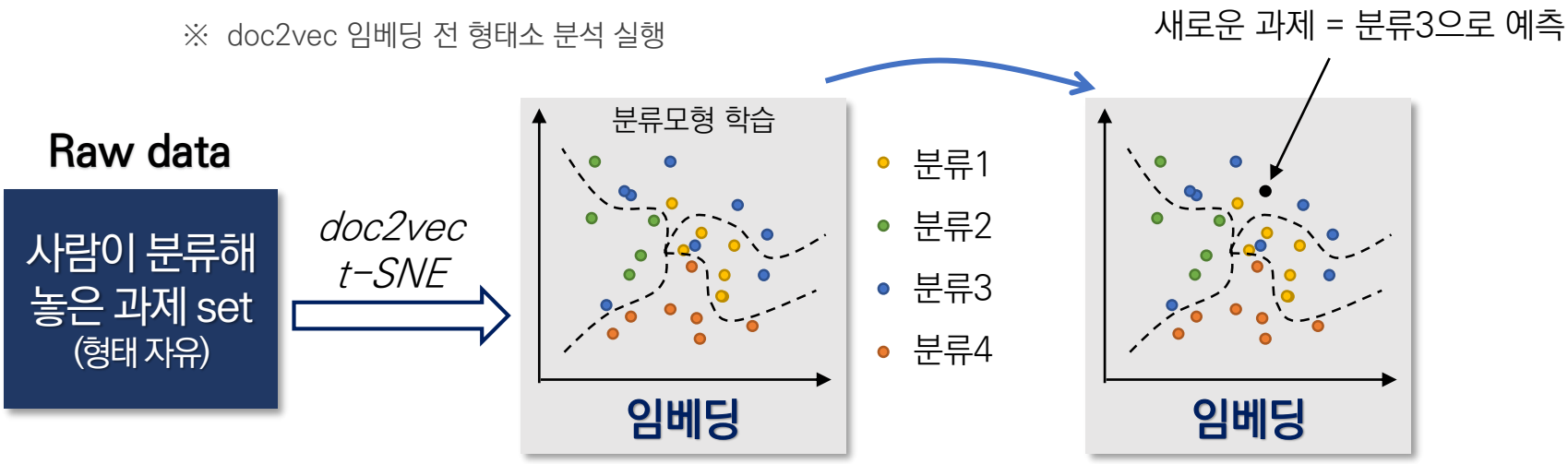
※ doc2vec 임베딩 전 형태소 분석 실행



시스템 동작 방식

- NTIS 과제를 doc2vec 으로 임베딩하여 각 과제를 숫자 벡터로 대응시킴
  - 과제명, 연구목적, 연구내용, 국문키워드, 영문키워드를 하나의 긴 텍스트로 이어서 그 과제를 대변하는 텍스트로 활용

※ doc2vec 임베딩 전 형태소 분석 실행



분류체계 학습 기반 과제 자동분류

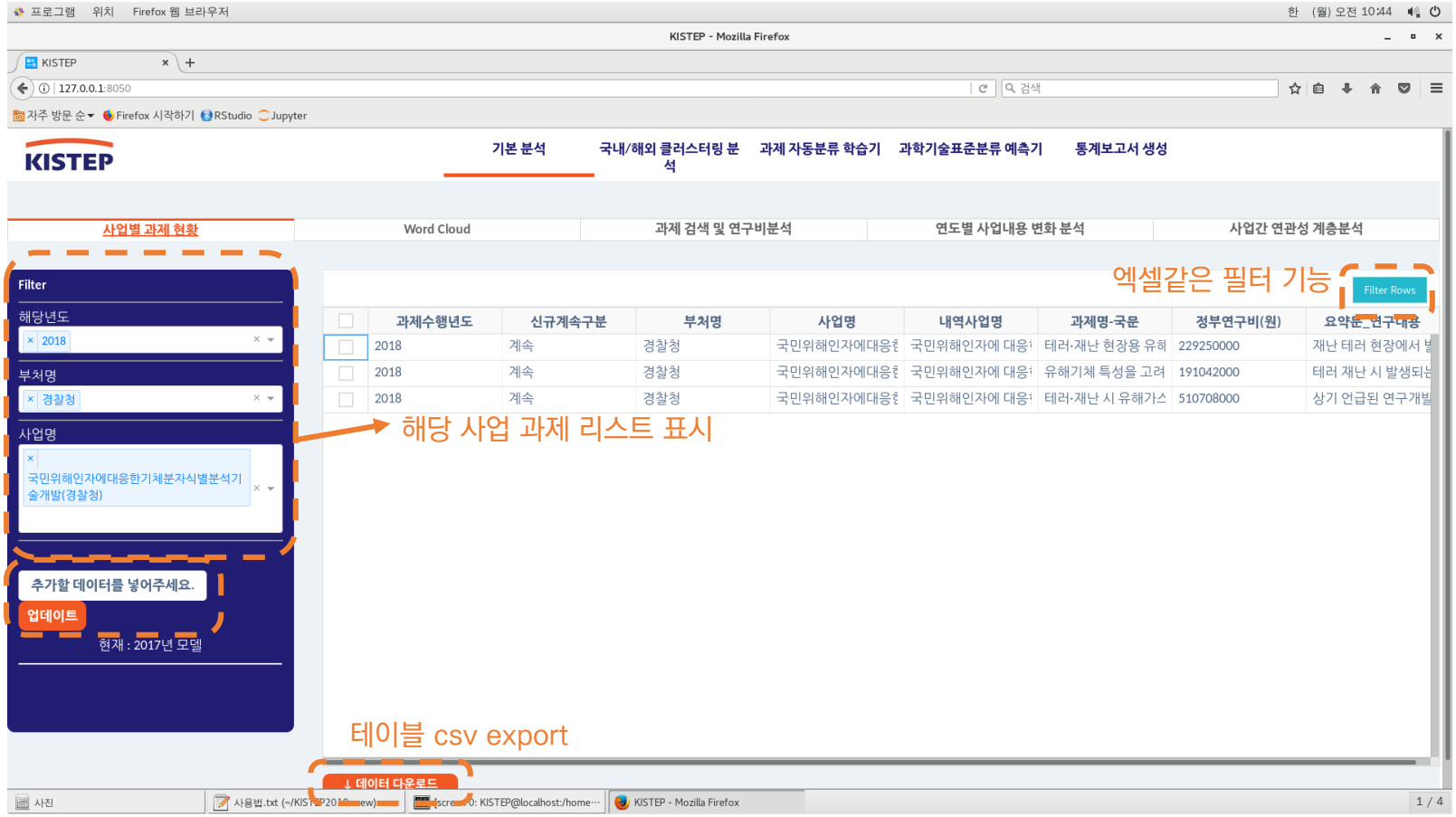
시스템 개발 spec.

