서울대학교 중앙도서관

도서대출 빅데이터 정책연구과제
결과보고서

목차

발간사 4

연구정책과제 개요 6

PART. I
연구과제
네트워크 임베딩을 활용한 도서 및 친구 추천 시스템 설계 10

PART. II
연구과제
서울대학교 중앙도서관 도서 대출 빅데이터 분석 24

PART. III
연구과제
서울대학교 도서관 데이터 분석 40

PART. IV
연구과제
다이나믹 / 반응형 탐색적 자료분석 50

부록 72

발간사

본 연구과제는 우리 도서관이 작년 한해 역점 사업으로 추진했던 빅데이터 기반 지식정보 플랫폼 〈LikeSNU 시스템 구축 사업〉에 앞서 도서관이 보유하고 있는 빅데이터의 활용 가능성을 확인하기 위해 시작되었습니다. 본 연구에서는 데이터를 활용하여 어떻게 하면 학내 구성원들에게 유용한 시스템을 구축할 수 있을지에 대한 방향성 확립과 데이터 활용에 관한 구체적인 방법론에 대한 연구도 함께 이루어졌습니다.

학내/외 최고의 데이터 전문가들이 참여한 본 연구정책 과제에서는 지난 10년 간 서울대에서 이루어졌던 가장 중요한 지적 활동 중의 하나인 단행본 대출 기록 200만 건을 가지고 탐색적 데이터 분석, 머신 러닝, 데이터 모델링 및 시각화 등을 통해 데이터를 여러 각도에서 살펴보고 활용 방안을 모색하였습니다. 그리고 흙 속에서 진주를 발견하듯 그 결과물로 참여한 연구원분들 모두 전문 지식과 혜안을 가지고 우리 도서관 대출 데이터에 관한 중요한 인사이트를 제시해주셨습니다. 그 연구 내용을 정리하여 본 책자로 출간하게 되어 뜻깊게 생각합니다.

본 연구과제가 시작점이 되어 앞으로 서울대가 생산하고 있는 가장 중요한 생산물인 완성된 지식뿐만 아니라 이를 생산하기 위한 과정에서 이루어지고 있는 모든 지적 활동들을 총 망라하여 지식정보 플랫폼을 구축하겠다는 이정표를 세울 수 있었습니다. 그 결과물로 작년 LikeSNU시스템을 구축하여 서비스 중에 있으며 곧 2차 사업이 시작될 예정입니다.

LikeSNU시스템은 앞으로 계속해서 데이터 확장과 기능 고도화를 통해 교수, 학생, 연구자가 새로운 지식 창조와 교육 및 학습 · 연구 전 과정에 활용할 수 있는 선도적 시스템이 될 것이라 기대합니다. 뿐만 아니라 서울대학교를 넘어 국내/외 연구자와 일반인들에게도 서울대의 지식 빅데이터를 공유하여 대국민 지식 공헌에 이바지하는 가치 있는 시스템이 되리라 확신합니다.

본 연구정책과제는 서울대학교 본부의 적극적인 관심과 지원 덕분에 가능하게 되었습니다. 오세정 전 총장님과 이현숙 전 연구처장님께 감사의 말씀을 드립니다. 훌륭한 연구결과를 제출해주신 참여 연구원 분들께도 깊이 감사드립니다. 서울대 중앙도서관이 빅데이터를 활용한 새로운 지식 창조의 교두보로 거듭나는 데에 본 연구가 도움이 되길 바라며, 앞으로도 학내외 많은 분들의 관심과 지원을 기대합니다. 감사합니다.

2023. 5.

서울대학교 중앙도서관장

장 덕 진

연구정책과제 개요

연구과제명

서울대학교 중앙도서관 대출 빅데이터 분석을 통한 연구지원 활용 방안 연구

참여연구원

책임연구원	서울대학교 중앙도서관장 장덕진(서울대학교 사회학과 교수)
공동연구원	서울대학교 정보화본부장 김홍기(서울대학교 치과대학 교수)
	서울대학교 대학혁신센터 데이터통합관리부장 이준환(서울대학교 사회학과 교수)
	서울대학교 자연과학대학 조교수 박건웅
	한국과학기술원 인문사회융합과학대학 조교수 김란우
	숭실대학교 스마트시스템소프트웨어학과 조교수 윤진혁

연구목표 및 내용

연구 목표

○ 서울대학교 학내에서 문헌을 활용한 연구동향을 분석하여 새로운 연구지원 서비스 개발 방법론 모색

연구 내용

○ 서울대학교 도서대출 데이터 200만 건(2015. 8. ~ 2022. 5.)을 분석하여 연구동향을 파악하고 이를 활용한 연구지원 방법론 모색

○ 중앙도서관이 전통적인 도서관 기능을 넘어 데이터 기반의 지식·정보 플랫폼으로 자리매김 하기 위한 기반을 마련

○ 데이터 분석을 통해 서울대인들이 독서를 통해 만들어온 지식지도(Knowledge Map)를 파악

○ [사람 X 대출]로 이루어진 데이터 분석을 통해 독서로 연결된 서울대 구성원의 커뮤니티를 파악, 이 커뮤니티는 지적 관심의 유사성, 이질성, 연결성 등 다양한 기준으로 찾아질 수 있음

○ [사람 X 대출]로 이루어진 데이터 분석을 통해 서울대 구성원으로 연결된 도서들의 커뮤니티를 파악, 실제 사람들의 독서를 통해 지식의 분야가 어떻게 나뉘는지, 어느 분야와 다른 어느 분야가 가깝거나 먼지, 융합을 위해 서로 다른 분야를 연결해주는 분야는 무엇이고 구체적으로 어떤 도서인지, 학생들에게 무슨 책을 읽혀야 할지 파악

○ 분석결과가 즉시 도서관 서비스로 활용될 수 있도록 분석용 코드 등 실질적인 도구 개발

연구방법

○ 다양한 통계 모델과 데이터 분석 방법을 활용하여 도서 대출 빅데이터로부터 인사이트를 발견하고, 문헌을 활용한 연구 동향 사례 검토와 시사점 도출

