



영재교육 담당교원 직무연수

제4차 산업혁명의 핵심: 융합시대의 빅데이터 및 빅데이터 활용법

2023년 1월 | 경상국립대학교

송지훈 | 대학원 기술경영학과



- | 강의 오리엔테이션 & 융합시대의 빅데이터
- | 교육 분야에서 빅데이터 활용
- | 빅데이터 실습

대학 교육의 패러다임 변화 - 학생 중심에서 수요자 중심으로

□ 정부의 정책: 평생학습 진흥방안

- 제5차 평생교육진흥 기본계획 (2023~2027년)
- 세계경제포럼(WEF)은 2025년까지 전 세계 근로자의 50%가 재교육이 필요하다고 강조
- 적극적인 학습 = 미래핵심역량
- 정부가 주도하던 공급자 중심의 정책방식을 정책 수요자인 국민의 관점으로 전환
- AI·빅데이터 등 디지털 기술을 적극 활용한 개인별 맞춤형 학습으로 전환

평생학습 상시플랫폼으로서 대학의 역할을 확대합니다!

1 성인 역량 향상을 위한 대학의 역할 강화

- 재교육·향상교육 목적의 학점제공이 가능한 마이크로디그리 과정 신설
- 성인학습자 전담대학 확대 및 역량·스킬 중심 K-MOOC 단기과정 개발·확대

2 기업 수요 맞춤형 교육을 위한 대학의 역할 강화

- 신산업·신기술분야 재교육·향상교육에 대한 인센티브 마련 추진
- 기업-대학이 공동개발·운영하는 매치업(Match업)등 교육과정 확대 및 고도화

3 지역주민의 평생학습을 위한 대학의 역할 강화

- 다양한 학위·비학위 과정을 제공하는 커뮤니티 칼리지 모델 도입
- 기초지자체-전문대학 간 컨소시엄 구성, 지역주민 맞춤형 고등직업교육거점 지속 확대

대학 교육의 패러다임 변화 - 학생 중심에서 수요자 중심으로

□ 정부의 정책: 평생학습 진흥방안

**3050 생애도약을 위한
평생학습을 지원합니다.**

1 3050 대상 종합적인 평생학습 지원 강화

- 상담 정신·심리·직업상담 등 각종 상담을 지원
- 학습컨설팅 전문적인 학습·진로설계 등 무료제공
- 학습비 대학 등 다양한 평생학습에 사용할 수 있는 원·패스 카드 지급
- 학습시간 평생학습휴가 및 평생학습휴직 제공 추진
- 학습콘텐츠 각 부처·지자체 특성을 살린 3050 맞춤형 콘텐츠 확대

2 체계적인 지원을 위한 기반구축

- 국가 전담기관을 지정·운영, 타부처·지자체 등과 연계·협력체계 구축
- 30~59세 국민의 특성·수요·효과 등을 분석·연구하는 중점 연구소 지정·운영

**인공지능 기반
맞춤형 평생학습을 제공합니다.**

1 인공지능 기반, 평생학습 원스톱 플랫폼 구축

- 평생학습 정보탐색, 학습이력 활용 등이 가능한 평생학습 원스톱 플랫폼 구축(~'24년)
- 학습자의 학교 및 공공·민간 평생학습기관 등과 연계·협력하여 평생 학습 원스톱 플랫폼에 학습데이터 누적 자동화 추진

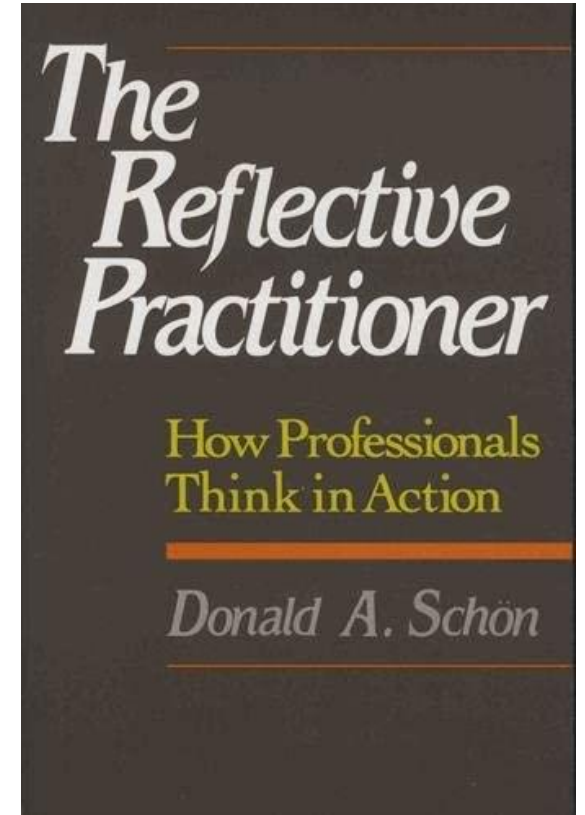
2 평생학습 이력 데이터를 다양하게 활용

- 평생학습 참여·이력 포인트 적립 및 학습비 전환·사용 가능한 '평생학습포인트제' 도입
- 평생학습 빅데이터 및 민·관의 인공지능 연계, 학습이력 기반한 취업 지원 서비스 제공

대학 교육의 패러다임 변화 - 학생 중심에서 수요자 중심으로

□ 누가, 무엇을 어떻게 가르치는 것이 좋은가?

- 1) 이론을 현장에 적용할 수 있도록 교육하여야 하고;
 - 2) 실무에서 이론의 효과를 검증할 수 있어야 하고;
 - 3) 이론이 현장에서 어떤 문제점을 가지고 있는지 이해할 수 있어야 하며;
 - 4) 이론과 실무에서의 차이를 통해서 새로운 개념, 이론, 인사이트 (insight)를 만들어 내는 순환적 과정을 포함하고 있어야 한다
- 이러한 제안점에 기반해서, 현재 운영 중인 수업의 형태, 교수법(pedagogy) 및 학생들에 대한 평가법이 “**Reflective Practitioner**”를 만들기에 충분한 것인지? 생각해봐야 하지 않을까?
- “**교육자**” 또한 스스로 자신의 수업에 대한 문제의식을 갖고, 비판적으로 성찰하며, 학습자가 갖고 있는 경험과 배경, 목적에 관심을 가져야 하지 않을까?



『MIT 교수, 미국의 철학자』
(1983년 저서)

영재교육의 목적



□ 영재 교육의 목적: 특별법인 「영재교육진흥법」에서 영재교육의 정의와 기본 골격을 제시

1조 영재교육의 목적

개인차원

- 개인의 타고난 잠재력 계발을 통해 자아실현하는 것을 목적으로 함

국가차원

- 국가와 사회의 발전에 이바지하는 인재 양성을 목적으로 함

정말 국가와 사회의 발전에 이바지하는 인재의 양성이 현 교육시스템으로 가능한가?

사회 : 사회일반

'SKY' 수시 합격자 2206명 등록 포기...이유
는 '여기' 가려고

중앙일보 | 입력 2022.12.20 20:57 업데이트 2022.12.21 00:00

뉴스 | 사회

“삼성·SK보다 의대?”...반도체학과 합격
자 69% ‘등록 포기’

입력 2022-12-21 11:24 | 업데이트 2022-12-21 11:25





교육의 패러다임 변화 - 해외의 사례

□ Work-based learning

- Work-Based Learning (WBL) is an educational strategy that provides students with real-life work experiences where they can apply academic and technical skills and develop their employability.

TOOL

Work-Based Learning Framework

Work-based learning supports a continuum of lifelong learning and skill development for a range of workers and learners—K-12 students, young adults, college students, adult jobseekers, and incumbent workers.

By Deborah Kobes , Charlotte Cahill , Kyle Hartung

REPORT/RESEARCH

Making Work-Based Learning Work

Work-based learning helps alleviate a common issue for jobseekers: meeting a "relevant work experience" prerequisite that is hard to gain outside of the workplace.

By Charlotte Cahill

~~(구글 검색 결과)
고용주와 학교를 연결하여
학생들에게 특별한 학습 경험을
제공하는 활동을 가리키는 용어~~

학생이 본인의 커리어 설계에 있어
다양한 실무적 경험을 쌓음으로써 직업
선택의 폭을 넓히며, 취업 이후에도
지속적인 교육 경험을 쌓아 발전할 수
있도록 돕는 교육 과정



Problem-based learning 또는 Project-based learning

□ PBL - 전통적 교육환경과 다른 대안적 교육 패러다임

구분	(전통적인) 강의식 수업	PBL 수업
교수자, 학생 역할	교수자: 지식 전달자 학생: 수용자	교수자: 촉진자 & 코치 학생: 능동적인 참여자
수업 진행 순서	개념 및 원리의 전달 → 추상적이고 구조적인 문제 풀이	맥락적이고 실제적인 문제 제시 → 학습 활동
평가 방법	결과 중시 "처방"으로서의 교육	과정 + 결과 중시 "경험"으로서의 교육

- 학교지식 (정답이 있는 결정적인 논리) → 성찰적 실천 (실제적인 상황을 제시하고 학습자로 하여금 성찰적 사고와 활동을 할 수 있도록 촉진 해야 함)
- 학습자가 중심이 되는 학습환경의 구현이 중요
- 교육이론의 발전 → "창의", "융합", "비판적 사고", "STEAM", "메이커 교육" 등



융합시대의 교육

□ 융합시대

- 1) '초연결성(Hyper-Connected)'과 '초지능화(Hyper-Intelligent)'를 통해 모든 것이 상호 연결
- 2) 여러 복합적인 문제들을 융합적 사고로 해결하며, 유연하게 대처 및 적응할 수 있는 능력이 요구됨
- 3) 왜 학습자 중심인가?: 학습자는 각자의 경험에 따라 이미 각양각색의 모습과 결을 지닌 사고체계(인지틀, 이해틀, 신경회로, 스키마)를 지니고 있음
- 4) 제4차 산업혁명 시대의 도래와 더불어 PBL이 재조명 받고 있음
 - PBL을 통해 학생들이 경험할 수 있는 역량과 제4차 산업혁명 시대를 향해 가고 있는 현 사회에서 요구되는 역량이 다각도로 일치하고 있음
 - 대학원: 현장문제 해결형 수업의 운영

융합시대의 교육



□ 대학원에서의 PBL 교육

- 1) 기업의 실무적 문제에 대한 이해 및 기술/서비스 혁신을 위한 이론적 토대 마련
- 2) 프로젝트 진행에 의한 기업 문제 해결 및 경험 축적

Why	What	How	What if
<p>왜 이 문제를 해결해야 하는가?</p> <ol style="list-style-type: none">1. 환경분석2. 문제의 정의	<p>결과물을 도출하기 위해 필요한 지식과 정보는 무엇인가?</p> <ol style="list-style-type: none">1. 기술에 대한 이해2. 고객의 니즈 파악을 위한 기법3. 혁신의 종류 및 그 기대 효과에 대한 이해	<p>어떠한 방법으로 결과물을 도출해 낼 것인가?</p> <ol style="list-style-type: none">1. 신제품 개발 방법론 적용2. 제품 설계 툴 활용3. 제품 인터페이스 설계	<p>최종적으로 제시한 결과물이 적용되었을 때 얻을 수 있는 기대효과는 무엇인가?</p> <ol style="list-style-type: none">1. 신제품에 대한 비즈니스적 배경 제공2. 제품 출시 전 주요 제반사항에 대한 검토

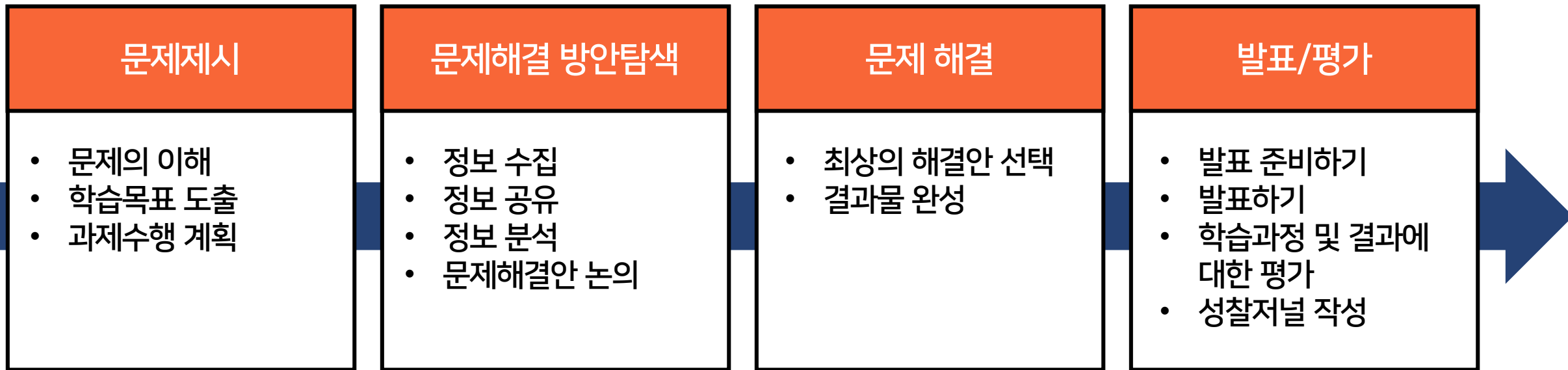




영재교육에 적용을 시킨다면 ...?