- 개발 언어 및 구동 환경 : Python 3.6.3 / Linux(CentOS 7) PC
- 기본 동작 구조
  - Raw data(DB: sqlite3)는 미리 doc2vec, word2vec으로 임베딩(300차원)
  - 기능 호출 시 필요한 벡터들을 fetch해서 분석하고, 결과를 표 또는 그래프로 시각화하여 출력
    - ※ 사용된 알고리즘 : t-SNE, k-means, LDA, MLP, SVM, GBM, GLM, Random forest 등
    - ※ 사용된 파이썬 라이브러리 : pandas, flask, numpy, pickle, plotly, dash 등
  - 과제 자동 분류기는 원하는 데이터(csv파일)를 업로드해서 학습, 한번 학습한 데이터는 나중에 불러오기 가능
- GUI 구현 방식
  - Flask 라이브러리를 이용해서 가상 웹서버 실행 → 웹브라우저로 로컬IP 접속해서 사용
  - 파이썬으로 html 코딩, 테이블 및 그래프는 각각 dash, plotly 라이브러리(오픈소스) 이용

III. 각 기능별 소개 및 활용 예시

① 각 기능별 소개

1.1. 사업별 과제 현황



연도, 부처, 사업명 선택

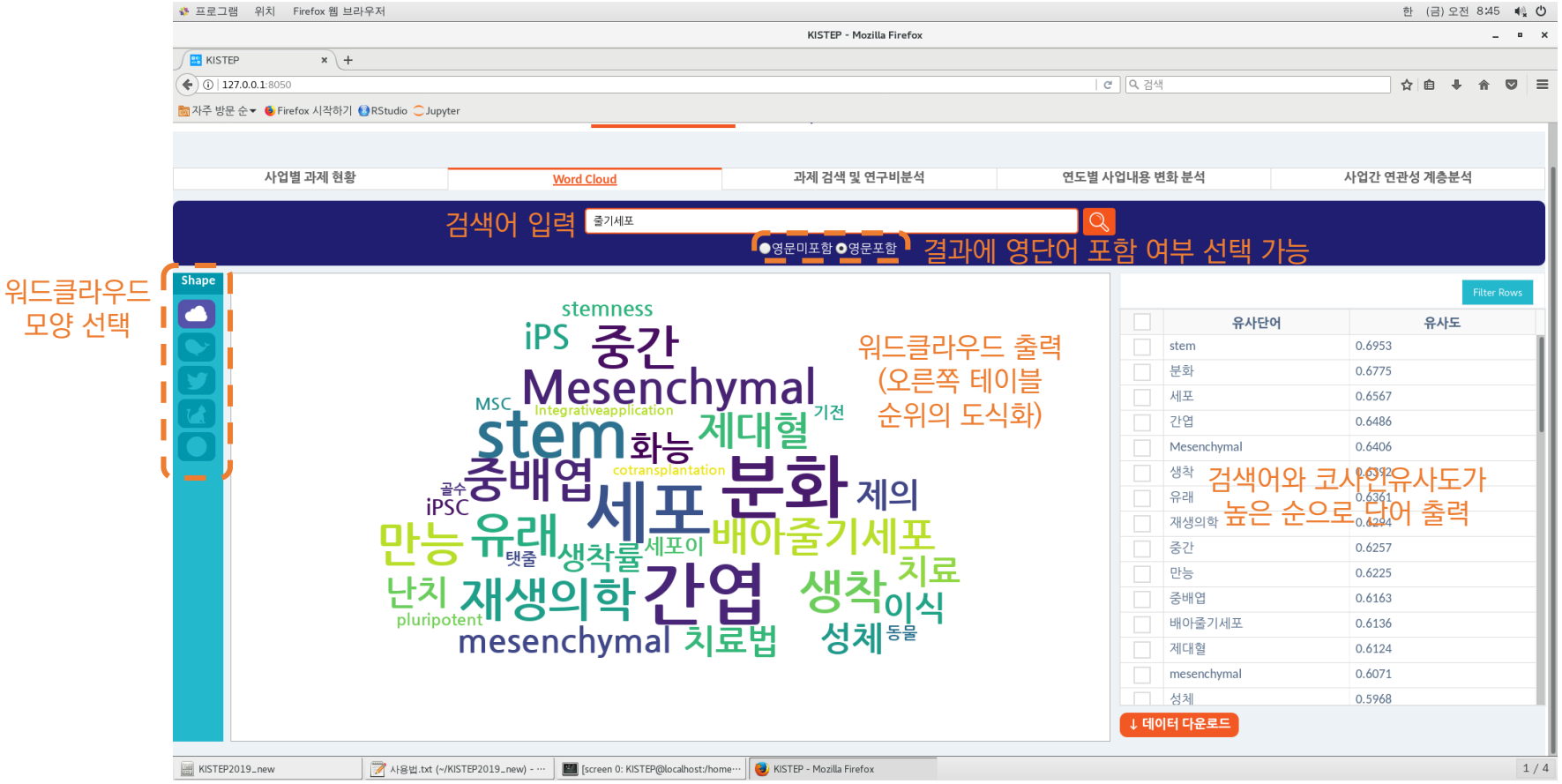
조사분석 데이터 추가 (doc2vec 학습) 버튼

해당 사업 과제 리스트 표시

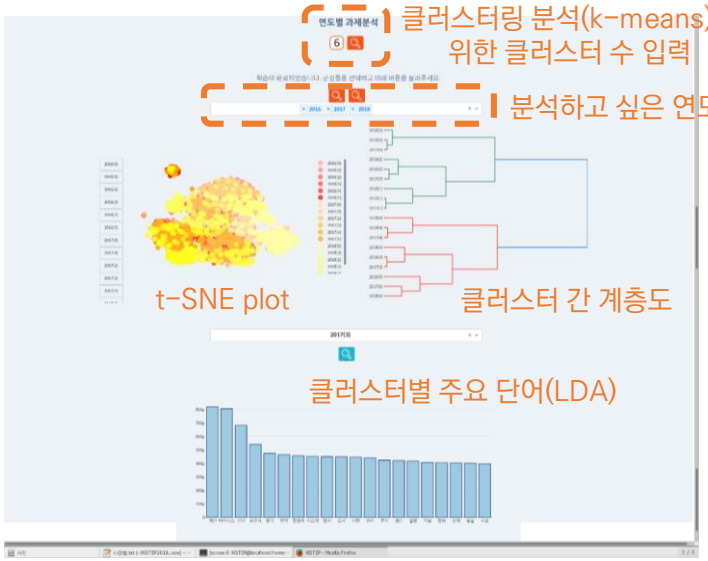
엑셀같은 필터 기능

테이블 csv export

1.2. Word Cloud



1.3. 연관과제 현황

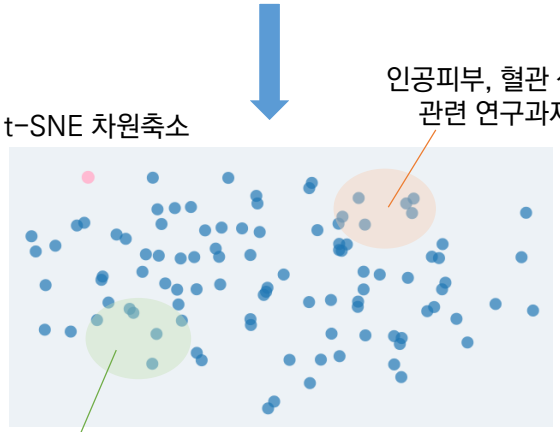


- 코사인유사도 기반의 유사(연관)과제 검색 기능
  - 검색어와의 유사도 순으로 과제를 정렬
- 검색 결과의 연구비 통계 제공
- 검색 결과의 클러스터링 분석 기능
  - 총 클러스터 수 : 정의한 클러스터 수 X 연도 수