○ 외부 공개 데이터를 접목시켜 문헌 활용도를 측정하고 새롭고 창의적인 관점으로 연구동향에 대한 분석을 시도하여 이를 활용한 연구지원 방안 모색

PART. I
연구과제

네트워크 임베딩을 활용한 도서 및 친구 추천 시스템 설계

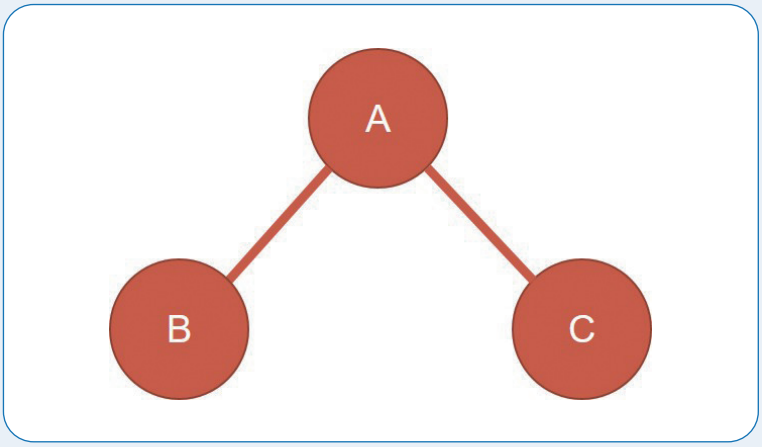
연구배경
연구내용
연구결과
향후계획

네트워크 임베딩을 활용한 도서 및 친구 추천 시스템 설계

숭실대학교 스마트시스템소프트웨어학과
윤진혁

연구과제

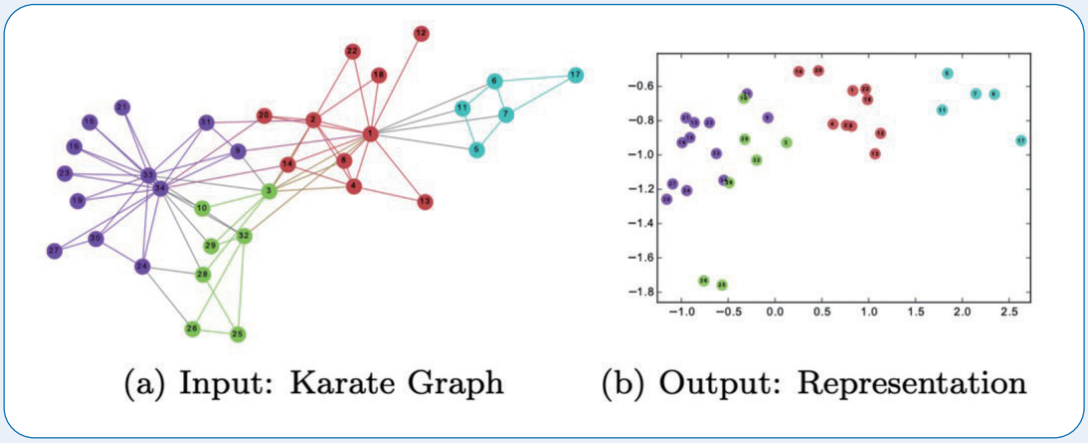
네트워크 구조는 추상적이고 복잡한 관계와 상호작용을 데이터로 표현하는 데에 적절하다. 또한, 복잡한 문제를 더 간단한 표현으로 단순화하고 반대로 관계의 종류를 정의하여 다양한 요소들의 상호작용을 컴퓨터가 처리 가능한 형태로 만들 수 있다. 하지만 네트워크의 경우 유클리드 공간으로 표현할 수 없으며 무한대의 차원을 가지는 프랙털 구조를 가진다.



[그림 1] 네트워크 구조

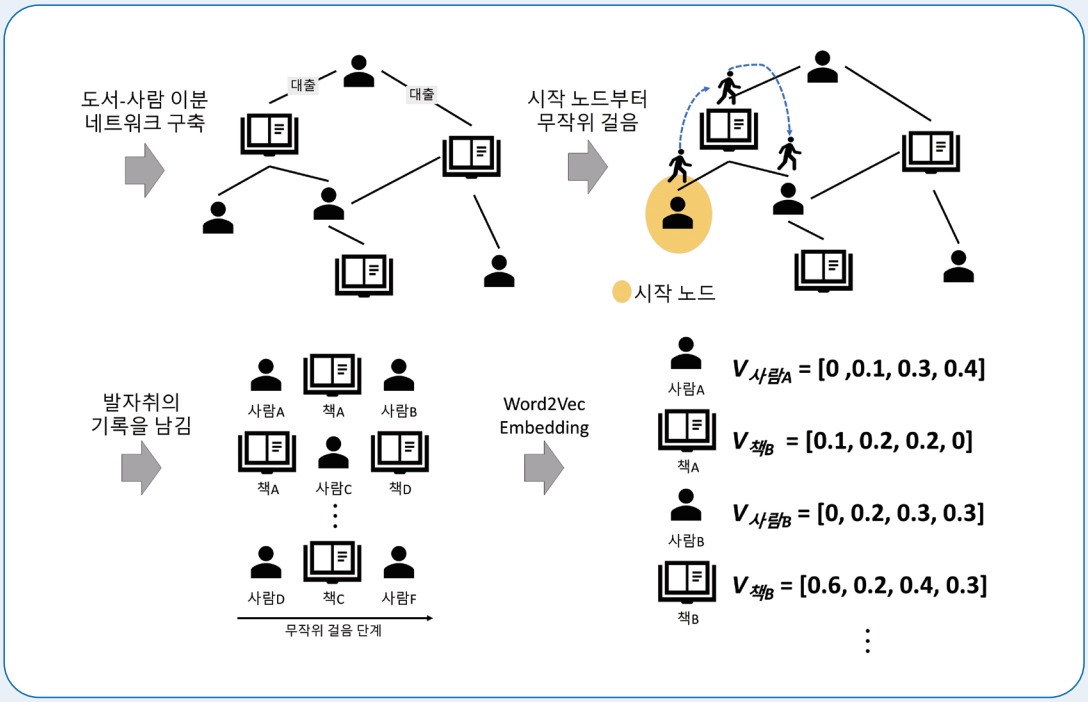
또한 네트워크의 경우 직접적인 연결이 없는 경우에 대한 관계 추정이 어렵고 무관계함을 가정하나, 실제로는 샘플링 문제 등으로 관측되지 않은 링크나 아직 만들어지지 않았으나 차후에 연결이 생길 가능성이 있는 노드쌍이 존재할 수 있다. 네트워크의 희소성으로 인하여 소수의 노드만 연결이 많이 존재하고, 대부분의 노드는 실제 정보가 적어 올바른 정보를 유추하는 것이 어려울 때가 많다.