□ 초·중·고교에서의 PBL 교육

- 1) 우리 주위에 일어나는 다양한 현상에 대해 배우는 것
- 2) 스스로 문제를 발견하고 사고와 탐구를 통해 실현 가능한 해결책에 도달하는 과정 → 자신의 생각을 만들어 내는 과정



- 학생들이 직접 문제에 참여하고 경험함으로써, 어떤 흥미를 느끼고 지식을 구성했을 때 학습효과가 더 뛰어남
- 교과지식과 문제해결 전략을 동시에 학습



영재교육에 적용을 시킨다면 ...?


- Reflective practice(성찰적 실천)을 통해 획득할 수 있는 가장 중요한 역량은
 - 자신의 생각, 행동을 의심할 수 있고;
 - 의심하고 있는 자신의 행동 및 생각을 해결할 수 있는 대안을 찾을 수 있어야 한다고 주장
- 교수자 입장에서 생각해보면 [...]:
 - PBL 교육의 평가체계에 대한 고찰
 - 새로운 생각과 다양한 의견을 수용하는 방식에 대한 교수법 연구



융합시대

우리가 체감하는 것 이상으로 사회는 무서운 속도로 변화에 직면하고 있다.

21세기 현대사회의 산업 패러다임의 주요 변화 중 하나로 “융합화”를 꼽을 수 있다.

- 융합 (融合, convergence) - 녹을 **융** + 합할 **합**: 서로 다른 두개의 무언가를 녹여서 하나로 합치는 일
- 융합화: 다양한 기술적 가치 및 산업 간의 창조적 결합을 통해 새로운 부가가치를 창출 → 
- 모든 분야의 경계가 무너지고, **기술**과 **경영/사회학**이 융합되고 있다.
- 우리는 이미 “**경계를 넘나드는 사람**” (boundary crosser)의 역량을 필요로 하는 시대에 살고 있습니다. (새로운 핵심 경쟁력으로 부상)
- 기술적 소양만 갖춘 “영재”는 경쟁력이 상대적으로 낮다고 판단 될 수 있습니다. (**기술적** 소양 + **경영/인문학적** 소양 + **디지털 문해력**)

시대적 흐름에 대한 이해가 중요 → 빅데이터 활용 능력



데이터에 대한 사람들의 관심이 왜 커졌을까?

데이터를 적절히 가공해서 의미 있는 정보를 추출하고, 업무 처리시
의사결정을 위한 지식으로 활용

다양한 분야(교육)에 응용할 수 있는 인사이트를 얻을 수 있음

데이터의 양, 종류, 처리속도 기술발전으로 접근성이 좋아짐 (셀프 분석)



우리는 데이터 경제 시대에 살고 있다

데이터 경제(Data Economy): 데이터의 활용이 다른 산업의 촉매역할을 하고 새로운 제품과 서비스를 창출하는 경제

- 데이터의 활용이 경제 및 사회발전에 중요한 원동력
- 데이터 분석 및 인공지능의 광범위한 활용

미래에는 전 산업 분야에서 디지털화·데이터화가 급속히 진행되며 데이터 자본을 기반으로 기존 시장의 질서를 재창조

데이터를 잘 활용하면 생산성이 높아지고 새로운 서비스와 일자리의 창출을 기대 할 수 있고 다양한 산업분야는 지능화를 기반으로 산업의 혁신을 촉진.

데이터가 혁신적인 지식·상품·서비스 창출을 위한 투입요소(자본재)로 활용 ("data as the new currency")

데이터의 유통·거래는 2차적 가치의 재생산 등 고부가가치의 확산의 출발점 → 데이터에 대한 가치 부여



빅데이터 시대

왜 데이터라고 하지 않고 빅데이터라고 할까?

인간과 데이터: 데이터는 누가 만드는 것일까?

- ✓ 인간 - 표현하고자 하는 본능 ("sensible things")
- ✓ 센서 - 사물에 탑재된 센서가 스스로 주변을 탐지해 데이터를 창출, 다양한 종류의 센서 존재

그 결과 다양한 데이터가 존재 (numeric, categorical, relational, audio, video, text data 등)

센서들을 연결 시킨 것 → 사물인터넷 (Internet of Things)

센서들이 우리의 스마트한 생활을 가능하게 하기 위해, 많은 양의 데이터를 생성
(스마트 홈, 스마트 오피스, 스마트 팩토리, 스마트 시티)

디지털 환경에서 발생하는 대량의 모든 데이터 → 디지털 정보량의 기하급수적 증가

빅데이터란 단순히 대용량 데이터만 의미하는 것이 아니라 관리 도구의 능력을 넘어서 데이터에서 가치를 추출하고 결과를 분석하는 기술



빅데이터에 대한 다양한 시각

빅데이터 출현 배경

- Web/Mobile 활동의 증가

트위터 · 페이스북 등 SNS의 급격한 확산
기업들의 데이터 수집 증가 (구글, 아마존, 페이스북, 애플은 핵심 서비스를 통해 방대한 데이터를 수집하고 활용)

- 모바일 기기 사용의 증가

통화 및 문자 메시지 사용 증가
전자기기에 스마트 센서가 추가되고 이로부터 로그가 수집

- 새로운 IT 기기의 출현

RFID 리더기, 다양한 센서 기반 측정기 → 데이터 생산을 더욱 확대

- 멀티미디어 콘텐츠 증가

소비가능한 콘텐츠의 수요가 꾸준히 증가 (다양한 데이터 생산 원천)



빅데이터의 활용 가치

빅토르 신베르거(2015, 옥스퍼드대 교수)

“빅데이터는 새로운 시각으로
세상을 보게 해 주는 안경”

마윈(2015, 알리바바 회장)

“1930년대엔 사람들이 ‘보이지 않는 손’이
시장에 있다고 믿었기 때문에, 그래서
시장경제가 이긴 것” 하지만 손에 데이터를 쥐고
있는 지금의 우리는 예전엔 보이지 않던 그 손을
볼 수 있게 됐다”

제프 베조스 (2017, 아마존 CEO)

“데이터가 상품이나 서비스로 언제
중요해질지 알 수 없기 때문에 우리는 절대로
데이터를 내다버리지 않는다”

손정희(2017, 소프트뱅크 회장)

“데이터는 산업혁명 시대의
석유같은 자원이다”

빅데이터에 대한 다양한 시각

최근 빅데이터의 의미

- ✓ 대용량 데이터의 수집, 저장, 플랫폼, 분석기법 등을 포괄 하는 용어로 변화

최근 빅데이터 분석(Analytics)의 의미

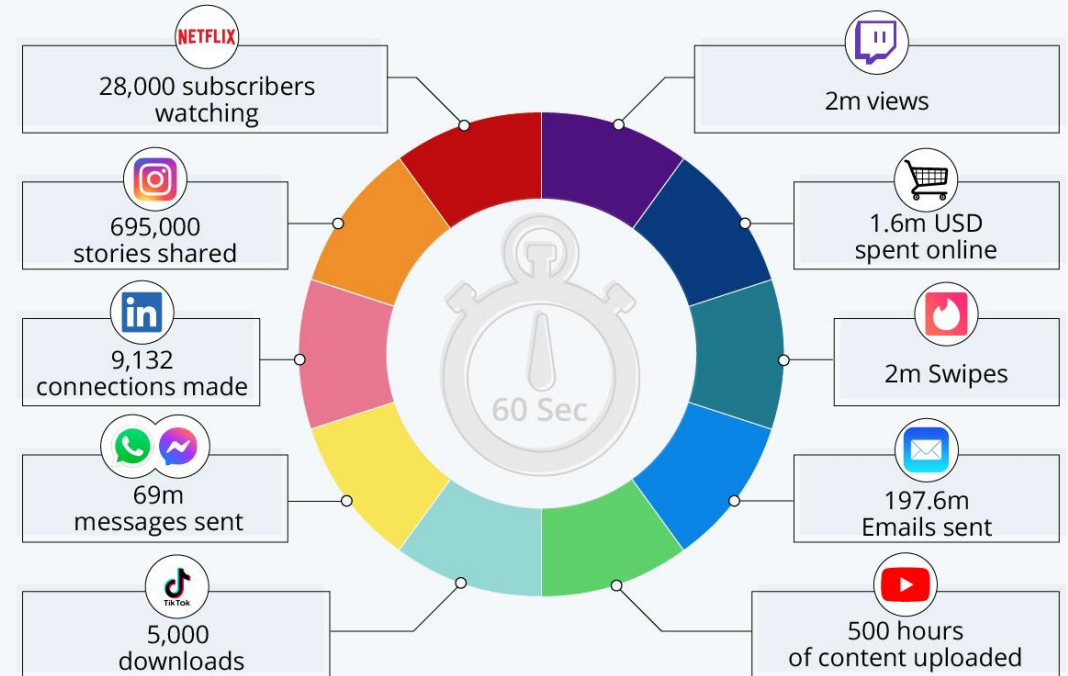
- ✓ 단순 데이터 조회 및 리포팅의 생산과정이 아닌, 데이터에 근간한 통계분석, 트렌드 예측, 최적화
- ✓ 의사 결정을 지원하기 위한 데이터의 광범위한 활용 (탐색적 분석, 정량적 분석, 사실에 근거한 경영)

빅데이터의 활용이란?

- ✓ 생성된 지식을 바탕으로 능동적으로 대응하거나 변화를 예측하기 위한 정보화 기술을 총칭

A Minute on the Internet in 2021

Estimated amount of data created on the internet in one minute



Source: Lori Lewis via AllAccess



statista



스몰데이터 vs 빅데이터 - 공통점과 차이점

빅데이터란?

빅데이터는 전통적인 데이터 처리 방법으로 처리할 수 없을 정도로 대규모이거나 복잡한 데이터를 나타냄.

빅데이터는 흔히 '3V'로 불리는 볼륨(Volume), 다양성(Variety), 속도(Velocity)라는 특성을 가짐.