※ 위의 케이스의 경우 6 X 3 = 18개 클러스터 출력

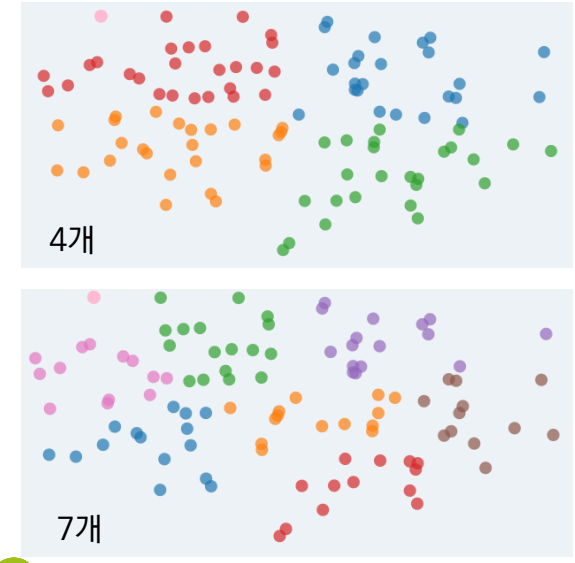
클러스터링 분석 방법론

예시 :
(검색어) 동물의 체내에서 사람의 장기를 생산하는 연구
(2017년 과제 대상)
유사도가 가장 큰 100개 과제 출력(출력 과제 수 지정)



k-means
클러스터링
클러스터 수 지정

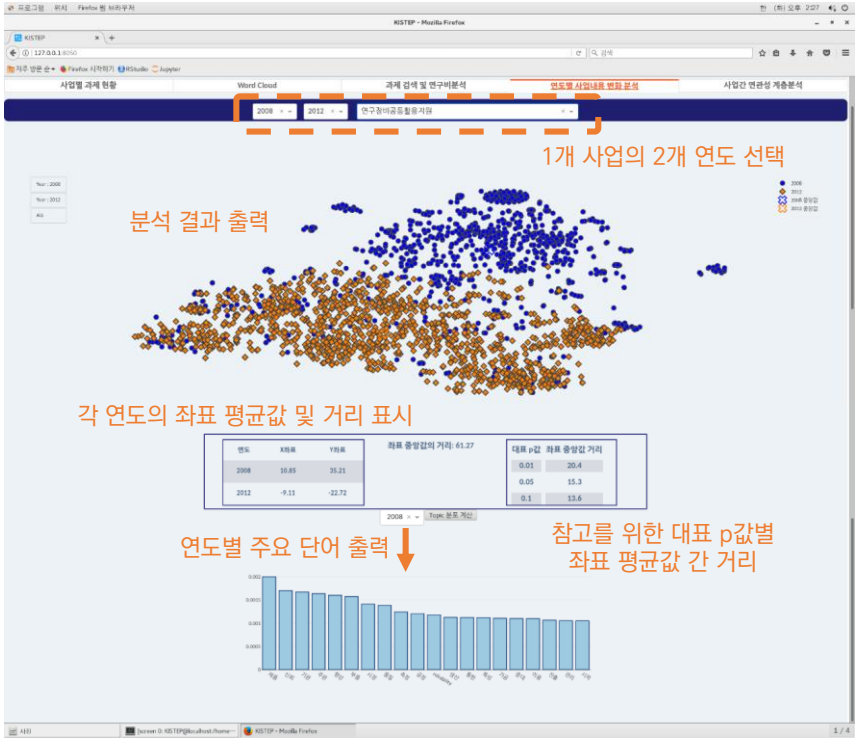
k-means 클러스터링 :
근처에 있는 점들을 묶어 같은 색으로 표시



클러스터별 특성이 나타나는지 확인
→ 최적 클러스터 수 결정

동일한 분석을 PubMed 논문에도 수행 가능
→ 국내/해외 동향 비교 가능

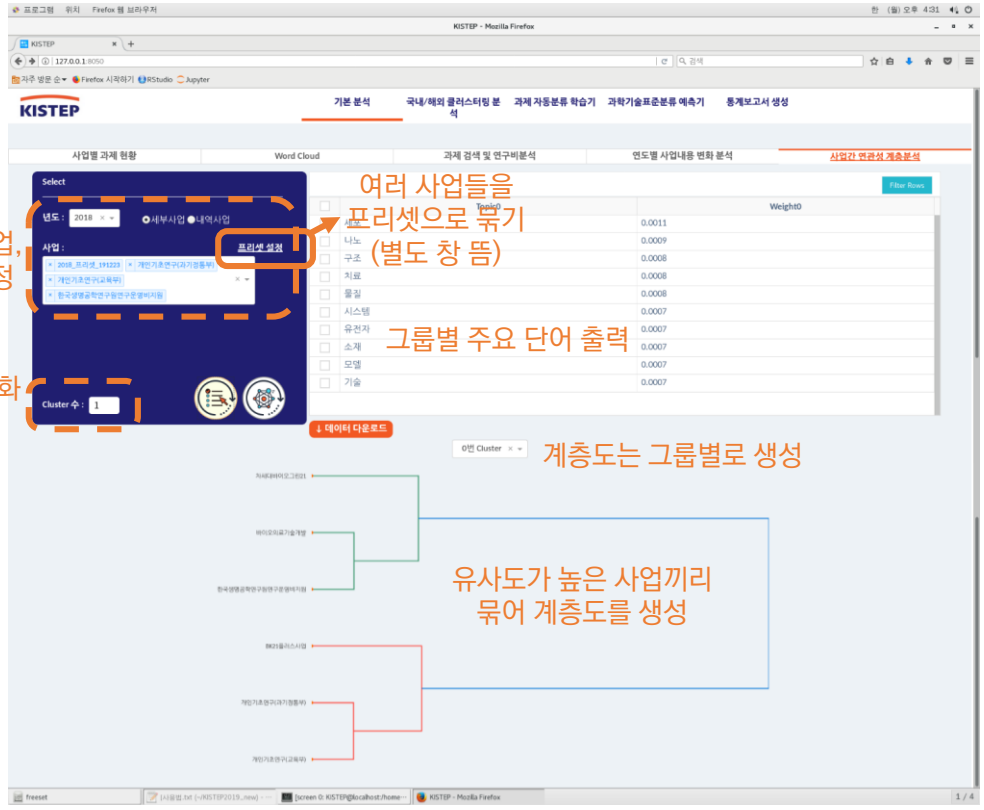
1.4. 연도별 사업내용 변화 분석



- 특정 사업이 2개 연도 사이에 얼마나 변화하였는지 분석
  - 해당 doc2vec 벡터 fetch
  - t-SNE로 차원 축소
  - 연도별 좌표 표시하고 평균값(중앙값) 간 거리 계산
- 중심극한정리를 이용해서 유저가 얻은 평균값 간 거리의 통계적 유의미성을 제공
  - 사업 & 연도 조합 ranking 매김
  - 중심극한정리로 p값 계산
  - 대표 p값(0.01, 0.05, 0.1)별 평균값 거리 제공

    ※ 상대적으로 얼마나 연도 간의 변화가 컸는지 레퍼런스로 활용
- 연도별 주요 단어를 출력하여 내용 파악
  - 벡터 plot을 드래그하여 특정 과제들의 주요 단어만 출력도 가능
  - LDA(k=1) 결과 단어 weight 순으로 나열