이러한 단점을 보완하고자 Graph Neural Network(GNN)과 Graph Embedding의 개념이 도입되었다. 한 예로 Recurrent GNN은 한 노드의 feature와 그 이웃한 노드와 상호 간 링크의 현재 feature를 정의하고 이를 재귀적으로 업데이트하여 규칙성이 없는 Graph를 벡터 공간상에 표현하는 형태로 학습한다.



[그림 2] Graph Embedding

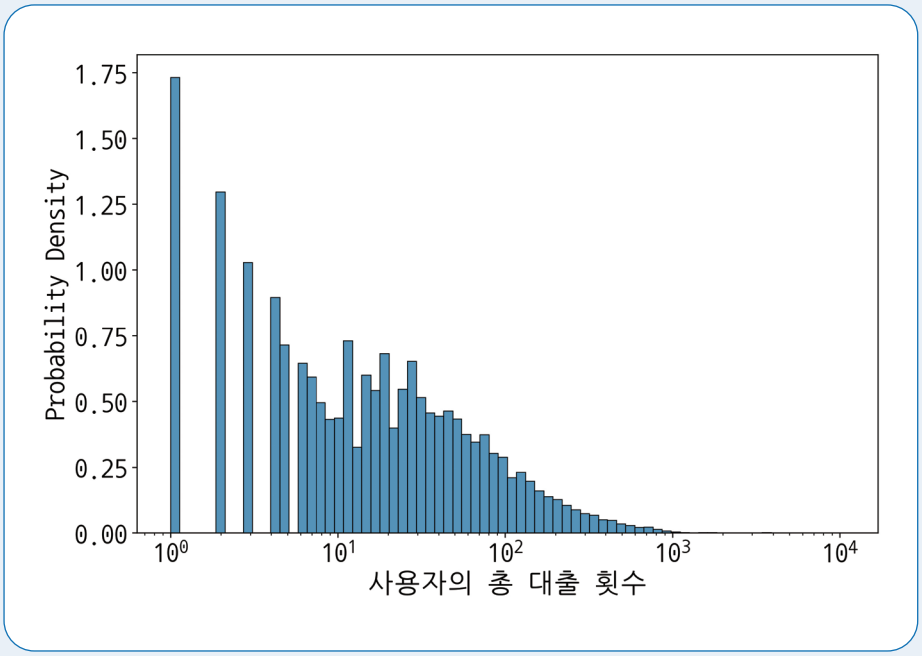
본 연구에서는 도서관 대출 데이터를 활용하여 이분(도서-사람) 혹은 삼분(도서-사람-강의) 네트워크를 구축하고, 이를 Random Walker를 사용한 방식으로 Embedding하여(그림 3) 도서관 서비스에 활용할 수 있는 방법론을 탐색하였다.



[그림 3] 도서-사람 이분 네트워크의 구조와 임베딩 방법

연구내용

먼저 실제 대출 기록에서 희소성 문제가 발생하는지, 두꺼운 꼬리 분포가 나타나는지를 확인하기 위하여 사용자별 대출 기록의 수 분포를 확인해 보았다. 대출 기록이 있는 사람 중 최소의 대출 횟수는 1회였으며, 최대 대출 사용자의 대출 건수는 10,613건으로 나타났으며 1인당 평균 4.38권을 대출하였고 표준편차는 126.11권으로 나타났다. 또한 대출 기록이 있는 책은 총 420,915권 (mms_id기준) 이었다. 대출 건수의 분포는 [그림 4]와 같이 나타났으며 꼬리가 두꺼운 형태의 분포가 나타났으나 그 형태가 멱함수 꼴을 나타내지는 않는 것으로 추정된다.



[그림 4] 도서관 이용자 1인의 도서 대출 수의 분포

다음으로 [그림 3]의 과정으로 Embedding한 사용자 및 도서의 vector가 실제를 잘 반영하는지에 대한 분석을 진행하였다. word2vec model의 경우 embedding에 사용하는 parameter에 따라 결과가 다르게 나타날 수 있다. 본 연구에서는 sequence의 길이를 30으로, 한 노드당 20번의 random walker를 출발시켜 50차원의 벡터 공간에 embedding하였다. embedding을 위해서 길이 5의 word window를 사용하였고, skip-gram with negative sampling 모델로 계산하였으며 5번의 epoch 수행하여 최종 모델을 도출하였다. 재현성을 위해 embedding을 위한 무작위 함수의 seed는 4,090을 사용하였다.