볼륨은 대규모 크기를 의미하며, 다양성은 데이터 포맷의 다양화 (정형, 비정형) 그리고 속도는 신속하고 효율적으로 처리되어야 하는 특성을 의미.

구분	스몰데이터	빅데이터
분석목적	<ul style="list-style-type: none">표본분석을 토대로 모집단의 특성을 추론비교집단 간 특성 차이, 조치수단 간 효과 차이 비교를 통해 관심대상 집단·수단을 선별하는데 중점	<ul style="list-style-type: none">대규모 데이터에 숨어있는 패턴을 발견하고 규칙을 도출함도출된 패턴 및 규칙을 이용해 개별 대상 별 (고객·제품 등) 액션방안 마련에 중점
데이터 수집	<ul style="list-style-type: none">서베이나 포커스그룹 인터뷰 활용으로 상당한 조사기간이 수반됨	<ul style="list-style-type: none">대규모 내부거래 처리 데이터, 음성, 동영상, 외부공공 및 소셜데이터 활용으로 실시간에 가까운 분석
분석 기법	<ul style="list-style-type: none">기술통계 및 차이분석, 회귀분석, 구조방정식 등 추론통계 중심	<ul style="list-style-type: none">군집분석, 연관분석, 분류분석, 예측분석 등 정형 및 텍스트 마이닝, 딥러닝 기법이 추가됨



스몰데이터 vs 빅데이터 ?

구분	스몰데이터	빅데이터
분석도구	<ul style="list-style-type: none">상용 통계분석 패키지 활용 (SPSS, SAS 등)	<ul style="list-style-type: none">R, 파이썬 등의 오픈소스 및 클라우드 기반의 분석도구 활용
적용분야	<ul style="list-style-type: none">제조, 금융, 유통, 관광, 보건, 행정, 국방 등 공공 및 민간 전분야	<ul style="list-style-type: none">동일한 적용분야를 가지며, 보다 다양한 데이터소스 및 분석기법 활용으로 분석근거의 정확성과 예측력이 향상됨
사용자	<ul style="list-style-type: none">통계학 및 머신러닝 전문가에게 의뢰를 통한 분석	<ul style="list-style-type: none">업무실무자 스스로 셀프분석을 통한 의사결정에 활용

전통적인 접근방법과 최근의 분석기법을 결합해, 균형감 있는 분석을 실시해야 함

쉽게 사용할 수 있는 알고리즘과 오픈데이터 활용의 보편화 → 편의성 증가

빅데이터의 활용은 우리가 논리적으로, 인간의 구조화된 지식 속에서 판단할 수 없었던 많은 숨겨진 패턴이나 지식들을 도출할 수 있게 도와 줄 수 있다.



빅데이터 시대의 데이터의 근본적 변화



기업들이 빅데이터에 많은 관심을 가지고 있는 이유

- ✓ 사용할 수 있는 데이터의 양은 2년마다 2배 이상으로 폭증
- ✓ 고객 행위(customer behavior), 기후 패턴(weather pattern), 금융 시장(financial market), 학습 패턴(learning pattern)에서의 행위 등 파악가능
- ✓ 여타의 현상들에 대한 통찰을 제공해주는 잠재력을 가짐
- ✓ 작은 데이터 셋에 비해 더 많은 패턴들과 흥미로운 현상들을 보여줌

Veracity: 데이터의 정확성

신뢰할 수 있는 데이터 인가? 데이터 품질이 균일한가?

Value: 데이터의 가치

수집된 데이터의 유용성

(The actors producing the data are not necessarily capable of putting it into value.)



5V 모델



그렇다면 필요한 빅데이터 융합 시대에 필요한 역량은?

1) 획일적이지 않은 문제 인식 역량 (“주어진 문제를 푸는데 능숙한 사람이 아닌, 문제를 스스로 정의할 수 있는 역량”)

- 인문학적·감성적·비판적인 상황 해석을 더해 기계와 차별화된 관점으로 문제를 인식할 수 있는 능력
- 복잡한 문제와 직면했을 때 기술적 뿐만 아니라, 문화적 이해와 감성적 해석
- 다양한 자료를 탐색하고, 학습을 통해 문제와의 관련성을 찾을 수 있는 역량
- 일반적인 틀에서 벗어나 문제의 핵심을 해석할 수 있는 역량

2) 다양성의 가치를 조합하는 대안 도출 역량

- 다양한 사람들에게 창의적 의견과 지식을 추출(유인)해낼 수 있는 역량
- 다양한 사람들의 의견을 종합하여 결론을 도출하는 기준과 과정을 설계할 수 있는 능력 (협업 능력)
- 다양한 네트워크의 인적자원을 활용해 대안을 도출해 수행할 수 있는 역량
- 다양한 유형의 정보를 체계적으로 조합하여 지식화 할 수 있는 능력 (시스템적 사고)



그렇다면 필요한 빅데이터 융합 시대에 필요한 역량은?

3) 기계와의 협력적 소통 역량

- ICT 기기의 특성과 그로부터 발생하는 디지털 정보를 이해하고 활용 할 수 있는 능력 (디지털 문해력)
- 첨단기기를 정교하게 조작하거나 감수·보정할 수 있는 능력
- 기계로부터 얻을 수 있는 정보와 사람의 의견을 체계적으로 연결하고 종합할 수 있는 능력

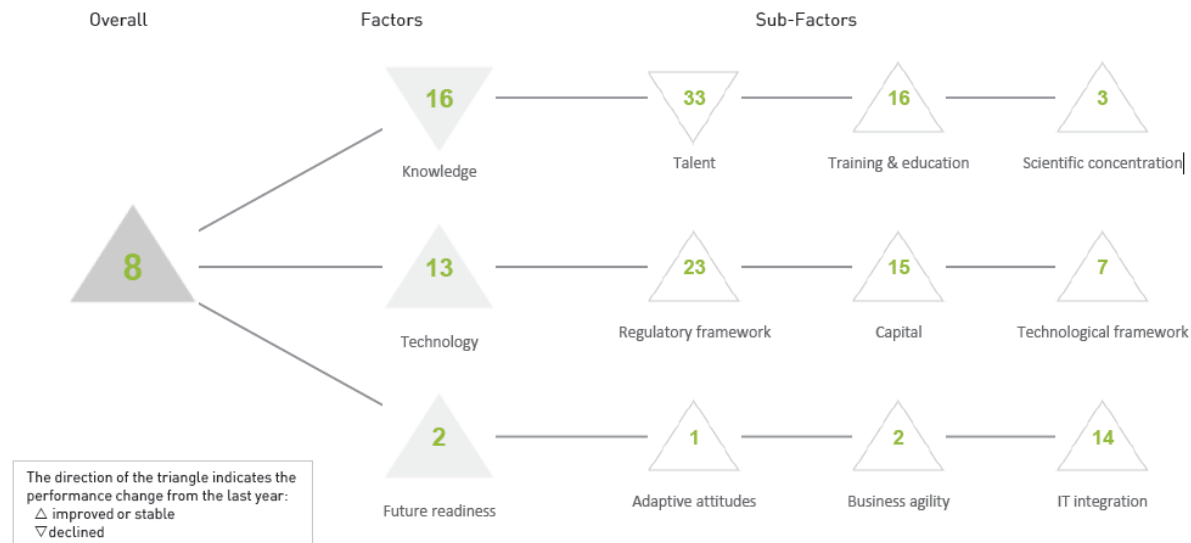
대한민국의 빅데이터 역량

□ 2022 스위스 국제경영개발연구원(IMD) 디지털 경쟁력 평가 결과

DIGITAL TRENDS - OVERALL

KOREA REP.

OVERALL PERFORMANCE (63 countries)



Training & education Rank

Employee training	34
Total public expenditure on education	42
Higher education achievement	04
Pupil-teacher ratio (tertiary education)	30
Graduates in Sciences	11
Women with degrees	20

Business agility Rank

Opportunities and threats	35
World robots distribution	03
Agility of companies	16
Use of big data and analytics	34
Knowledge transfer	30
Entrepreneurial fear of failure	02

교육 분야에서 빅데이터 활용

□ 빅데이터 기반으로 운영되는 교육 시스템

- 미네르바 대학(Minerva University) – 미국 San Francisco에 본부가 위치
- “캠퍼스가 없는 혁신 대학”, “혁신 교육의 대명사”, “세계가 캠퍼스인 미래형 대학”
- 100% 온라인 수업으로 운영 + (반년마다) 세계 7대 도시 순회 교육
- 전 세계에서 우수한 학생들을 모집 중
- 경쟁률은 100:1 → 세계적인 명문인 Harvard University 보다 들어가기 힘들다는 평가를 받고 있음 (합격률 1.6%)
- 그렇다면 수업의 운영은 어떻게?

소규모 세미나 형식 + 모두가 참여하는 토론 수업

학교와 연계된 기업에서 프로젝트를 수행 (아마존, 애플, 우버)



MINERVA
UNIVERSITY

혁신 학교 미네르바스쿨 개요

설립	2010년 설립. 2014년 신입생 29명으로 시작. 매년 150~180명씩 신입생 입학. 5월 말 첫 졸업생 배출
수업 방식	별도 캠퍼스 없이 4년 교육 과정 동안 세계 7개국 돌며 기업 인턴십, 프로젝트 등에 참여하는 현장 실습형 교육. 수업은 100% 온라인 강의로만 이뤄지며 실시간 토론 방식
등록금	3만950달러(약 3600만원, 현지 기숙사·책·생활비 등 포함)
분야	예술·인문, 경영, 컴퓨터과학, 자연과학, 사회과학

교육 분야에서 빅데이터 활용



□ 빅데이터 기반으로 운영되는 교육 시스템

▪ 미네르바 대학(Minerva University)의 수업 운영 방식



▪ 독자적인 “온라인 강의 플랫폼” 소유

▪ 해당 플랫폼을 통해 실시간으로 소통하며 진행되는 수업 방식

▪ 수업은 녹화되며, 교수가 추후에 피드백을 제공

▪ 발표량이 적은 학생은 **빨간색**으로 표시

▪ **초록색**은 활발한 수업 참여도를 나타냄



빅데이터 분석을 기반으로 수행
알고리즘이 자동으로 토론 참여를 측정

교육 분야에서 빅데이터 활용





교육 분야에서 빅데이터 활용

□ 미네르바식(式) 교육의 특징

- 세계 각지를 돌며 기업과 비영리단체·공공기관 프로젝트를 진행하고 현장 경험을 쌓음
 - 실시간 토론 수업을 통해 비판적 사고, 창의성, 커뮤니케이션 능력
 - 설립자는 “왜 기업들은 인재난을 겪고 있을까?” 하는 물음에서 시작
 - ✓ 현재의 교육방식 → 미래 인재 양성에 있어 제 역할을 하지 못함
 - ✓ 환경/기반에 대한 투자가 아닌, 미래 사회에 필요한 인재를 양성하는 **교육 본연의 목표에 집중**
 - 수준별 맞춤 학습(Cross Contextual Scaffolding)
 - 완전히 능동적인 학습(Fully Active Learning)
 - 체계적인 피드백(Systematic Formative Feedback)
- 플랫폼을 바탕으로 맞춤형 교육경험 제공
체계적으로 사고하는 훈련과 빠르게 의사결정을 내릴 수 있는 지혜
- “**에듀테크(Edu-Tech)**”의 적극적인 활용: 빅데이터·인공지능 등 정보통신기술을 활용한 차세대 교육 혁신

교육 분야에서 빅데이터 활용

□ 한국에서도 시도되고 있는 교육 모델

교육부 인가받은 '한국판 미네르바大'...태재 대학교, 내년 문 연다

중앙일보 | 입력 2022.01.25 12:56 업데이트 2022.01.25 13:11

홍지유 기자

구독



- 한샘 창업주가 사재 3000억원을 들여 설립을 추진 중인 대학교
- 미네르바 대학교와 유사하게, 입학하면 4년 동안 한국/미국/중국/일본/러시아의 5개국을 돌며 100% 온라인 학습을 받음
- 교육 목표: 세계 경영인재 육성 → 해외를 주무대로 활동하는 경영인재가 과연 현 시점 꼭 필요한 교육 모델일까?