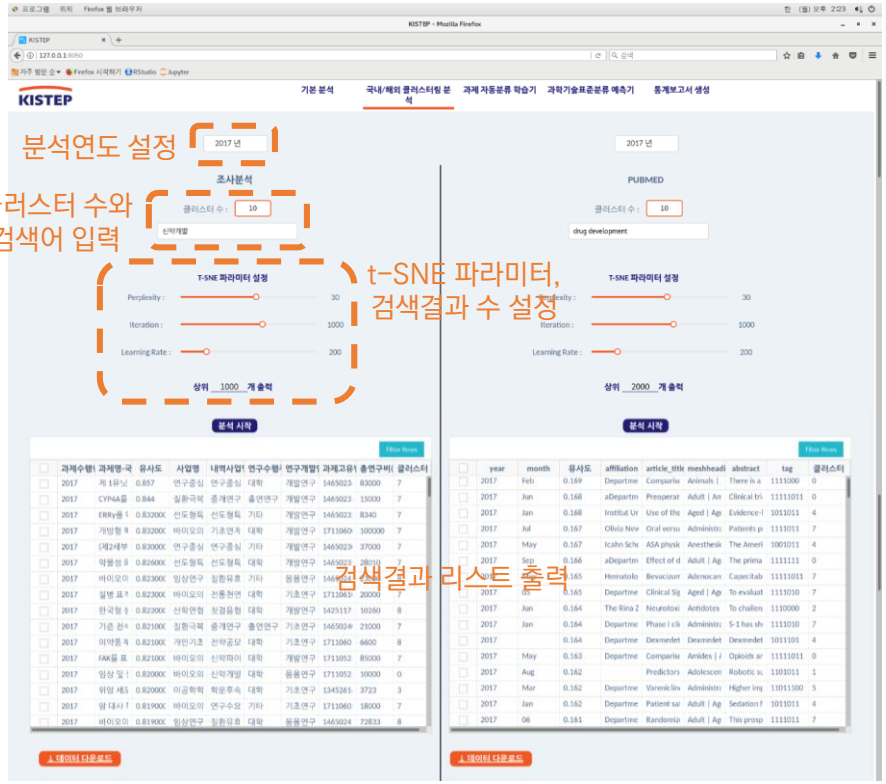
1.5. 사업간 연관성 계층분석



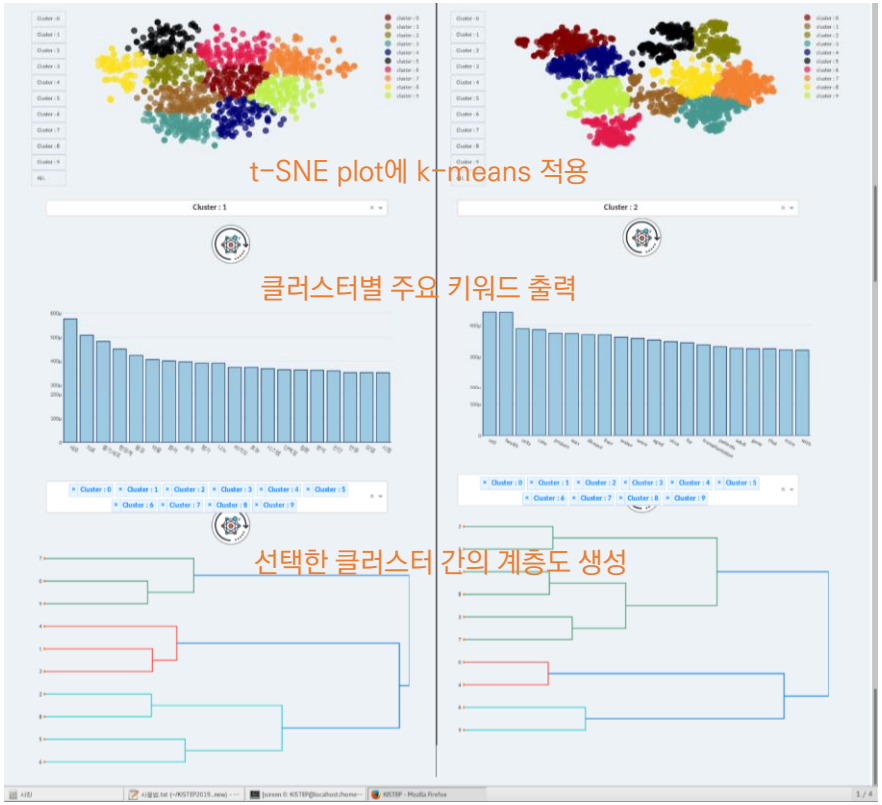
- 여러 사업 간의 상호 유사성을 분석하고 비슷한 사업들의 군을 도출
  - 세부사업/내역사업 단위 분석 가능
  - 사업들에 포함된 세부과제들의 벡터 평균값을 사업을 대표하는 벡터 값으로 간주
  - 이 벡터값들의 코사인유사도를 기준으로 유사한 사업들을 묶음

    ※ 계층도(dendrogram) 생성
- 사업 그룹(클러스터) 수를 지정하여 덩어리를 미리 나눌 수 있음
  - 클러스터별 주요 단어(LDA 이용) 출력
  - 클러스터별 계층도 생성
- 자주 input 값으로 사용하는 사업들을 프리셋으로 지정할 수 있음
  - 프리셋과 다른 개별 사업들을 섞어 분석할 수 있음

2. 국내/해외 클러스터링 분석



조사분석(국내 과제)과 PubMed(해외 논문)를 병렬적으로 분석



3. 과제 자동분류 학습기

특정 기술분야가 A, B, C 세 분류로 이루어져있다고 가정하고,
기존에 세부과제들을 A or B or C 로 분류해놓은 데이터가 있을 시,
이를 학습하여 새로운 과제가 A, B, C 중 어떤 분류인지 예측

기존에 분류된 데이터
(예: 13~17년 데이터)

과제명	분류
과제1	A
과제2	B
과제3	A
과제4	C
⋮	⋮
과제n	B

학습항목(feature) 선택:
과제명, 연구목표, 연구내용 등

5개 머신러닝 알고리즘
(MLP, GBM, RF, SVM, GLM)

임의의 과제를
A, B, C 중 하나로
분류해주는 모델

새로운 과제 데이터
(예: 18년 데이터)
(과제a, 과제b, ⋯, 과제z)