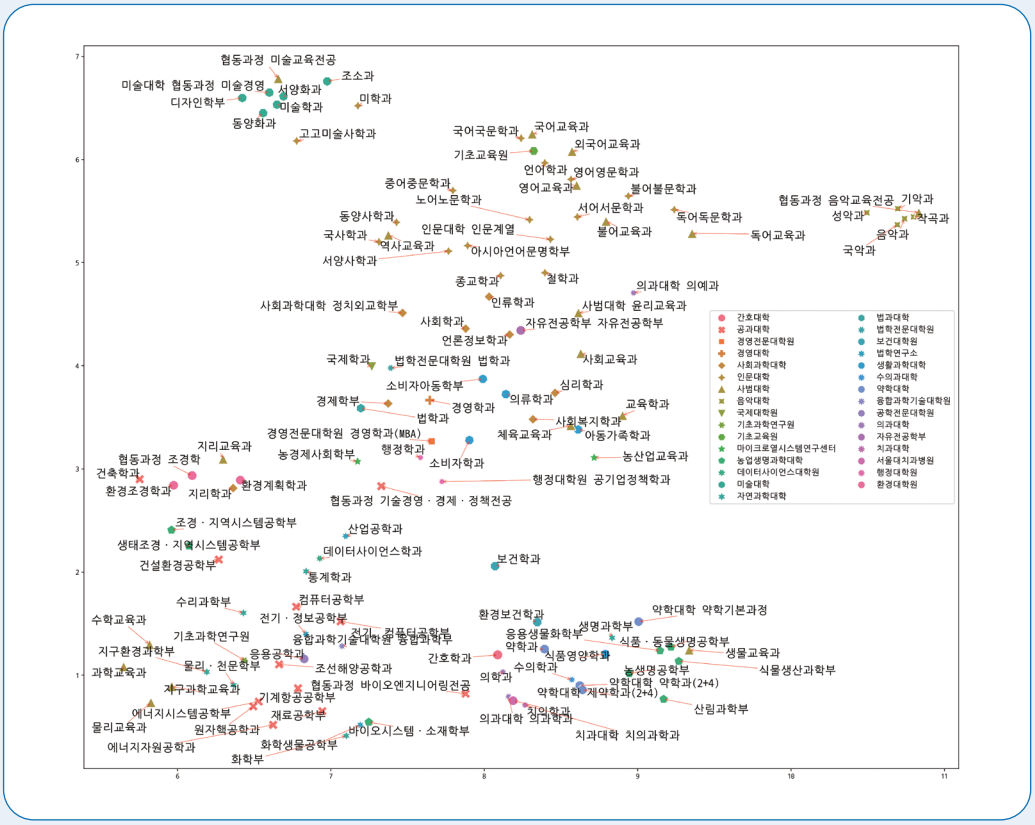
[그림 5]의 경우 위에서 구축한 이분 네트워크에서 구한 임베딩 벡터의 학과별 유사도를 2차원 공간에 표시하였다. 학과 벡터의 경우 학과에 속한 모든 학생의 벡터를 길이 1로 L2 정규화를 수행한 이후, 이렇게 정규화한 벡터를 평균하여 구하였다. 2차원 공간에 표시하기 위해 이렇게 구한 평균 벡터를 Uniform Manifold Approximation and

Projection (UMAP) 이라는 manifold learning 기법을 사용하여 차원 축소하였다.

[그림 5]에서 먼저 유사 학과가 근처에 위치하는 경향성을 확인할 수 있었다. 예를 들어 우측 상단에는 음악계통학과(협동과정 음악교육전공, 기악과, 성악과, 작곡과, 국악과) 등이 위치하였으며 좌측 상단에는 미술계통학과(협동과정 미술교육전공, 조소과, 미학과 등)가 위치함을 확인할 수 있었다.

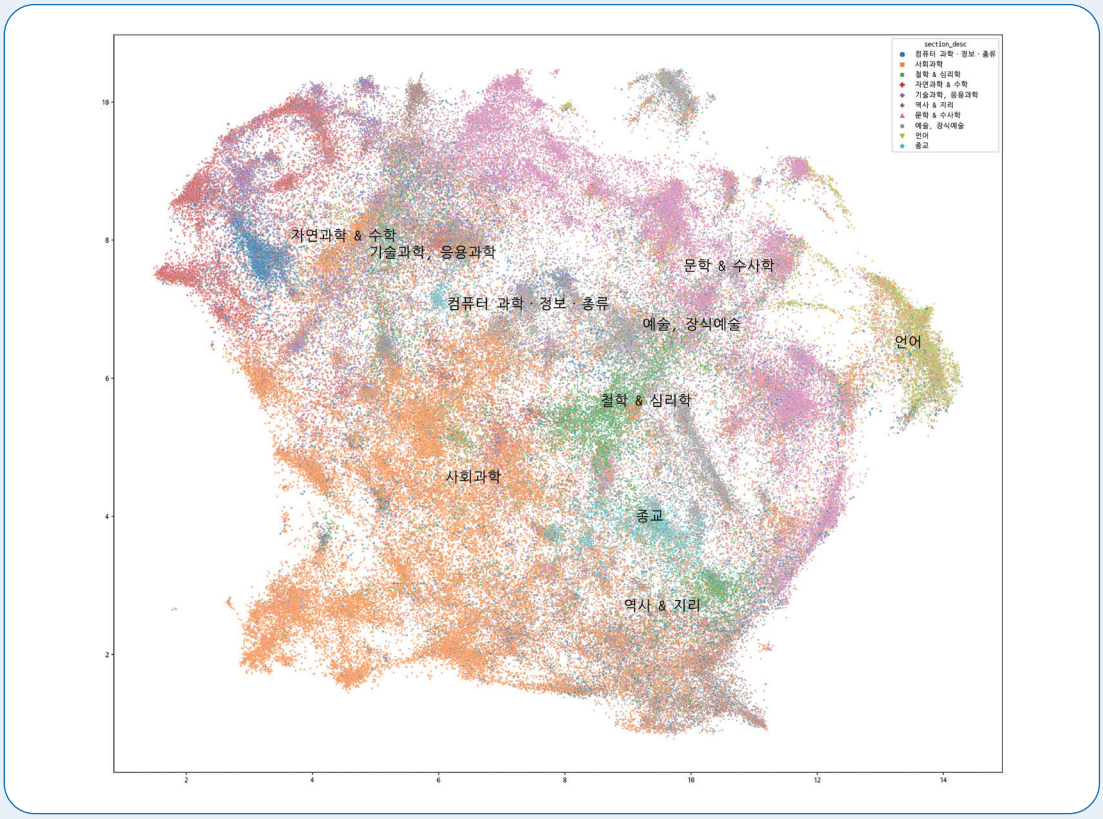
학과의 소속 단과대학보다는 유사 전공인 학과가 근처에 위치함을 확인할 수 있었다. 예를 들어 환경 및 건축 계통 학과(지리학과, 지리교육과, 환경계획학과, 환경조경학과, 건축학과 등)는 단과대학에 무관하게 모두 좌측의 유사 위치에 모여 있었으며, 유사하게 생명 및 의약학 관련 학과들(수의학과, 간호학과, 학과, 생명과학, 생물교육, 환경보건 등)은 모두 우측 하단에 위치한 것을 확인할 수 있었다.

교육 계통 학과들은 서로 모여있지 않고 전공이 유사한 학과와 비슷한 위치에 놓여있음을 확인할 수 있었다. 예를 들어 지구과학교육과는 지구환경과학부와, 물리교육과는 물리천문학부와 유사한 위치에 있었는데, 이를 통해 도서대출 패턴은 교육계통 학과와 전공이 유사한 학과가 비슷함을 확인할 수 있었다.