교육 분야에서 빅데이터 활용

□ 빅데이터를 활용한 학교 교육

- 빅데이터의 활용: 1) 학습자에게 최적화된 학습 자료를 제공할 수 있어 수준별 맞춤 학습이 가능
2) 교수자 지원을 위한 기반기술의 역할 담당
- 교육행정정보시스템(National Education Information System; NEIS)의 활용
 - ✓ 온라인 수업, 디지털교과서 등의 콘텐츠 이용 현황 분석을 통해 보다 효과적인 교육 프로그램 제공
 - ✓ 빅데이터: 학습자 중심의 교육을 실현하기 위한 학습 분석을 가능하게 하는 것
 - ✓ 어떻게, 어떠한 목적을 가지고 데이터를 수집하는지가 중요 → 빅데이터는 분석을 통한 결과에 의미를 부여하는 해석을 통해 가치가 창출
 - ✓ 미국 애리조나주립대(ASU, Arizona State University)의 경우 빅데이터·AI 활용해 학생 중도탈락률을 단축
 - ✓ 애리조나주립대: 연구 경험 있는 학부생, 중도탈락 ↓ 전공-취업 매치 ↑

국내외 교육 분야에서 빅데이터 활용

□ 활용 분야

- 1) IR(Institutional Research, 대학기관연구): 대학 차원의 전략계획 수립, 정책개발 및 의사결정을 지원하는 일체의 활동
- 2) 지능형(적응형) 학습서비스: 다양한 서비스 플랫폼에서 만들어지는 학습자의 성취도, 학습과정 상에서 기록되는 학습패턴, 선호도 정보, 학습자의 관심사들을 데이터화해 분석
 - 학습 지식 구조
 - 문항추천
 - 학생 성취도 분석
 - 학생 성취도 예측
- 3) 소셜 빅데이터 분석: 사회적 여론과 목소리를 반영해 국가의 현안 해결과 정책 수립에 반영
- 4) 공공 분야 빅데이터 활용: 교육문제 현안 해결 가능성 탐색



IR(Institutional Research, 대학기관연구)

□ 활용 예시

- (언급한) 애리조나 대학의 사례
- 대학의 입학생 모집 전략
 - ✓ 입학생 모집을 위한 핵심성과지표(Key Performance Indicators)를 분석하여 광범위한 홍보 대신 선택과 집중 전략 수립
 - ✓ 특정 지역과 시장을 겨냥한 마케팅 전략 수립이 가능 (지리적 정보를 고려해 학생 선발 시 등록율에 대한 예측)
 - ✓ 사회 여러 분야와 마찬가지로 소셜 미디어를 활용해, 잠재적 입학생들의 관심과 학업에 대한 행동모델 등을 분석 (미국 대학 입시 담당자들 중 40%는 지원자의 소셜 미디어를 추가적으로 분석)
 - ✓ 빅데이터를 활용해 학업 중 발생하는 문제에 학교가 개입하여 해결 가능

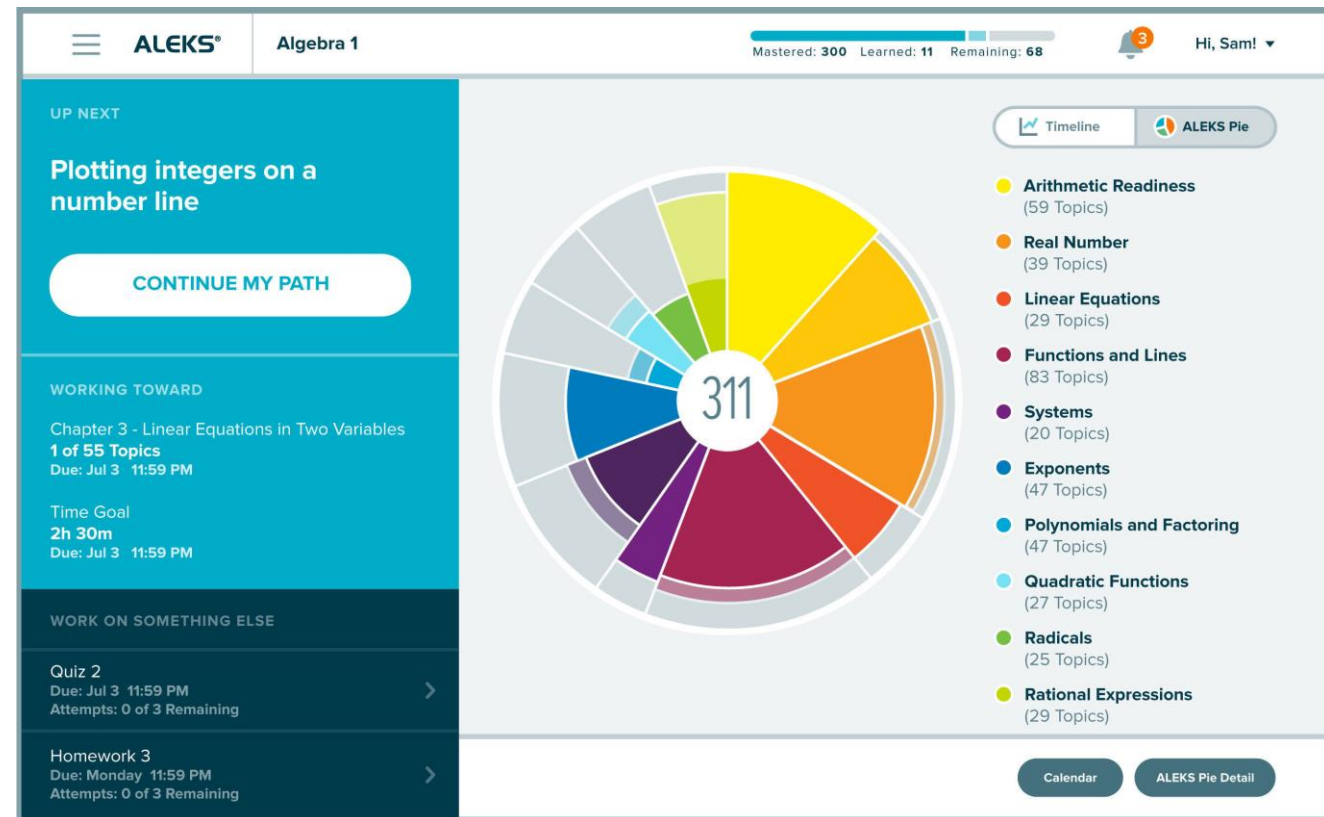
국내외 교육 분야에서 빅데이터 활용



□ “에듀테크(Edu-Tech)” 예시
블랙보드(blackboard)

알렉스(ALEKS: Assessment and Learning in Knowledge Spaces)
→ 초·중·고등 교육

Students			
My Students (313)		Search for student by name or ID	
313 Results Terms: Current Term +2 Clear filters			
Student	Probability of Passing	Courses	Cumulative GPA
Anthony J. M. Quinn 10920559	83 %	CMSC205 CMSC206C	2.79
Diana Randall 10668196	81 %	CMSC205 CMSC206C	2.69
Katherine Alsop 10799053	78 %	CMSC206C CMSC205	2.84
Mary Peake 10431326	76 %	CMSC205 CMSC206C	2.34
Yvonne Reid 10623609	66 %	FREN0101 HIST0101 CMSC205 +1	2.87
Richard May 10330743	66 %	FREN0101 HIST0101 CMSC205 +1	2.74
Claire Karagory 10462794	65 %	GERM0101 CMSC205 CMSC206C	3.46
Connie Chu 10692730	64 %	GERM0101 CMSC206C CMSC205	2.15
Gavin Taylor 10112562	64 %	GERM0101 CMSC205 CMSC206C	2.74



교육 분야에서 빅데이터 활용

□ “에듀테크(Edu-Tech)” 시장의 가파른 성장

- 국내 에듀테크는 빠르게 성장 하고 있는 시장
- 국내 에듀테크 시장은 공교육 보다 “**사교육 시장**”을 중심으로 발전
 - ✓ 민간분야: 학습자의 학습 활동을 진단하고 학업성취를 높일 수 있는 적응형 학습 서비스가 주류를 이룸
 - ✓ 공공분야: 적응형 학습 분석서비스, 교육 행정 지원 서비스, 그리고 머신러닝을 활용한 데이터 분석 연구로 구분이 가능
- 원격 수업의 확산 → 얻어진 데이터를 어떻게 분석하고 활용할 것인가에 대한 사회적 관심과 요구가 높아지고 있음

초/중/고 교육분야에서도 빅데이터에 대한 이해를 기반으로 커리큘럼을 설계



교육기관에서도 빅데이터를 교육 혁신을 위해서 어떻게 활용할 수 있는지 고민해야 함

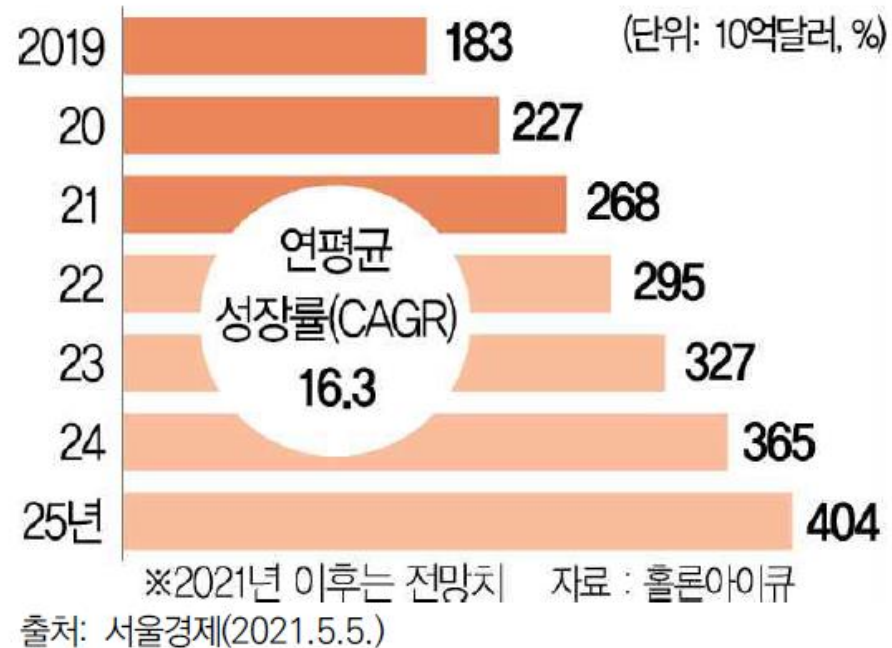
교육 분야에서 빅데이터 활용



□ “에듀테크(Edu-Tech)” 시장의 가파른 성장

- 에듀테크가 교육 분야 디지털 전환(DX)의 핵심으로 성장
- 코로나19로 발생한 대면 교육의 공백을 온라인 교육 플랫폼, 맞춤형 학습 등 에듀테크 서비스를 활용하여 해소

| 전세계 에듀테크 시장 규모 전망치 |



| 교육 분야 지능형 서비스 활용 사례 |

구 분	사 례
맞춤교육 지원	1. EBSi 인공지능 학습진단 서비스, 「DANCHOO」 공공 국내
	2. AI기반 맞춤형 수학 교육 플랫폼, 「Knowre」 민간 국내
	3. AI기반 맞춤형 학습서비스, 「Classting」 민간 국내
	4. 교육 검색 플랫폼, 「QANDA」 민간 국내
	5. AI기반 맞춤형 교육 프로그램, 「Quizalize」 민간 영국
학습경험 혁신	1. 탐구중심 과학교육의 실천, 「지능형과학실」 공공 국내
	2. AR 독서 서비스, 「AR인터랙티브북」 민간 국내
	3. VR 교육 프로그램, 「Vantage Point」 민간 미국
	4. 기업 VR 교육, 「Immerse Virtual Enterprise Platform」 민간 영국
학습격차 해소	1. AI 로봇 기반, 「발달장애아동 스마트안심케어 서비스」 공공 국내
	2. 정부지원 튜터링 서비스, 「National Tutoring Programme」 공공 영국
	3. 자폐 아동 학습을 돕는 사회적 보조로봇, 「KIWI」 민간 미국
	4. 시각장애인을 위한 오디오북, 「RealSAM」 민간 호주