예측

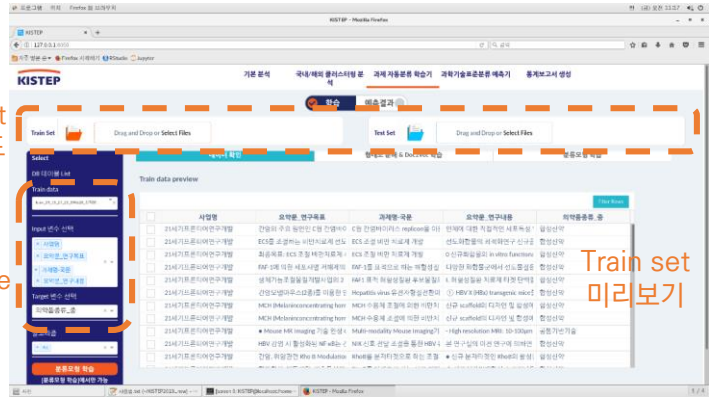
과제명	분류 (예측)
과제a	C
과제b	B
과제c	C
⋮	⋮
과제z	A

매년 추가되는 조사분석 데이터에 대해 분류를 실시할 시 전문가 판단의 참고자료로 활용 가능

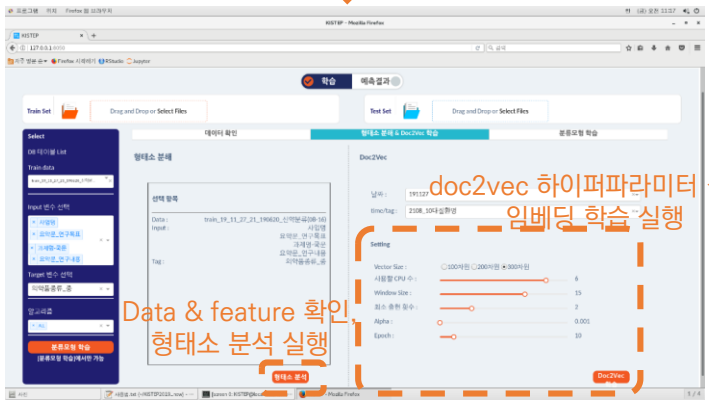
3. 과제 자동분류 학습기

Train/test set 업로드

Train set 불러오고 학습 feature 선택



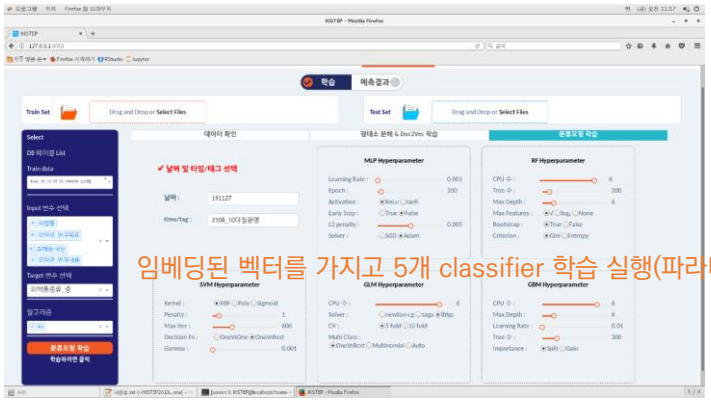
Train set 미리보기

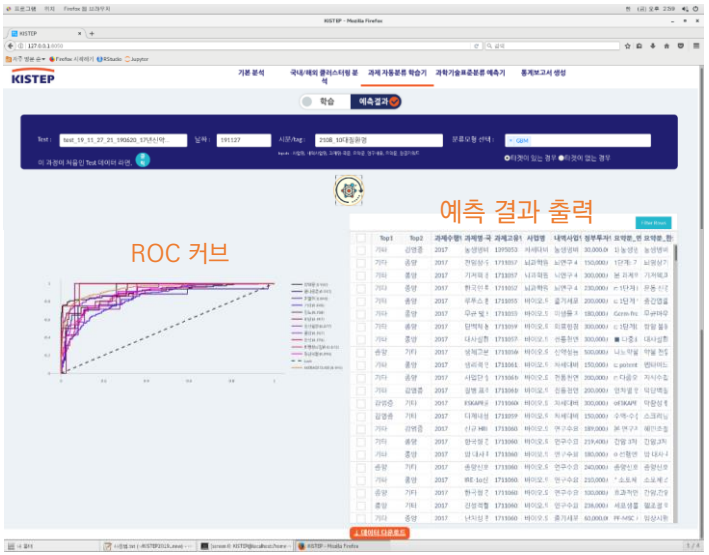


Data & feature 확인, 형태소 분석 실행

doc2vec 하이퍼파라미터 설정, 임베딩 학습 실행

임베딩된 벡터를 가지고 5개 classifier 학습 실행(파라미터 조정 가능)





ROC 커브

예측 결과 출력

III. 각 기능별 소개 및 활용 예시

② 국내 전반적 투자현황 분석

활용 시나리오

- 어떤 담당자가 유전체 분야의 투자현황을 알고 싶다고 가정
  - 유전체에 대해 대략적으로는 알지만(바이오 전공자), 전문가는 아님
    ※ 주요 기술적 용어들의 뜻은 알고 있음
- 알고 싶은 내용 :
  - 전체적인 사업 현황, 세부 분야
  - 연도별 투자 비중 변화
    ☞ **투자방향 수립, 신규사업 검토 등 업무 수행 시 기본 배경지식**
    ☞ **전문가 의견을 효과적으로 이해 가능**

- 유전체 분야를 효과적으로 이해하기 위해 분석 시스템을 활용
  - 관련 사업(예산요구서), 과제, 기술 문서 등을 일일이 읽어 학습하기 전, 전체적 구도를 파악
- 주요 사업 및 관련 과제를 검색하고, 세부분야에 대한 군집 분석을 수행

주요 키워드 및 사업 파악

- “유전체” 관련 주요 키워드(워드클라우드)
  - “유전체”와 유사도 순으로 나열
    ※ 염기서열을 시퀀싱(해독)하고, 무슨 형질이 발현되는지 연구하고, 전사체, 단백체, 대사체 등도 포괄되며, 생물(생명)정보학, 의학, 육종 등에 활용
- “유전체” 관련 과제검색을 통한 주요 사업 도출
  - 검색결과의 내역사업 단위 예산액 제공
  - 유전체 관련 주요 사업 도출

포스트게놈신산업육성을위한다부처유전체사업
바이오의료기술개발
차세대바이오그린21
한국생명공학연구원연구운영비지원
질환극복기술개발
암연구소및국가암관리사업본부연구운영비지원
극지연구소연구운영비지원
첨단의료기술개발
한국식품연구원연구운영비지원
산림과학연구
GoldenSeed프로젝트
뇌과학원천기술개발
이공학학술연구기반구축
생물자원발굴및분류연구
형질분석연구
…

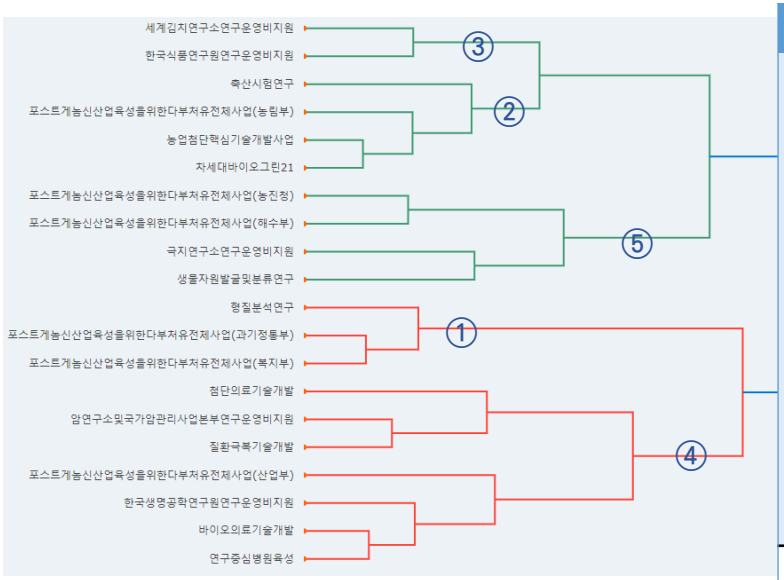
“유전체” 워드클라우드

유사사업 그룹화 분석

- 유전체 관련 주요 사업들을 그룹화(계층분석)
  - 비슷한 사업들끼리 그룹화된 계층도 생성
  - 그룹별 주요 키워드로 사업군별 내용을 파악

※ 과제 클러스터링 결과 및 예산요구서 참고 등을 통해 입체적인 이해 필요

유전체 주요 사업 계층도('18년)