[그림 5] 이분 네트워크의 학과별 유사도. 학과의 유사도는 학과 학생의 normalized vector의 평균값을 사용하였으며, 시각화에는 데이터에 100명 이상 존재하는 학과만 사용하였다.

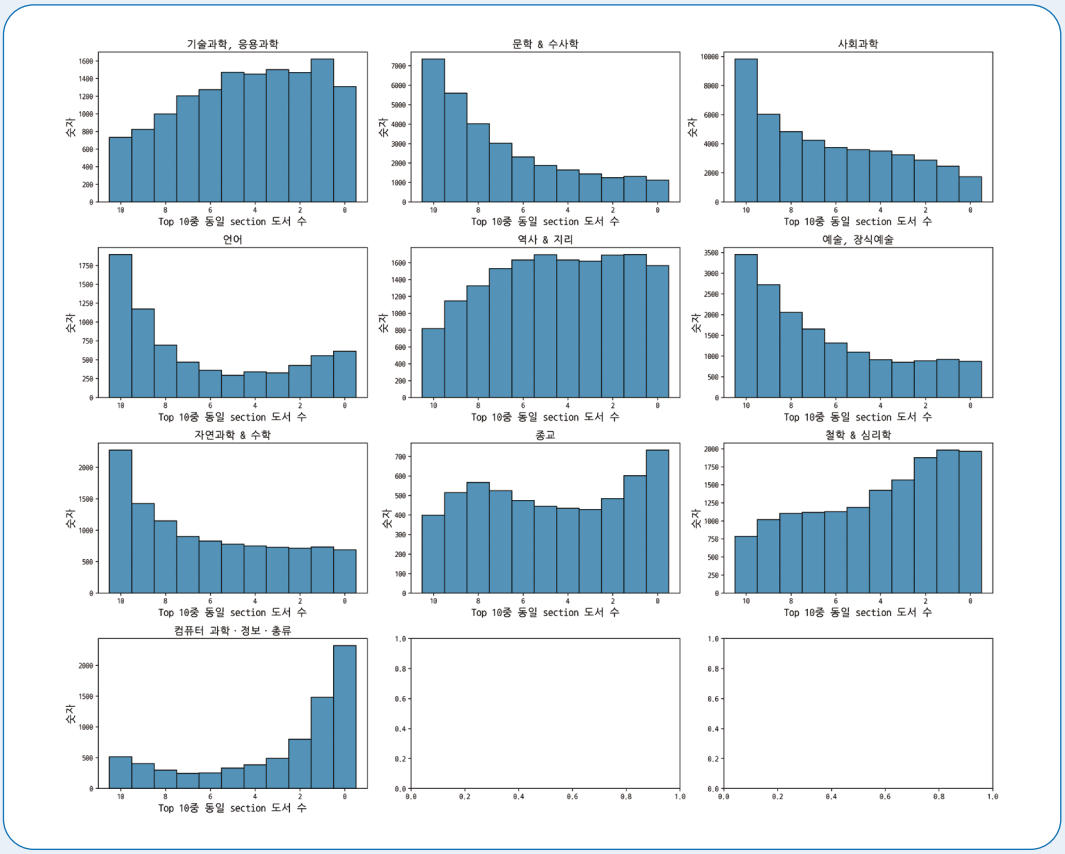
[그림 6]의 경우 위에서 구축한 이분 네트워크에서 구한 도서간 임베딩 벡터 유사도를 시각화 하였다. 역시나 2차원 공간에 표현하기 위하여 UMAP을 사용하였다. 각 도서의 색상은 도서관의 카테고리 분류 코드중 section code 10가지를 사용하였다. 또한 텍스트로 표기된 분야는 각 분야 도서 전체의 평균 벡터 위치에 두었는데, 개략적인 각 분야의 중심점 위치를 뜻한다. 아래 그림에서 유사한 섹션의 도서가 모여있는 것을 확인할 수 있는데, 이는 개인의 도서 대출 패턴은 비교적 일관적이며 본인이 관심있는 분야의 도서를 지속적으로 대출한다는 점을 뜻한다.



[그림 6] 이분 네트워크에서 구성한 도서별 대출 패턴 유사도

도서와 사용자 모두 고유한 vector를 가지기 때문에 이러한 vector를 기반으로 한 추천 시스템을 구축할 수 있다. 가장 단순하게는 자신과 가장 비슷한 vector를 가지는 아이템을 추천하는 방식이 있다.

[그림 7]에서는 cosine similarity를 기반으로 각 도서별 10개씩의 연관 도서를 추천받아 이들이 어떤 section에 분류되어 있는지 통계를 보여준다. 재미있게도 2가지 다른 패턴으로 구분되는 것을 관찰할 수 있는데, 문학&수사학, 사회과학, 언어, 예술&장식예술, 자연과학&수학의 경우 동일한 주제의 책을 더 많이 추천해주는 경향을 확인할 수 있었으나, 반대로 기술과학&응용과학, 역사&지리, 종교, 철학&심리학, 컴퓨터과학 관련 서적은 다른 분야의 책을 더 많이 추천해주는 것을 확인할 수 있었다.