교육 분야에서 빅데이터 활용

□ “에듀테크(Edu-Tech)” 시장의 가파른 성장

- 인공지능, 빅데이터 등 지능정보 핵심기술을 기존 이러닝(e-Learning)에 접목하여 고도화한 개념
- 학습자 개별 맞춤형 서비스의 제공이나 실시간 양방향 커뮤니케이션을 제공
- 가상현실(Virtual Reality, VR) 기술을 활용한 실감형 콘텐츠로 영역을 점차 확대 추세

□ 민간 교육 분야에서의 빅데이터 활용 사례

- 산타토익 (Riiid Tutor, 리이드 튜터)
 - ✓ TOEIC 대비를 위한 지능형 튜터링 서비스 (문항추천 엔진)
 - ✓ 학생 수준을 미리 진단하고 예측해 취약 분야를 집중 보강할 수 있도록 관련 학습 자료를 추천
 - ✓ 다른 영역으로 확대 중: 공인중개사 시험 대비 서비스, SAT와 GRE 관련 서비스

지능형(적응형) 학습서비스

□ Riiid 웹사이트 상의 내용

- AI와 빅데이터를 활용한 교육 혁신



Research in AI Education

Knowledge Tracing

Modeling of a learner's knowledge level by predicting the learner's ability to solve unsolved questions

지식 수준 측정

Student Diagnosis

Extracting assessment questions for the best prediction of user's score and ability

학생 수준 진단

Score Prediction

Overall assessment of a learner's level of knowledge and skills based on the interaction data and metadata

점수 예측

Dropout Prediction

A prediction model aimed to identify learners at risk of leaving a learning session to help optimize engagement

탈락자 예측

Content Recommendation

A recommendation system designed to optimize learning gains along the learning process

콘텐츠 추천

Vocabulary Recommendation

Recommender designed to identify a user's vocabulary level and suggest new words to best fill in the knowledge gaps

단어 추천



지능형(적응형) 학습서비스

□ Riiid 웹사이트 상의 내용

- 단순 서비스를 제공하는게 아니라, 관련 분야 연구(논문발표) + R&D(특허출원)를 통한 혁신

<p>PAPER</p> <p>Fast Fine-tuning of Pre-trained Language Models for Content-based Collaborative Filtering</p> <p>View More →</p>	<p>PATENT</p> <p>Method, apparatus, and computer program for operating machine learning framework</p> <p>View More →</p>	<p>PATENT</p> <p>Methods, devices and computer programs for estimating test scores</p> <p>View More →</p>
<p>PATENT</p> <p>Method, device, and computer program for providing personalized educational content</p> <p>View More →</p>	<p>PATENT</p> <p>How to analyze the data</p> <p>View More →</p>	<p>PATENT</p> <p>Machine learning methods, devices and computer programs for providing personalized educational content based on learning efficiency</p> <p>View More →</p>

- 그렇다면 대학에서는?
- 공공교육기관에서는 어떤 입장을 취해야 할까?



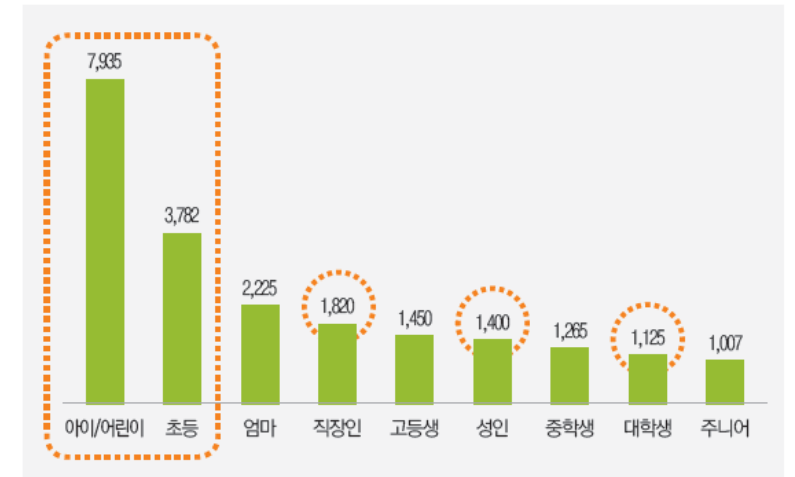
지능형(적응형) 학습서비스

□ 민간 교육 분야에서의 빅데이터 활용 사례

영역 구분	기업명	주요 특징
빅데이터와 인공지능을 활용한 학습 진단 및 분석	웅진 (웅진씽크빅 AI학습)	글로벌 에듀테크 기업인 Kidaptive의 ALP(Adaptive Learning Platform)를 도입하여 학습자의 흥미를 유도하면서 학습에 몰입할 수 있는 서비스 제공 문제 풀이시간 등 학습활동 패턴을 분석하여 학습자의 습관을 교정하는 데 도움 제공
	노리 (KnowRe)	수학 과목에 특화된 맞춤형 학습 서비스이다. 온라인 개인교사 표방 빅데이터와 인공지능 기술을 활용해 학생들의 수준에 맞는 문제를 제공하고, 반복적으로 틀리는 문제에 대해 부족한 개념을 파악하고, 이를 집중적으로 학습할 수 있는 콘텐츠 제공
	마타수학	빅데이터와 인공지능 기술을 활용하여 학습자의 취약한 수학 개념을 분석하고, 맞춤형 문제 추천 서비스; 학습자 맞춤형으로 최적의 학습경로를 제시
	천재교육 (밀크T)	빅데이터와 인공지능 기술을 활용하여 학습자의 학습 유형을 분석하여 맞춤 시간표와 학습법 추천; 유명 강사진의 강의를 제공하고, 전문가의 1:1 학습관리 서비스 제공

소셜 빅데이터 분석

- 빅데이터 분석을 통한 새로운 타겟 고객층 발굴
- “스테디톡” 원어민 화상영어 서비스
 - ✓ 핵심고객군 분석: SNS 데이터를 활용해 화상·전화영어 언급 직업 군 분석
 - ✓ 화상·전화영어와 함께 언급되는 서비스 선택기준에 대한 키워드 분석
 - ✓ 화상·전화영어 언급량을 채널별로 분석



[화상 / 전화영어 언급 직업군 분석]

분석개요

- 분석대상기간 | 2014. 10. 1 ~ 2015. 9. 30
- 정보출처 | 뉴스, 트위터, 커뮤니티, 블로그, 카페

평균
수업시간



평균
수강기간



직장인 / 대학생



17.5분

4.3개월

초등학생



29.4분

11.8개월

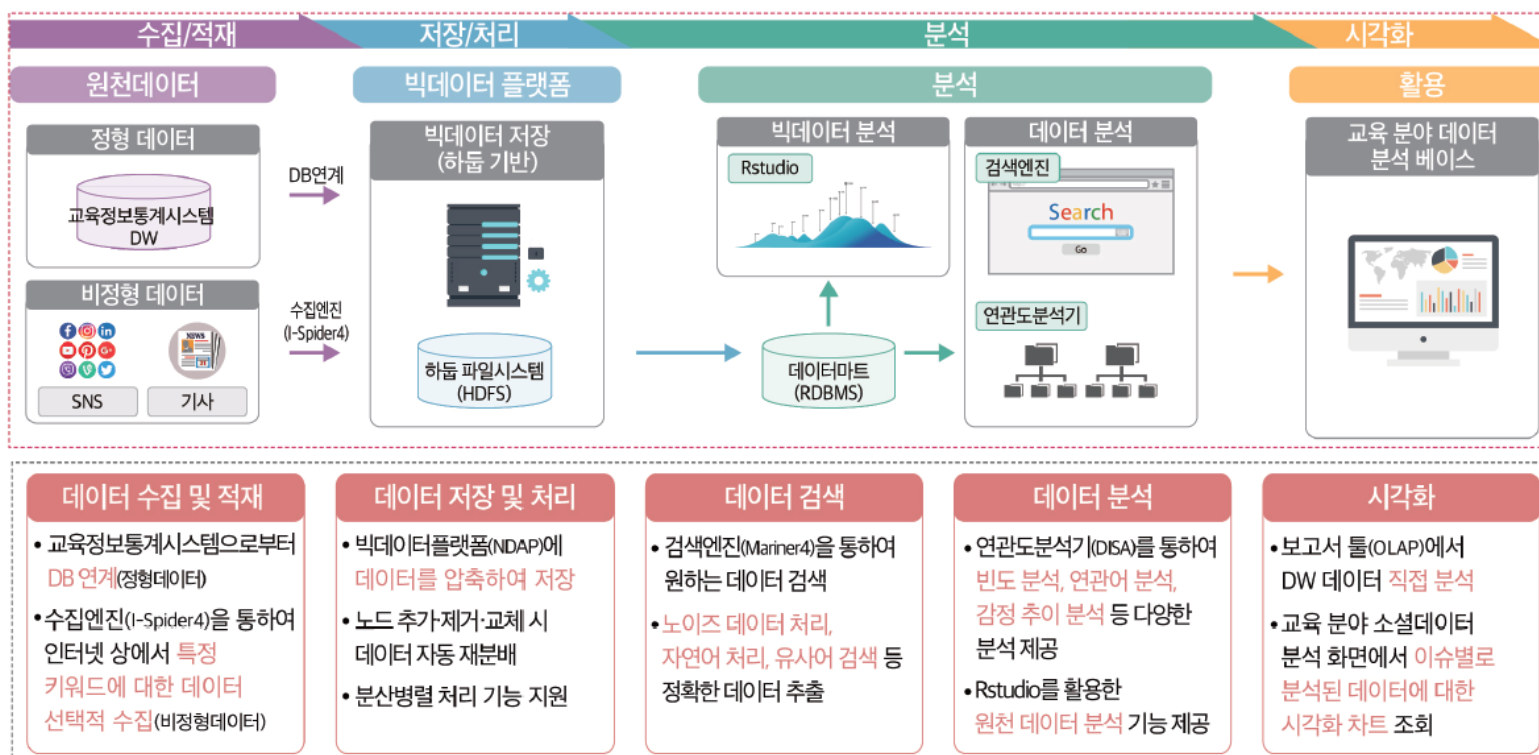
[직장인/대학생 vs. 초등학생에 대한 수업데이터 분석]