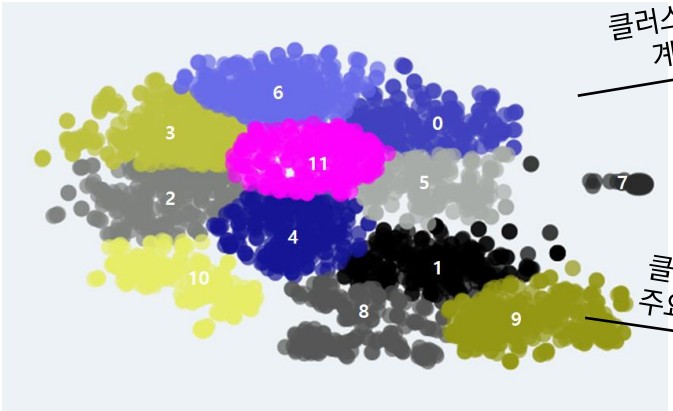
그룹별 주요 키워드

그룹1	그룹2	그룹3	그룹4	그룹5
유전체	돼지	김치	줄기세포	유전체
치료	한우	미생물	세포	해양
폐암	유전자	유전체	진단	수산
유전자	사료	프로바이오틱스	바이오	정보
데이터	작물	바이옴	물질	유전자
위암	평가	식품	유전자	생물
세포	형질전환	유용	면역	육종
유방암	정액	마이크로	치료	저항
종양	질병	전통	질환	식물
전이	선발	장내	분화	작물
질병치료	**육종**	**마이크로바이옴**	**유전자치료**	**생물유전자원**

과제 단위 주제별 군집화 분석

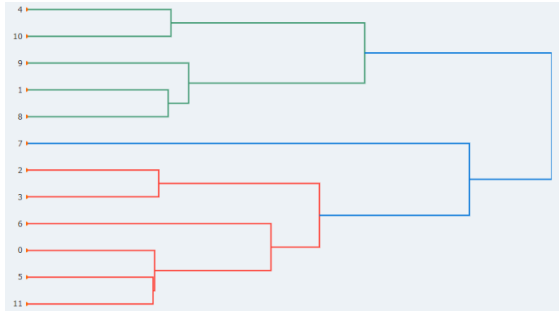
- “유전체” 관련 과제들의 과제 단위 군집화(클러스터링)
  - 유사한 과제들을 그룹화 하고 그룹별 주요 키워드를 출력
  - 최적 검색결과(과제) 수 및 클러스터 수 설정 → 2,700개/12개
    - ※ 유사도가 낮아짐에 따라 유전체와 관련성이 낮은 과제 등장 → NTIS 검색결과 수 정도로 잡는 것을 권장
    - ※ 최적 클러스터 수 : 분류결과를 보고 trial and error를 통해 적정 개수 찾기
  - 클러스터의 계층도와 클러스터별 주요 키워드를 기준으로 그룹화

클러스터링 결과('17년 유전체 관련과제)

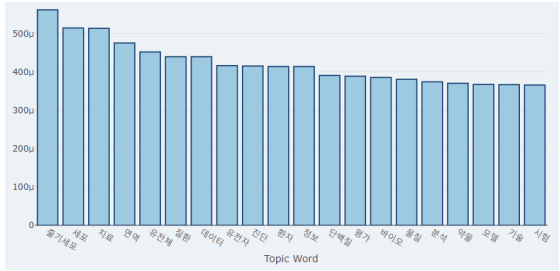


※ 각 점은 과제들이며, 클러스터 번호가 부여됨

클러스터들의 계층도



클러스터별 주요 키워드



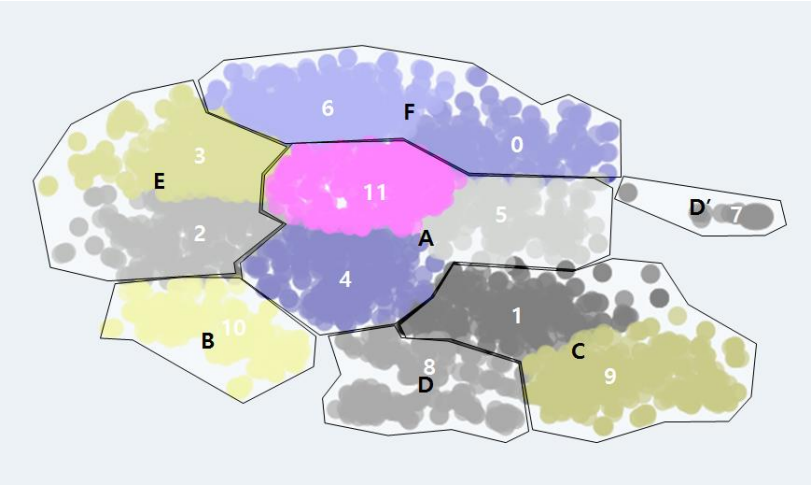
과제 단위 주제별 군집화 분석

- 위치가 인접한 클러스터는 내용상 차이가 적음 → 병합
  - 클러스터들의 계층도 상 위치와 주요 키워드 분포를 보고 주제를 판단

※ 키워드에서 주제가 잘 드러나지 않는 경우 클러스터에 포함된 과제들의 과제명 등을 참고

최종 군집화 결과('17년)

그룹	클러스터	주제
A	4, 5, 11	유전자·세포치료제 기초기전연구
B	10	마이크로바이옴 등
C	1, 9	동식물 육종
D	8	생물자원 유전정보 수집·관리
D'	7	해양생명자원
E	2, 3	유전자 기반 치료법 및 진단법 개발
F	0, 6	유전체 기반 치료기술

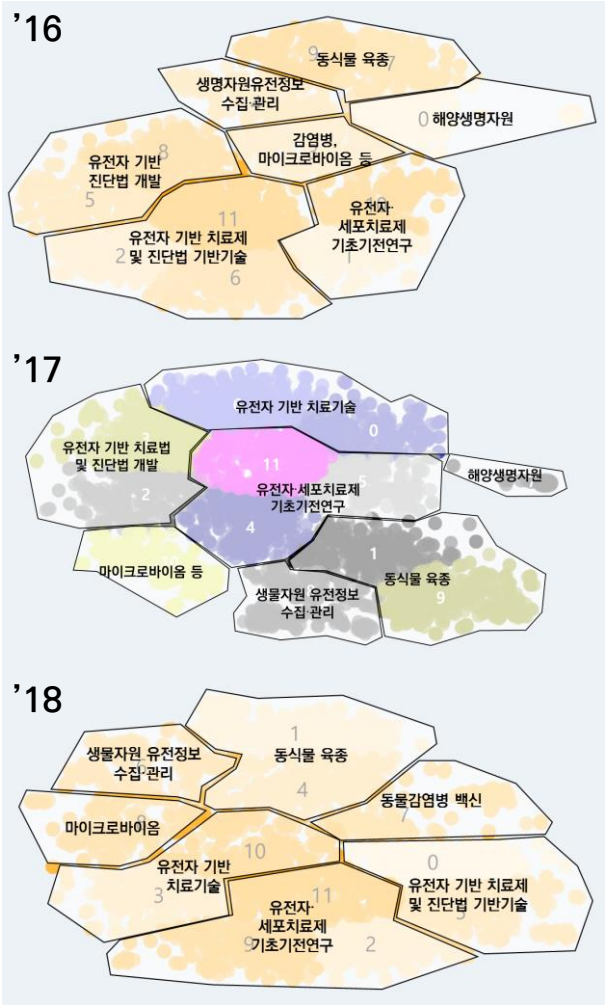


A, E, F : 레드 바이오
B, C : 그린 바이오
D, D' : 생명자원·정보

과제 단위 주제별 군집화 분석

- ’16~’18년 3개 년도에 대해 군집화 분석 수행
  - 그룹별 과제 수 및 연구비 비중 산출

그룹명	과제 수 비중			연구비 비중		
	’16	’17	’18	’16	’17	’18
유전자·세포치료제 기초기전연구	14.9%	23.9%	25.9%	15.6%	28.2%	28.4%
유전자 기반 진단법 개발	17.4%			23.8%		
유전자 기반 치료기술		18.6%	15.7%		17.1%	17.6%
유전자 기반 치료법 및 진단법 개발		19.9%			21.4%	
유전자 기반 치료제 및 진단법 기반기술	29.6%		17.7%	31.3%		18.2%
동식물 육종	13.8%	21.4%	20.4%	9.3%	16.0%	14.6%
생물자원 유전정보 수집·관리	7.8%	7.7%	7.6%	7.9%	6.6%	8.6%
동물감염병 백신			5.2%			4.2%
감염병, 마이크로바이옴 등	8.8%			7.4%		
마이크로바이옴 등		7.2%	7.5%		9.3%	8.4%
해양생명자원	7.7%	1.3%		4.7%	1.4%	