[그림 7] 이분 네트워크에서 도서 1개를 기반으로 타 도서를 추천하였을 때 동일 section의 도서가 추천되는 경우의 수

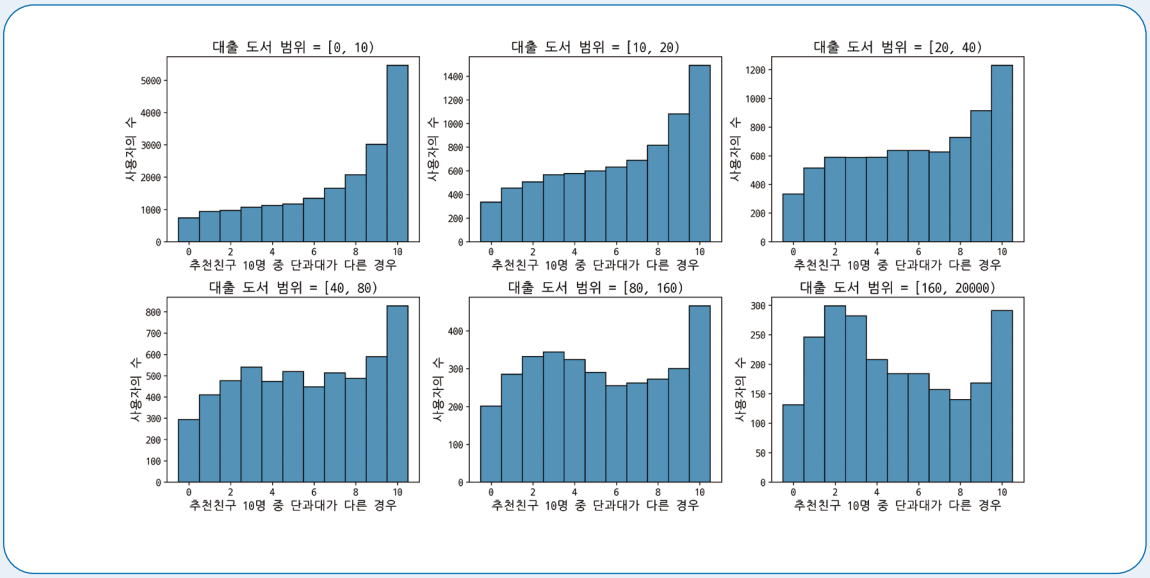
다음으로 추천시스템이 어떠한 사용자들에게 가장 가치가 높은지에 대해서 판단하기 위하여 추천시스템에서 추천해 준 도서 중 실제로 대출한 기록이 없는 도서가 몇 권이나 있나를 확인해 보았다.

[그림 8]에서 확인할 수 있듯, 대출한 도서의 수가 늘어날수록 추천 도서 중 더 많은 수의 도서가 기존에 대출 기록이 있는 도서로 나타났다. 하지만 80권 이상 160권 미만을 대출한 도서관을 자주 사용하는 사람의 경우에도 추천 도서 중 대출 기록이 없는 도서가 한 개라도 나타나는 경우가 대출 기록이 있는 도서들만 추천하는 경우보다 많았다. 160권 이상을 대출한 최상위 사용자의 경우 거의 모든 도서가 기존 대출 기록이 있는 도서였으며, 극소수의 미대출 도서만 추천된다는 것을 확인할 수 있었다.

추천 도서의 수를 늘린다면 이러한 고빈도 사용자에게도 도서를 추천할 수 있으나 전체적인 유사도가 낮아져서 추천의 질이 떨어질 것을 추측할 수 있다.

위와 유사한 방식으로 cosine similarity 기반으로 본인의 독서 대출 취향과 가장 유사한 친구를 추천해주는 시스템도 설계가 가능하다. [그림 8]은 가장 유사도가 높은 10명의 사용자를 추천해주는 경우 동일 단과대를 추천받는 경

우와 다른 단과대를 추천받는 경우를 사용자의 도서 대출 수에 따라 그린 그림이다. 도서의 대출 수가 늘어날수록 단과대가 같은 학생을 추천하는 비율이 높아지는 것을 알 수 있다. 도서 대출 수가 많은 경우 같은 단과대를 추천하는 비율이 올라가는 것은 도서관을 전공 관련 서적을 대출하는데에 주로 사용하는 학생의 비중이 높기 때문으로 추측된다. 동시에 추천받은 10명 모두가 단과대가 다른 경우의 비율은 일정 수준 이상으로 유지되는 경향도 보이는데, 이는 일정 비율의 사용자는 전공서적보다는 교양서적이나 기타 서적을 대출하기 때문으로 추정된다. 다만 위의 추정을 뒷받침하기 위해서는 도서가 교양서인지 전공서인지를 구분해야 하는데, 이는 현재 데이터에는 구분되어 있지 않아 파악이 어렵다.



[그림 8] 도서 대출 수에 따라 이분 네트워크 모델에서 추천한 친구 10인 중 단과대가 같은 경우와 다른 경우. 도서 대출의 수가 늘어날수록 같은 단과대를 추천하는 비율이 늘어나는 것을 볼 수 있다.