공공 분야 빅데이터 활용

□ 공공 교육 분야에서의 빅데이터 활용 사례

▪ 교육 행정 지원 서비스

- 지능형 교육정보통계시스템(Education Data System: EDS) → 빅데이터 분석 기능을 강화한 지능형 시스템 개통



공공 분야 빅데이터 활용

□ 공공 교육 분야에서의 빅데이터 활용 사례

1) 머신러닝을 활용한 데이터 분석 연구

○ 특성화고에 대한 취업률 예측 모델

- ✓ 2019년 기준 마이스터고 포함 전체 고등학교의 21.7%를 차지
- ✓ 그러나 취업률은 2017년 이후 하락세를 보임
- ✓ **목적:** 취업률이 낮을 것으로 예상되는 학교에 개선 방안 제시
- ✓ 머신러닝 알고리즘을 활용한 예측 모델의 개발
- ✓ 취업률에 미치는 영향 요인은 크게 (학교환경 요인), (교수학습 요인), (개인특성 요인)으로 구분이 가능
- ✓ 교수학습 요인: 동아리 활동, 방과후 학교, 기타 선택적 교육 활동, 교육활동 예산

공공 분야 빅데이터 활용

□ 공공 교육 분야에서의 빅데이터 활용 사례

1) 머신러닝을 활용한 데이터 분석 연구

○ 특성화고에 대한 취업률 예측 모델

- ✓ 어떤 변수들이 더 큰 영향을 미치는지를 파악할 수 있으며, 이를 바탕으로 개선 방안
에 대한 논의를 데이터 기반으로 진행할 수 있음

<표 III-20> 취업률 중위값 기준 주요 변인별 평균값 t검정 결과

변인 구분	취업률 중위값 이하(n=1,615)	취업률 중위값 초과(n=1,614)	t(p)
학생당 교육활동 지원 예산***	10,715,216원	17,933,147원	7.24(0.00)**
학생당 선택적 교육활동 예산***	9,554,510원	15,341,184원	6.04(0.00)*
동아리 참여비율***	75.05%	109.73%	17.76(0.00)**
학생당 학교시설 확충 예산	3,370,124원	4,903,929원	1.40(0.16)*
학생당 장학금***	282,505원	250,618원	-3.23(0.00)*

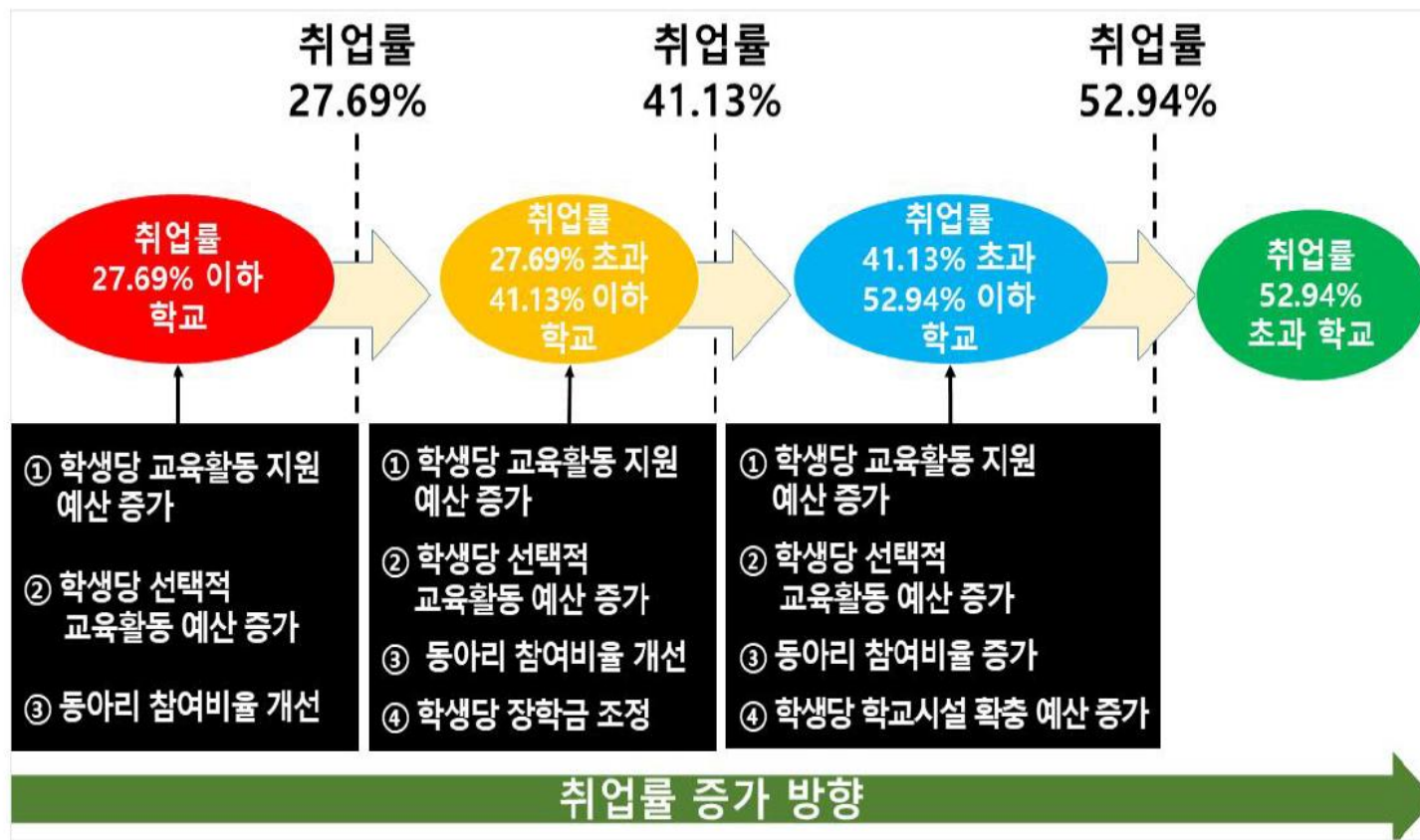
<표 III-16> 변인별 영향력 분석(중위값 기준)

연번	변인 구분	정보획득 지수 값
1	학생당 교육활동 지원 예산	0.1080
2	학생당 선택적 교육활동 예산	0.1019
3	동아리 참여비율	0.0893
4	학생당 학교시설 확충 예산	0.0545
5	학생당 장학금	0.0498
6	시군명	0.0495
7	학생당 학비감면액	0.0419
8	시도명	0.0340
9	급식비 지원 학생 수 비율	0.0339
10	학업중단율	0.0313
11	학생당 급식비 지원액	0.0296
12	남학생 수 비율	0.0157
13	여학생 수 비율	0.0157
14	학제명	0.0128
15	교장공모제여부	0
16	지역행정구분명	0
17	방과후학교 참여비율	0

공공 분야 빅데이터 활용

□ 공공 교육 분야에서의 빅데이터 활용 사례

1) 머신러닝을 활용한 데이터 분석 연구



공공 분야 빅데이터 활용

□ 공공 교육 분야에서의 빅데이터 활용 사례

2) 머신러닝을 활용한 데이터 분석 연구

○ 동영상 기반 학습환경에서 **학업성취 예측 모형**

- ✓ 동영상 강의의 보편화 및 동영상 강의의 교육 효과성 측정을 위한 학습자 로그 데이터 분석
- ✓ 온라인 환경에서 의미 있는 학습자 정보를 추출하고 학습 효과성을 평가하는 것에는 어려움이 존재
- ✓ **목적:** 실시간으로 기록되는 학습 로그 데이터를 활용한 학업성취 예측모델 개발
- ✓ 도출하는 공통된 주요 변수는 추후 교수설계에서 중요한 요소로 고려가 가능 → 동영상 학습 설계 환경의 개선
- ✓ 빅데이터가 아닌, 실험을 통한 데이터 획득으로 실무적인 적용성에 대해서는 별도의 검토 과정을 거쳐야 함
- ✓ 의미 있는 변수들을 로그데이터로 활용했는가?
- ✓ **하지만 이러한 실험은 학교별로/학급별로 충분히 해볼 수 있을 만한 “연구”**

공공 분야 빅데이터 활용

□ 공공 교육 분야에서의 빅데이터 활용 사례

2) 머신러닝을 활용한 데이터 분석 연구

- 동영상 기반 학습환경에서 **학업성취 예측 모형**
 - ✓ 상호작용이 가능한 동영상 학습 환경에서의 학업성취가 있을 것으로 예상
 - ✓ 코멘트 작성과 동영상을 일시정지 하는 행동이 학업성취 분류를 위한 주요 속성으로 해석
 - ✓ 일시정지: 동료 학습자가 작성한 코멘트를 확인 또는 북마크 설정과 같이 학습에 있어 적극적인 행동을 보임

표 3. 행동로그 지표 분류

분류	변수명	지표 의미
행동 빈도	Annotation.Freq	북마크 및 코멘트 열람 빈도
	Bookmark.Freq	북마크 설정 및 확인 빈도
	Comment.Freq	코멘트 작성 및 확인 빈도
	Condition.Freq	외부환경 조정 빈도
	Filtering.Freq	주석 필터링 빈도
	Pause.Freq	일시정지 빈도
	Play.Freq	재생 빈도
	Seek.Freq	동영상 탐색 빈도
	Slide.Freq	슬라이드 조정 빈도
	ReplyIn.Freq	동료 학습자 코멘트 답글 작성 빈도
	LikeIn.Freq	좋아요 누른 빈도
	ViewIn.Freq	동료 학습자 코멘트 확인
행동 시간	Paused.TTime	총 일시정지 시간
	Play.TTime	총 재생 시간
	Annotation.Time	북마크 및 코멘트 열람 시간
	Comment.Time	코멘트 작성 시간
	Condition.Time	외부환경 조정 시간
	Filtering.Time	주석 필터링 시간
	Seek.Time	동영상 탐색 시간
	Slide.Time	슬라이드 조정 시간
	Bookmark.MTime	북마크 설정 시간

공공 분야 빅데이터 활용

□ 공공 교육 분야에서의 빅데이터 활용 사례

3) 머신러닝을 활용한 데이터 분석 연구

- 머신 러닝을 활용한 이러닝 학습 환경에서의 **학습자 성취 예측 모형 탐색**
 - ✓ 이러닝 강의의 학습자의 데이터를 토대로 학습 성취 수준을 예측
 - ✓ 대학교육에서는 이러닝 및 MOOC, EdX 등을 개방형 플랫폼 도입으로 다수의 학습자를 대상으로 한 온라인 교육이 점차 증가하는 추세
 - ✓ **이러닝 강좌의 수강은 기하급수적으로 증가하고 있으나, 평균 이수율은 채 10%에 미치지 못하는 문제가 발생**
 - ✓ **목적:** 출결이나 과제 등 여러 요인들이 실제 학습 성과에 얼마나 영향을 미치는지 예측 모형을 개발해 분석
 - ✓ 이러닝을 중심으로 한 과학교육의 발전 가능성에 대한 시사점 제공
 - ✓ 주의 집중과 학습 지연 행동 방지를 위해서는 이러닝 후반 주차 운영에 매우 주의를 기울여야 함 (기말고사가 가장 큰 영향을 미침)

공공 분야 빅데이터 활용

□ 공공 교육 분야에서의 빅데이터 활용 사례

4) 머신러닝을 활용한 데이터 분석 연구

- 빅데이터를 활용한 학업중단 학생 예측 모델 (및 이를 통한 체계적인 관리 체계 구축)
 - ✓ 시·도 교육청 및 학교에서도 학업중단을 예방하기 위한 다양한 노력을 기울이고 있음
 - ✓ **목적:** 학업중단 위기에 놓인 학생들을 조기 발견
 - ✓ 학교생활에 부적응하는 잠재위험군의 학생들을 체계적으로 진단하는 건 매우 중요한 과제
 - ✓ **일반화 선형 모형 + 머신러닝 모델을 활용한 예측**
 - ✓ 학생 수준 변수 가운데 성별, 학업성취도, 출결사항, 체험활동 등이 학업중단에 영향을 미침
 - ✓ 학교 수준 변수로는 학교규모 등이 학업중단에 영향을 미침