과제 단위 주제별 군집화 분석

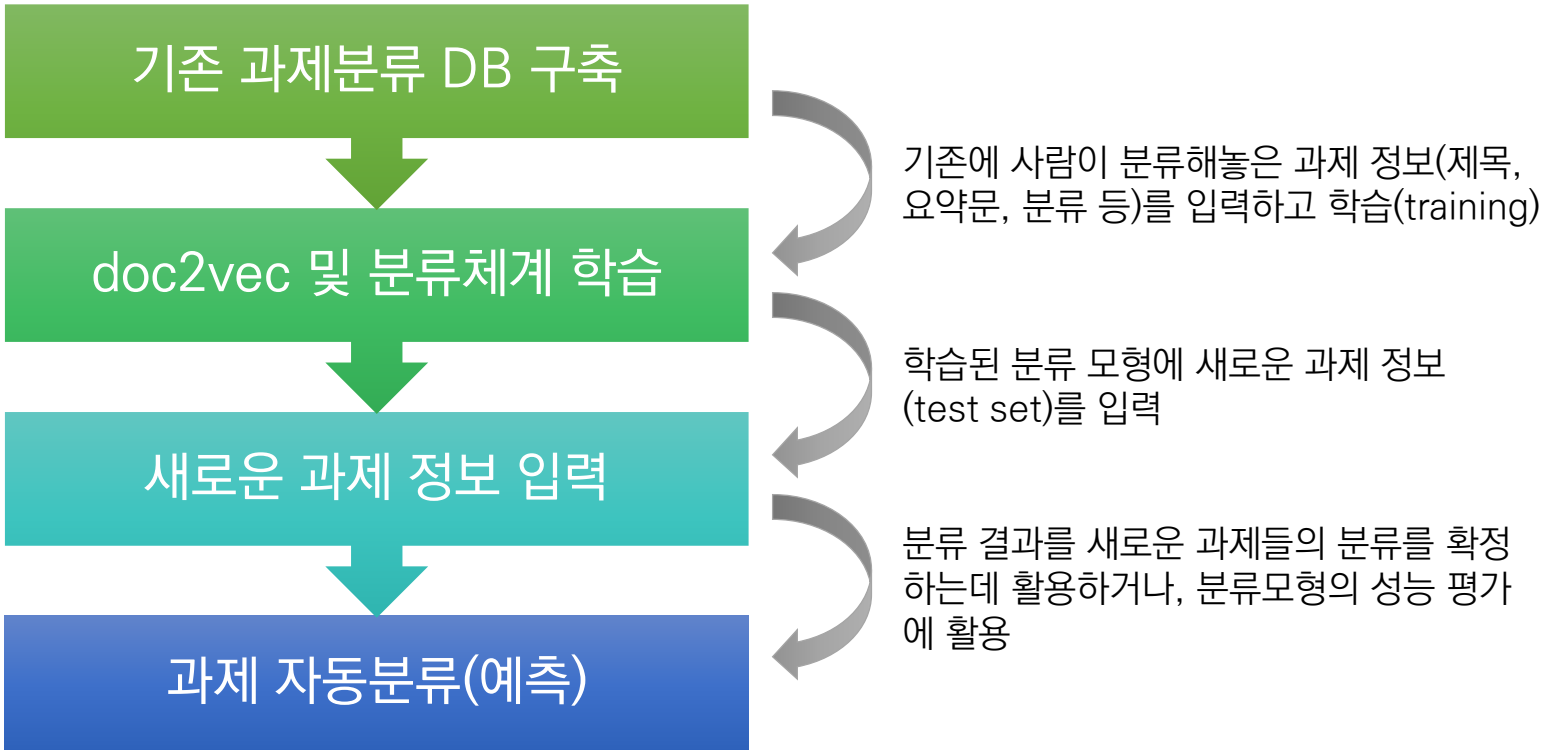
- ’16~’18년 3개 년도를 비교한 결과 :
  ※ 절대개수 또는 금액이 아닌 상대적 비중을 비교하는 것이 바람직
  - 유전자치료제 등 레드바이오에 해당하는 과제 : 60~70% 선에서 유지
  - 레드바이오 內 기초기전 연구의 비중이 증가하는 추세
    ※ 결과의 노이즈를 감안할 필요
  - 생물자원 유전정보 수집관리는 8% 수준으로 일정
    ※ 국가 역할로 일정하게 수행
  - 마이크로바이옴, 동물감염병 백신 분야 증가(비중 小)
- 세부사업 단위 연관성 계층분석 결과(p.32)와 유사
  - 유전자치료, 질병치료 관련 사업이 대부분
  - 육종, 생명유전자원 관련 사업 및 소수 마이크로바이옴 관련 사업 존재

III. 각 기능별 소개 및 활용 예시

③ 과제 분류체계 학습을 통한 자동 분류

활용 시나리오

- 기존에 분류되어 있던 과제를 동 시스템에 학습시킨 후, 새로운 과제들을 자동으로 분류
  - 사람(해당 분야 전문가)이 직접 분류하는 작업량을 경감



신약분야 과제분류 데이터

- 생명의료전문위 소관 정부R&D 사업(과제) 중 신약분야에 해당되는 과제들을 매년 분류한 DB 존재
  - 신약개발 단계, 질환 분야, 의약품 종류 3가지 기준으로 분류
  - '08년 부터 매년 전문가 분류 용역 수행(생명기초사업센터)

- '08~'16년까지 축적된 과제 데이터(7,318개)를 학습시켜, '17년 신규과제(520개)를 자동으로 분류
  - 각 과제의 과제명, 연구목적, 연구내용, 국문키워드, 영문키워드를 학습
  - '17년 신규과제의 전문가 분류도 수행하여 자동 분류가 몇 %의 정확도를 보이는지 평가
  - 총 6개의 학습 알고리즘을 테스트하여 비교함

※ Multilayer Perceptron(Deep Learning), Generalized Linear Model(GLM; 선형회귀), Random Forest, Gradient Boosting Machine(GBM), Support Vector Machine(SVM), Naïve Bayesian model

자동분류 결과

- Top1 예측 정확도 : 50~80% 수준 / Top2 예측 정확도 : 70~90% 수준
  - Top1 예측 : 가장 가능성이 높은 분류 1개만을 제시
  - Top2 예측 : 가장 가능성이 높은 분류 2개를 제시하고 그 중 정답이 있을 경우 정확히 예측한 것으로 간주함
- 질환 종류를 가장 정확히 예측하고, 의약품 종류나 개발 단계는 상대적으로 낮은 정확도를 보임
  - 하나의 과제에서 여러 개발 단계나 의약품 종류에 대한 연구를 수행할 수 있음
  - Top1 정확도는 의약품 종류가 더 높으나, Top2 정확도는 개발 단계가 더 높음
- 가장 높은 정확도를 보인 알고리즘은 Deep Learning(MLP)였음