연구결과

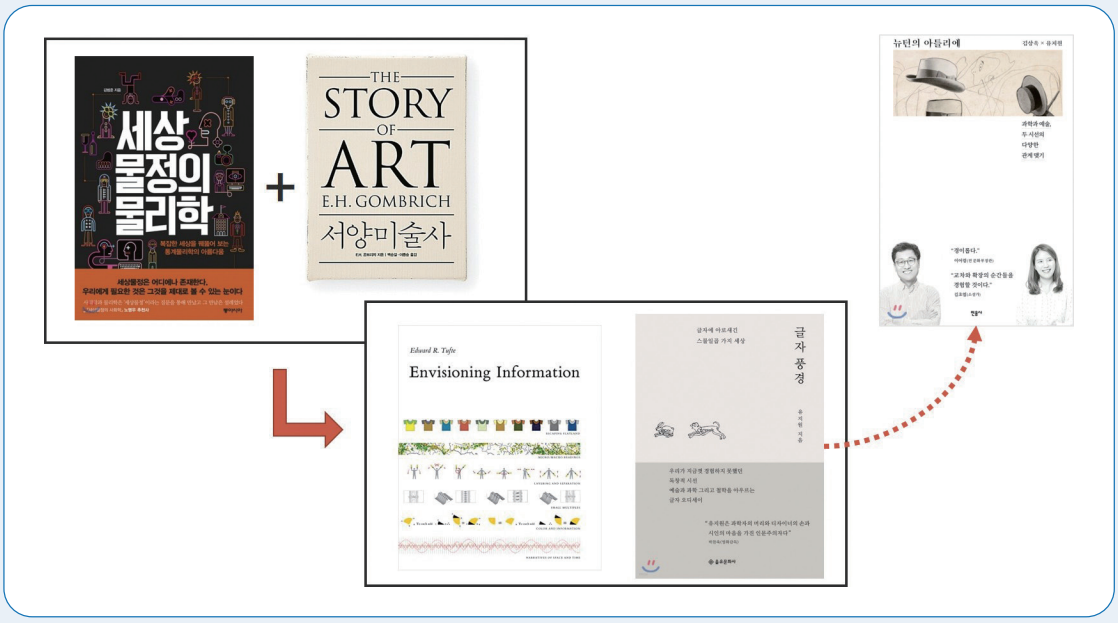
제안분석 모델 1 - 도서 대출 기록을 활용한 추천 도서 시스템

위에서 embedding한 데이터를 응용하면 도서 대출 기록이나 본인이 입력한 책을 기반으로 도서를 추천해 주는 시스템을 구축할 수 있다. 먼저 입력한 책들을 바탕으로 도서를 추천하는 시스템을 제작해 보았다. 입력한 도서들의 벡터의 길이를 동일하게 만드는 L2 정규화를 수행한 후 평균 벡터를 구하였고, 이 평균 벡터와 가장 유사한 방향성을 가지는 다른 책을 추천할 수 있다. 이 시스템이 제대로 작동하는지에 대한 예시를 찾기 위하여 복잡계 물리학의 대표적인 도서 3권을 입력으로 받아 추천 도서를 찾는 예시를 보았다. (그림 9)

[그림 9] 동일 분야로 입력하는 경우 도서 추천 예시 : 복잡계 네트워크 관련 대표 도서 3가지(세상물정의 물리학, 구글신은 모든 것을 알고 있다. 버스트 : 인간의 행동 속에 숨겨진 법칙)를 입력하는 경우의 추천 도서. 대부분이 복잡계 네트워크를 다루고 있다.

[그림 9]의 예에서 3가지의 유사 도서를 입력하는 경우, A.L. Barabasi, M. Buchanan, S. Strogatz 등이 쓴 복잡계 물리학에 대한 다양한 저서가 주로 추천되는 것을 확인할 수 있었으며 이 방법론으로 도서의 주제를 잘 찾을 수 있음을 확인하였다.

이를 확장하여 전혀 다른 주제의 도서 두 가지를 입력하는 경우도 확인해 보았다. 복잡계 물리학 저서인 〈세상 물정의 물리학〉과 서양미술사의 가장 대표적 교과서인 〈곰브리치 서양미술사〉 두 가지 책을 입력으로 사용한 결과에서는(그림 10) 정보를 효율적이고 심미성있게 시각화하는 방법을 다루는 E. R. Tufte의 〈Envisioning Information〉과 타이포그래피에 대해 다루는 유지원 교수의 〈글자풍경〉 두 권이 가장 높은 유사도를 보였다.



[그림 10] 도서 추천 예시 2: 복잡계 및 통계물리 서적(세상물정의 물리학)과 미술 서적(곰브리치 서양미술사)를 동시에 입력하는 경우의 추천 도서 예시

2위로 추천된 〈글자풍경〉의 작가인 유지원 교수의 경우 미술 전공이나 〈글자풍경〉의 출간 이후 경희대 물리학과의 김상욱 교수와 함께 〈뉴턴의 아틀리에〉 라는 교양 과학 서적을 출간하였다. 이러한 예시를 통해 임베딩을 통한 도서 추천이 어느 정도 사용자 취향에 기반한 도서 추천이 가능하게 함을 확인할 수 있다.

또한 이 모델은 단순히 도서만 입력할 수 있는 것이 아니라 사람의 정보도 같이 입력할 수 있는데(그림 11) 동양사학과 대학원생을 위의 〈곰브리치 서양미술사〉 및 〈세상물정의 물리학과〉 동시에 입력하는 예시를 보면 단순히 책을 입력하는 것과 다르게 사용자의 전공에 기반하여 역사에 관련된 도서를 추천하는 비율이 크게 늘었음을 확인할 수 있다. 이러한 도서-사용자 동시 임베딩을 통하여 도서를 추천하는 경우 텍스트 등을 통해 단순히 도서의 내용 기반으로 추천하는 것 보다 사용자의 취향에 더 잘 맞는 추천을 해 줄 가능성이 높으며, 도서관 만족도를 향상시킬 수 있다.

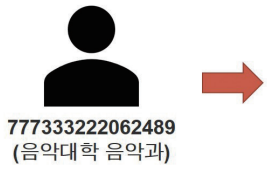
```
# 테스트
# 99391301702591: 세상물정의 물리학
# 99134781502591: 곰브리치 서양미술사
# 777333222026412: 동양사학과 대학원생
input_list = ["99391301702591", "99134781502591", "777333222026412"]
get_most_similar_books_from_multibook(input_list)[["title", "author"]]
```

	title	author
12	조선의 무기 /	강신엽.
8	鹽鐵論 : 소박하면 풍족해지고, 사치하면 기근이 온다 /	Huan, Kuan,
6	(시장으로 나간) 조선백자 : 분원과 사기장의 마지막 이야기 /	박은숙,
11	발해 대외관계사 자료 연구 /	장재진
10	고대 동북아시아 교통사 /	Wang, Mianhou
1	(옛날 이야기처럼 재미있는) 곰브리치 세계사 /	Gombrich, E. H.
0	군주론 /	Machiavelli, Niccol?,
13	고려의 북진정책사 /	장학근.
9	(신편) 발해국지장편 /	Jin, Yufu,
4	고려, 북진을 꿈꾸다 : 고구려 영토 회복의 꿈과 500년 고려전쟁사 /	정해은,

[그림 11] 도서 추천 예시3: 복잡계 및 통계물리 서적(세상물정의 물리학)과 미술 서적(곰브리치 서양미술사), 동양사학과 대학원생을 입력하는 경우의 추천 도서 예시. 위의 단순 추천과 책 목록이 달라져 역사학 관련 도서가 증가하였음을 볼 수 있다.