공공 분야 빅데이터 활용

□ 공공 교육 분야에서의 빅데이터 활용 사례

4) 머신러닝을 활용한 데이터 분석 연구

○ 빅데이터를 활용한 학업중단 학생 예측 모델 (및 이를 통한 체계적인 관리 체계 구축)

- ✓ 빅데이터 연구방법을 활용하여 학교교육에서의 중요한 성과 (Learning Outcomes)에 대한 예측이 가능
- ✓ 학생의 인적사항, 교육정보통계 자료 및 나이스 업무 자료 현황 분석을 통해 다양한 변수들을 고려
- ✓ 학생 대응 모델 추진을 위한 제도적 전략
 - a) 표준화된 학업중단 학생 대응 시스템을 개발하여 시·도 교육청에서 활용할 수 있도록 하는 방안 제언
 - b) 지역의 특성을 고려하여 독자적인 학업중단 학생 대응 시스템을 구축하는 방안 제언

교육 분야에서 빅데이터 활용

□ 민간/공공 교육 분야에서의 빅데이터 활용 사례

▪ AI 학습 분석 시스템 구축

동아일보 | 사회

사이버대 최초 AI학습분석시스템 구축... 재학생 자기주도 학습 강화

박서연 기자

입력 2022-12-15 03:00 | 업데이트 2022-12-15 03:00

HOME > 커뮤니티

세종사이버대학교, 역량기반 지능형 스마트러 닝시스템(ISLS) 오픈

이태훈 기자 | 승인 2022.12.12 16:31

가 < > □

- ✓ (사립) 대학 교육은 오프라인과 온라인의 경계가 점점 희미해지고 있음
- ✓ 학습경험 데이터를 기반으로 학업 중도 포기자나 탈락자를 예측해 학업을 중단하는 일이 없도록 예방
- ✓ 개별화된 관심도를 분석해 적성과 역량에 따른 교과·비교과 과목을 추천하는 등 맞춤형 교육을 제공
- ✓ 학습자별 학습지수 정보를 개별 학습자, 교수자, 관리자에게 제공해 필요한 시기에 집중적인 학습관리 지원

교육 분야에서 빅데이터 활용

□ 공공 교육 분야에서의 빅데이터 활용 사례

▪ 현장에서의 빅데이터 교육

✓ 학습자 및 교수자 모두 노력해야 하며, 이는 민·관 협력해 지속가능한 생태계 조성이 필요

“빅데이터 교육, 학생 정보문해력 키우는데 꼭 필요”

충남교육청, 첨단기술 활용한 교육과정 다양화 ‘눈길’

기사입력 2019-07-21 16:34:17 | 최종수정 2019-07-22 00:45:57 | 박근주 기자 | springkj@hanmail.net

100만 디지털 인재 양성...교육현장 어떻게 달라지나

대학 첨단분야 정원 확대 등 제도 개혁...비전공자 대상 융합과정 운영

초·중학교 정보·컴퓨터 수업 시간 대폭 확대...교원 전문성 향상도 추진

유아 놀이 매체로 활용 자료 개발...전 국민 대상 디지털 교육 기회 확대

2022.08.26 정책브리핑 윤세리



빅데이터 실습

| 토픽 모델링이란?

| 토픽 모델링 실습 (구글 Colab 활용)

<https://github.com/gnu-mot/tutoring>

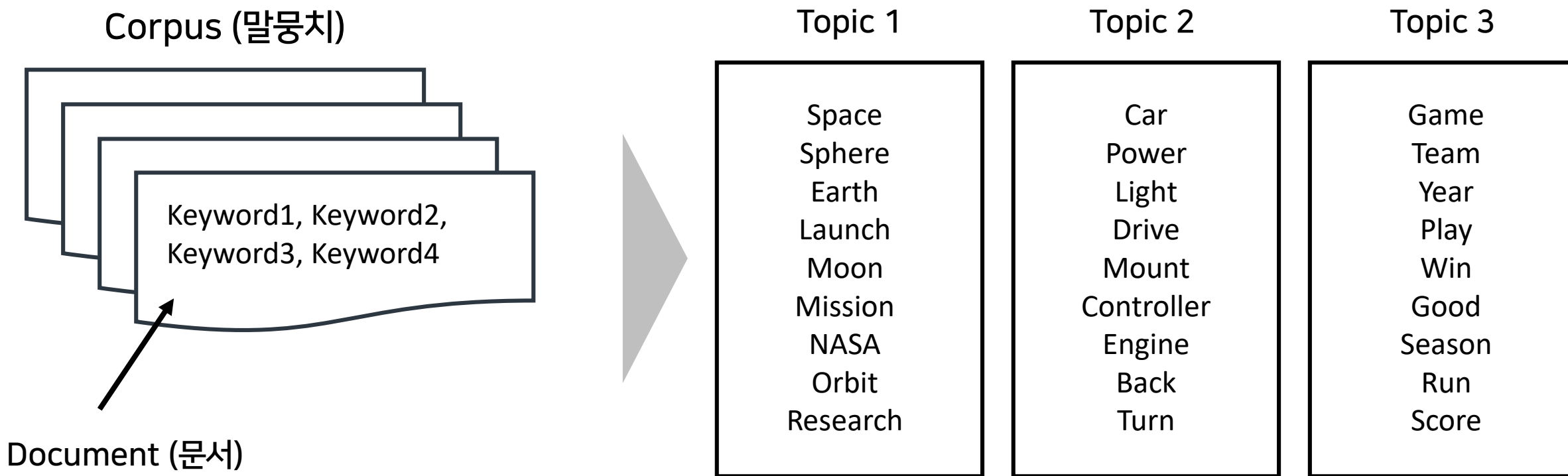
토픽 모델링이란?



토픽 모델링 (Topic modeling)

- 문서 집합에서 주제를 찾아내는 기법 → 문서내에 존재하는 다양한 키워드를 바탕으로 주제(topic)를 도출하는 통계적 분석 방법 (아이디어: “방대한 양의 문서가 존재할 때 누가 이것 대신 읽고 주제를 파악해줄 수 있을까?”)

1) 토픽차원에서 설명: 개별 토픽을 기준으로, 어떤 단어들이 빈번하게 등장을 하는가 (토픽별 단어의 분포)

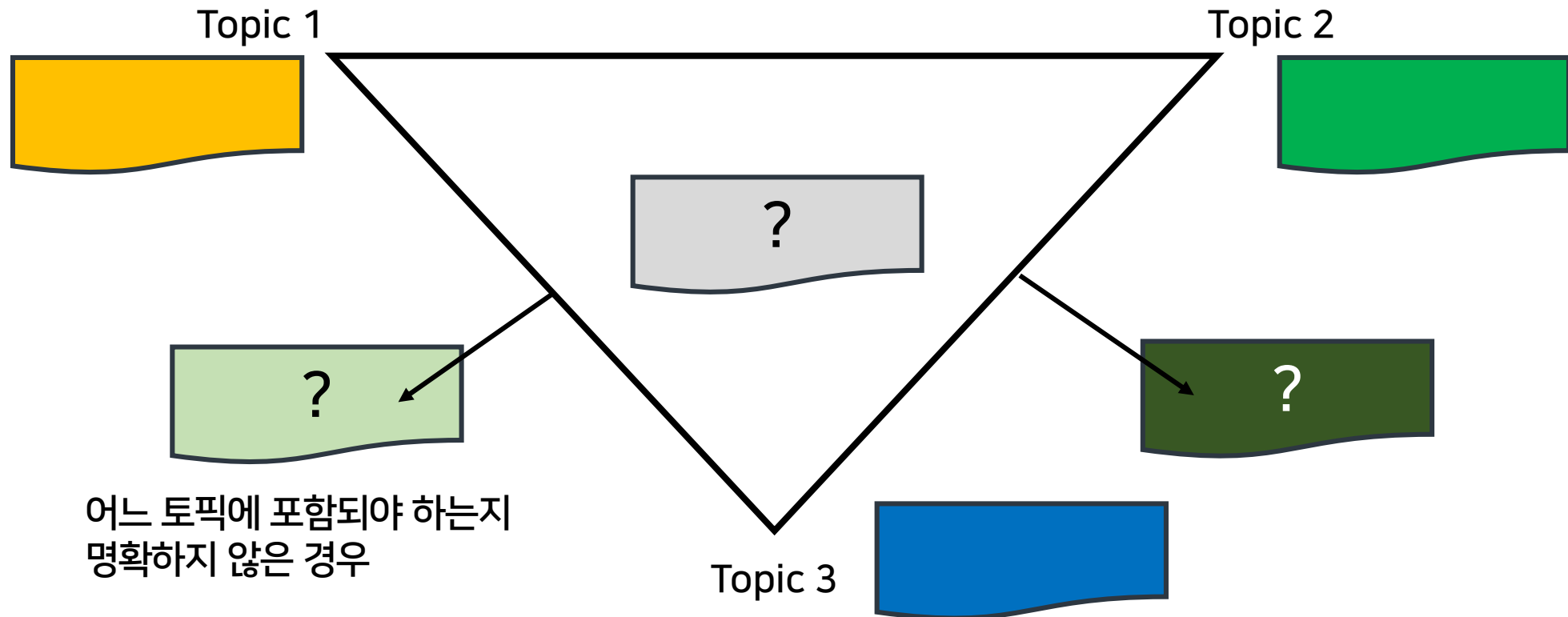


토픽 모델링이란?



토픽 모델링 (Topic modeling)

2) 문서차원에서 설명: 각각의 개별문서가 어떤 토픽을 더 많이 내포하고 있는지 파악
(특정 토픽이 해당하는 문서에서 차지하는 비중을 산출, 문서별 토픽의 분포)



토픽 모델링이란?

토픽 모델링을 이용한 핀테크 기술 동향 분석

김태경, 최희련, 이흥철*
고려대학교 산업경영공학과

A Study on the Research Trends in Fintech
using Topic Modeling

Table 3. Topic Modeling

no.	Topic Word
Topic 1	transfer, account, send, machine, ATM, automat, machine, device, money, bank
Topic 2	trade ,order, price, market, exchange, product, trader, match, rate, valu
Topic 3	modul, mobil, payment, card, nfc, chip, function. device, equip, phone
Topic 4	cpo, identifi, offer, provid, control, accept, financi, plan, custom, manag
Topic 5	network, secur, support, access, modul, embodi, comput, software, onlin, program
Topic 6	loan, amount, guarante, mortgage, lend, borrow, home, incom, benefit, rate
Topic 7	display, interfac, imag, machin, plural, graphic, screen, configur, type, view
Topic 8	custom, servic, provid, investment, option, allocation, analyze, monetary, bill, databas

수 있다. 그 결과 Topic 1 은 ATM(Automated banking), Topic 2는 거래 및 교환, Topic 3은 NFC(Near field communication), Topic 4는 금융 데이터 관리, Topic 5는 금융 소프트웨어, Topic 6은 주택 담보 대출, Topic 7은 디스플레이, Topic 8 자산관리, Topic 9는 보안, Topic 10은 인터넷 전문 은행, Topic 11은 경매 및 입찰, Topic 12는 재무 리스크 관리, Topic 13은 모바일 결제, Topic 14는 신용카드 결제, Topic 15는 금융데이터 분

“중요하지 않다”의 의미가 아니라, 상대적으로 특허 출원이 감소하고 있는 분야로 해석

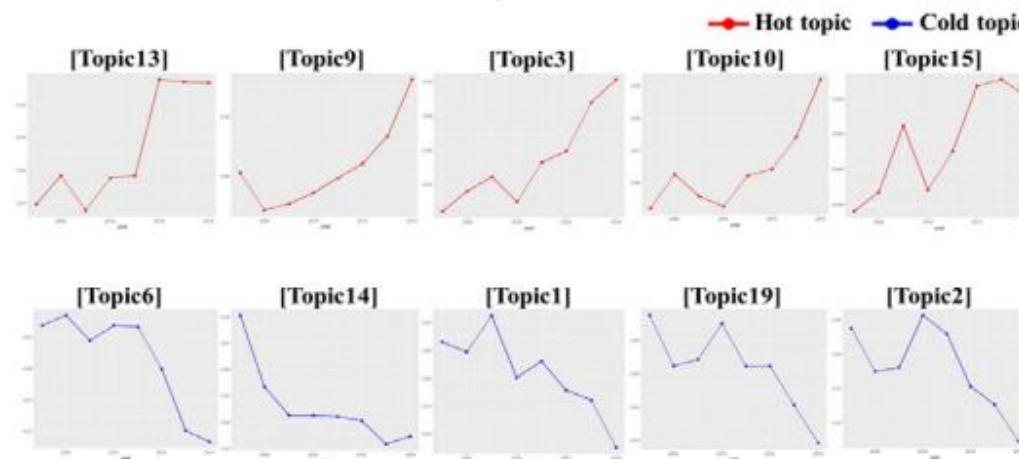


Fig. 4. Hot&Cold Topic of Total country

토픽 모델링이란?