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%

투자 포트폴리오 산출

- 전문가가 분류한 결과와 자동분류 알고리즘이 분류한 결과로 각각 분야별 투자 비중을 산출했을 때, 최대 4%p 이하의 오차를 보임

신약개발단계

구분			2017년 (모형예측)		2017년 (전문가분류)	
			연구비 (백만원)	비중 (%)	연구비 (백만원)	비중 (%)
타겟발굴 및 검증	타겟발굴및 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	신약플랫폼기술	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

의약품종류

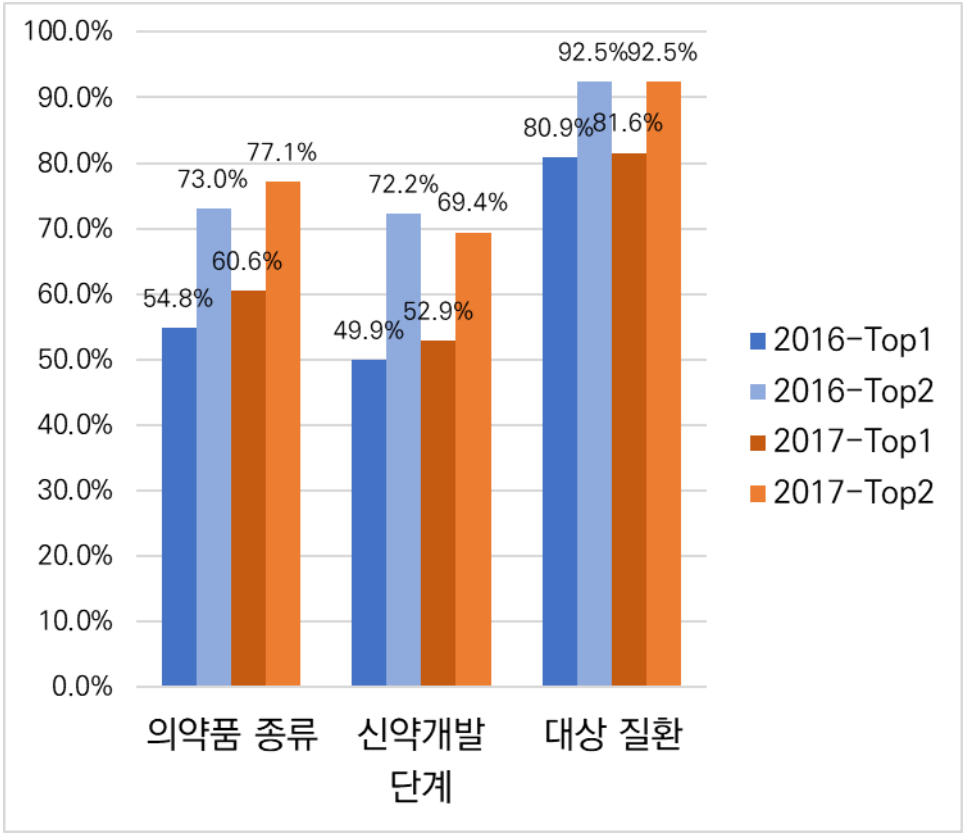
구분		2017년 (모형예측)		2017년 (전문가분류)	
		연구비 (백만원)	비중 (%)	연구비 (백만원)	비중 (%)
합성신약		91,619	26.4	99,204	28.6
바이오신약	단백질 치료제	22,432	6.5	18,888	5.4
	유전자 치료제	12,448	3.6	10,451	3.0
	세포 치료제	20,577	5.9	19,714	5.7
	백신	28,612	8.2	24,171	7.0
	항체	23,004	6.6	22,930	6.6
천연물신약		26,818	7.7	24,942	7.2
개량신약(합성)		11,880	3.4	10,756	3.1
바이오베터	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
합계		347,092	100.0	347,092	100.0

대상질환

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

학습데이터 증가에 따른 정확도 향상

- 2016-Top1,2 : '08~'15년 데이터를 학습하여 '16년 신규과제를 자동 분류한 결과
- 2017-Top1,2 : '08~'16년 데이터를 학습하여 '17년 신규과제를 자동 분류한 결과
  - 가장 좋은 성능을 보이는 알고리즘의 정확도 값을 취함
- 학습 데이터량이 많아질 수록 대체로 정확도가 향상됨



Ⅳ. 추진 경과 및 진행 상황

추진 경과

2017~2019년

- 과제분류, 클러스터링 알고리즘 등 기능 설계 및 개발
- 손쉬운 사용을 위한 그래픽 인터페이스 개발
- KISTEP 업무 활용기법 도출
  - 여러 기능들의 연계활용을 통한 R&D 현황분석 방법론
- (용역사) 애자일소다(AgileSoda)

2020년

- 원내 공동 활용을 위한 온라인화 (컨버전) 추진
- 데이터 확장, 성능 개선을 위한 연구 및 외부 전문가 자문
  - K2Base 메뉴로 추가 추진
  - 해외 부처 과제 데이터 테스트
- (용역사) 미소테크

※ K2Base 개발 및 유지보수 업체

현 진행 상황

- 분석시스템 온라인화 작업(K2Base 연동)
  - 온라인 이용을 위한 개발작업(front/back-end 분리 등) 완료
  - 당초 원내 서버 확충('21년)에 맞추어 서버 자원을 할당받아 서비스를 개시하고자 하였으나, 서버 OS 호환 관련 기술적 문제로 중지되어 있음
  - 개인 계정별 저장공간 등 활용 편의성 개선도 필요함

- 고도화 연구
  - 해외 정부부처 R&D 과제 정보 등 데이터 확장 가능성 확인
  - 일부 기능을 제외하면 정량적 성능 평가 기준 마련이 어렵다는 문제 존재함
  - doc2vec 보다 더 진보된 알고리즘(언어모델) 활용이 가능할 것으로 보이나, 벤치마크 사례를 찾기 어렵고 하드웨어 요구사항 등 전문적 자문이 필요

V. 향후 추진방향(안)

활용 및 고도화(개선) 추진 방향

- 원내 활용을 통해 개선사항 및 문제점, 추가 활용방안 등을 발굴
  - 분석결과의 신뢰성 확인
  - 원내 업무 활용도 평가
    ※ 현재는 기존 업무를 대체하는 성격보다는 보완하는 성격이 강하여 원내 활용 활성화를 위한 독려 필요
  - 추가 기능 발굴
  - 원내 업무의 자동화 & 업무 방식 자체의 진화 추구

- 성능 개선을 위한 최신 언어모델 적용 검토
  - Bert, Electra, GPT-3 등 최신 언어모델은 자연어처리 성능이 대폭 향상됨
  - 임베딩 품질의 향상 → 더 정확한 결과 & 더 많은 정보
  - 단, 보다 높은 컴퓨팅 성능이 요구됨

향후 추진 필요 사항

- 원내 활용을 위한 K2Base 탑재 재추진
  - 원내 서버 호환 문제에 대한 정확한 파악 추진
  - 혁신정보분석센터, 총무전산실, 미소테크 협의 필요
    ※ 현재 동력을 다소 잃은 상태
  - 개발 결과 사장 방지를 위해서라도 서비스 개시 필요

- 원내 차세대 업무 도구로 발전시키기 위한 고도화 방안 마련
  - 시스템 고도화를 위한 IT/빅데이터 분야 기업 협업/자문
  - 민간의 비즈니스 애널리틱스 사례 참조
  - K2Base를 중심으로, KISTEP 업무에 자연어처리 기술을 적용할 수 있는 방안 지속 발굴

KISTEP

감사합니다.