토픽의 네이밍은 자동으로 주어지는 것이 아니라 연구자의 분석/판단에 따라 네이밍이 결정 ("단점")

토픽 네이밍을 잘 하려면 "domain knowledge"가 필요

Table 1 20 topics and keywords in topics

[T1] Robotics	[T2] Text/Word searching	[T3] Video	[T4] Programming	[T5] Computer management
sensor power control robot fluid light system	user search context input display be interest	display frame motion screen camera video color	code processor program array element data address	product task system item tool work computer
[T6] Programming syntax	[T7] SQL	[T8] Text preprocessing	[T9] Network speed adm.	[T10] General networking
first second portion set parameter section end	data system input store event be set	document text speech web word file search	time level real high traffic line speed	network node neural data packet protocol computer
[T11] VLSI	[T12] Network administration	[T13] Object recognition	[T14] Problem solving method	[T15] Multimedia
signal input unit output circuit control pattern	server agent event client call request session	object region light subject scene interest detector	model rule state system process problem decision	content media audio video user music stream

토픽 모델링이란?

토픽 모델링 (Topic modeling)

- 방대한 양의 문서(비정형 데이터)에서 추상적인 토픽을 발견하기 위한 통계적 모델
- “특정 주제에 관한 문서에서는 특정 단어의 등장 빈도가 더 높을 것이다”라는 직관에 기반
- 문서에 숨겨진 의미구조를 발견하기 위한 방법 중 하나 → 여러 단어들의 동시 출현 횟수 및 확률이 중요
- 문서에서 발견되는 키워드 분포 분석을 통해, 문서들을 주제별로 분류가 가능 → 비지도 학습 알고리즘에 해당
- Clustering(군집화)의 경우 하나의 문서는 하나의 그룹에만 속할 수 있지만, 토픽 모델링 기법 중 LDA 방법을 적용하면, 하나의 주제가 여러 문서에 동시에 존재할 수 있는걸 나타낼 수 있음
- 주로 사용되는 토픽 모델링 기법으로는 LDA (latent dirichlet allocation, 잠재 디리클레 할당) 기법이 있음 (여러 기법이 있으나, 여기서는 LDA만 다룰 예정)



LDA

LDA (latent dirichlet allocation)은 대표적인 토픽 모델링 알고리즘 중 하나

- 주어진 문서에 대하여 각 문서에 어떤 토픽들이 어떤 분포로 존재하는지에 대한 확률 모형
- 문서내의 단어 (또는 키워드) 기반 토픽별 단어의 분포, 문서별 토픽의 분포 두 가지 모두 추정
- Latent → 문서내 여러가지 하위 주제들이 숨겨져 있다
- Dirichlet → 디리클레 분포 (단어가 한 주제에 속할 확률 값을 추정하는데 사용, 확률분포)
- Allocation → 할당 (문서에 주제를 할당, 단어를 주제에 할당)
- 하나의 문서는 여러 주제로 구성이 되어있고, 문서의 주제 분포에 따라 단어의 분포가 결정된다고 가정 (Each document as a mixture of topics; each topic as a mixture of words)

LDA



Topics

gene 0.04
dna 0.02
genetic 0.01
...

life 0.02
evolve 0.01
organism 0.01
...

brain 0.04
neuron 0.02
nerve 0.01
...

data 0.02
number 0.02
computer 0.01
...

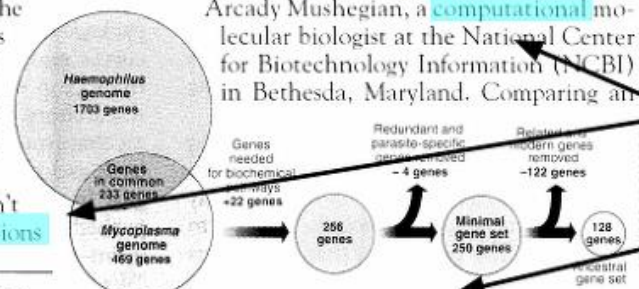
Documents

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions

"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

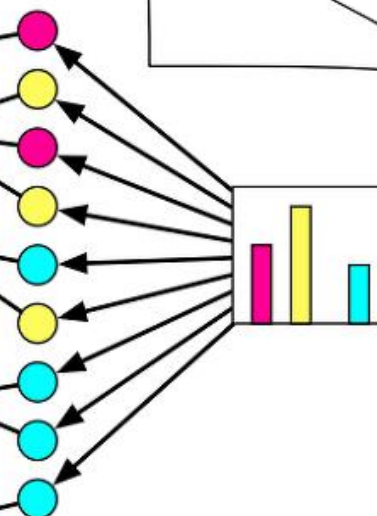


* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

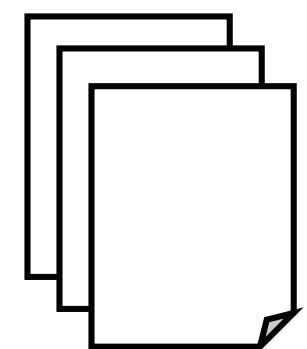
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

Topic proportions & assignments



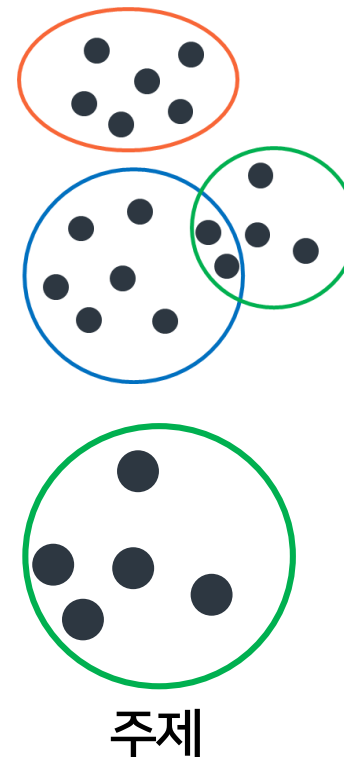
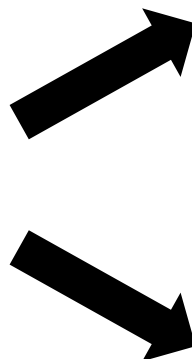
LDA



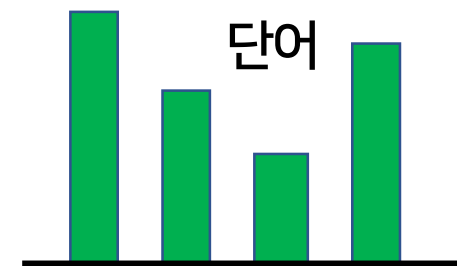
문서들의 집합



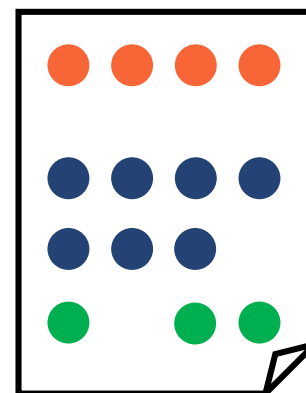
토픽 모델링



주제별 (k-topics)
단어들의 분포



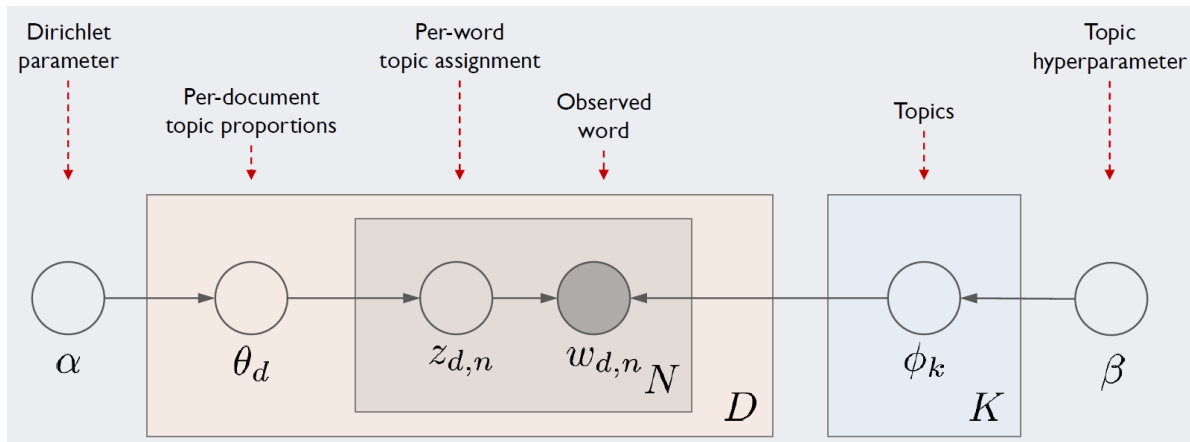
단어의 빈도



문서내 주제들의 분포

LDA의 원리

- 전체 문서들의 집합을 K 개의 주제로 표현이 가능하다고 가정할 때, 각각의 주제에 대한 단어 분포 (β_k)를 디리클레 분포 (β)로 정의
- 개별 주제의 비율을 디리클레 분포로 정의한 후, 각각의 단어에 대해 주제 비율로 주제를 배정하고, 개별 주제의 단어 분포에서 단어를 배정 (확률적으로 추론하는 기법, inference)



Graphical model representation of LDA

- D = 말뭉치 전체 문서 개수 (number of documents)
 - N = d 번째 문서의 단어 수
 - K = 전체 토픽 수 (사용자가 토픽 개수를 지정, 통계적 기법 또는 직관에 의해 결정)
 - $w_{d,n}$ = d 번째 문서에 등장한 n 번째 단어, **관찰 가능한 변수**
 - $z_{d,n}$ = d 번째 문서에 n 번째 단어가 어떤 토픽에 해당하는지를 설명하는 값
- Θ = the probability distribution of topics in documents (문서별 주제의 비율, 모두의 합 = 1)
 - ϕ = the probability distribution of words in topics (각 토픽에서 어떤 단어들이 얼마나 분포하는지를 나타냄, 모두의 합 = 1)
 - α, β = 하이퍼파라미터 (hyperparameter)
 - α = 문서들의 토픽 분포를 얼마나 밀집되게 할 것인지에 대한 설정 값
 - β = 주제내 단어들의 토픽 분포를 얼마나 밀집되게 할 것인지에 대한 설정 값



경청해 주셔서 감사합니다