제안분석 모델 2 - 도서 대출 기록을 활용한 독서 모임 친구 추천 시스템

임베딩 모델에서는 도서 이외에도 사용자도 모두 벡터가 주어져 있으므로 사용자들의 독서 취향에 맞는 친구를 추천해주는 시스템도 구축할 수 있다. [그림 12]는 임의로 선정된 음악대학 대학원생과 가장 유사한 친구 10인을 추천하는 경우를 예시로 들었다. 이런 경우 유사한 전공의 사용자가 주로 추천이 되나, 공과대학이나 사범대학 소속의 대학원생 또한 추천이 됨을 확인할 수 있다. 학과 기준이 아닌 독서 취향을 통해 친구를 추천하므로 단순히 전공을 넘어 본인과 관심사를 공유하는 친구를 추천해줄 수 있다.



	user_id	group	college	department
14	777333222108498	대학원생	음악대학	음악과
3	777333222056278	대학원생	사범대학	협동과정 음악교육전공
15	777333222114036	학부생	음악대학	국악과
10	777333222075609	대학원생	음악대학	음악과
12	777333222077357	대학원생	공과대학	전기·정보공학부
17	777333222119068	대학원생	음악대학	음악과
5	777333222062407	대학원생	음악대학	음악과
19	777333222131976	학부생	음악대학	국악과
22	777333222190484	강사	음악대학	국악과
18	777333222122354	학부생	음악대학	국악과

[그림 12] 친구 추천 예시: 음악대학 대학원생을 입력하는 경우 유사 학생들이 주로 검색된다.

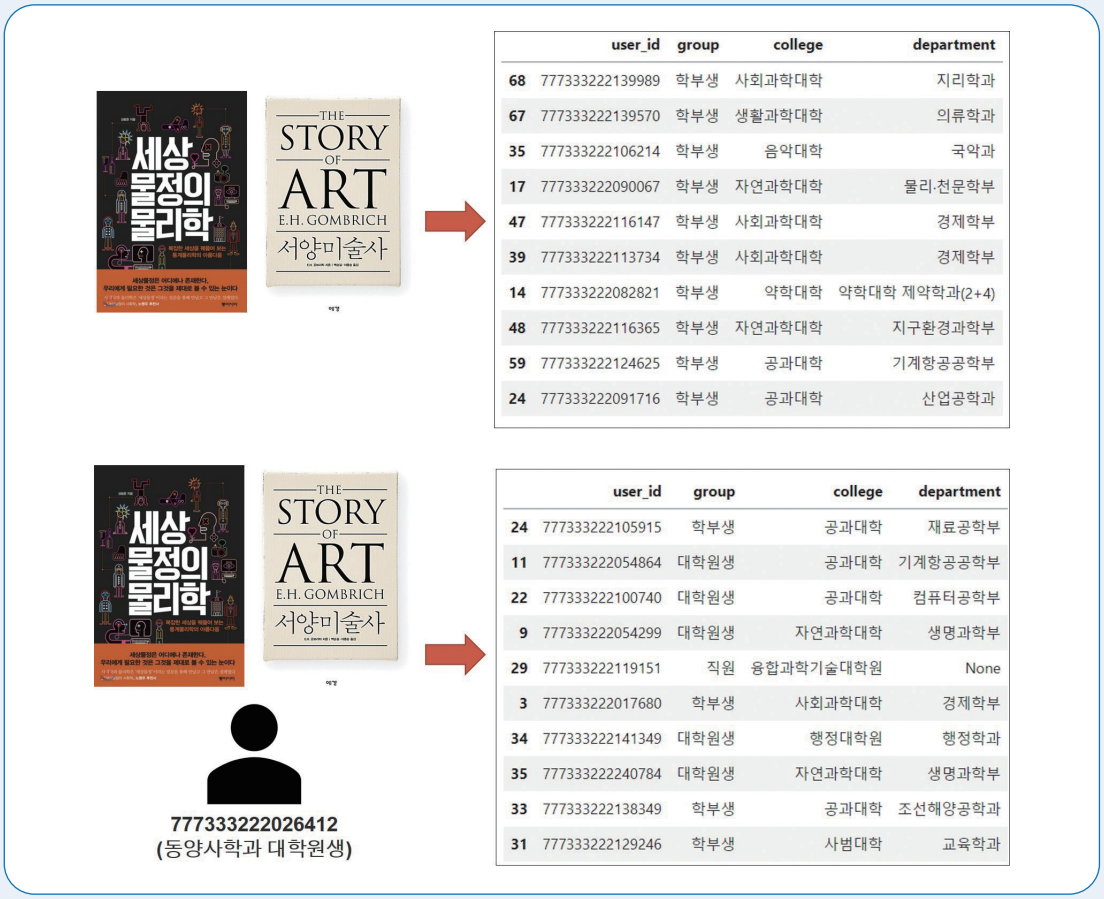
또한 이 모델은 도서의 정보를 입력하여 도서와 관련있는 사용자를 추천해줄 수도 있다. 예를 들어 도서관에서 분야별 추천 도서 목록을 선정하여 관심있는 사용자에게 발송하는 경우 등에 사용할 수 있다. 제안분석모델 1에서 사용된 복잡계 물리학 관련 3가지 도서를 입력하는 경우를 다시 고려해 보자(그림 13). 3개의 도서는 모두 물리학자가 집필한 물리학 관련 도서로 볼 수 있다. 하지만 이러한 도서에 관심을 보일 것으로 예상되는 사람들은 물리학과나 자연과학 대학 이외에도 사회학, 인문대학, 융합과학기술대학원 등 다양한 전공을 가지고 있음을 확인할 수 있다. 이는 사회학이나 경제학 등에서도 복잡계 네트워크 관련 기법을 사용하는 연구가 다수 존재하므로 이러한 분야적 관심도에 기반한 것이라 추정된다.

```
# 테스트
# 99391301702591: 세상물정의 물리학
# 99127554202591: 구글신은 모든 것을 알고 있다
# 99172392802591: 버스트 : 인간의 행동 속에 숨겨진 법칙
input_list = ["99391301702591", "99127554202591", "99172392802591"]
get_most_similar_users_from_multiitem(input_list)[["user_id", "group", "college", "department"]]
```

	user_id	group	college	department
19	777333222138625	학부생	자유전공학부	자유전공학부 자유전공학부
12	777333222090067	학부생	자연과학대학	물리·천문학부
16	777333222116147	학부생	사회과학대학	경제학부
18	777333222138028	학부생	자연과학대학	물리·천문학부
26	777333222190626	대학원생	사범대학	사회교육과
11	777333222086661	대학원생	공과대학	재료공학부
14	777333222109767	대학원생	융합과학기술대학원	None
15	777333222113734	학부생	사회과학대학	경제학부
22	777333222173944	회원제	인문대학	협동과정 인지과학전공
13	777333222099700	학부생	공과대학	재료공학부

[그림 13] 친구 추천 예시 2: 복잡계 네트워크 관련 대표 도서 3가지(세상물정의 물리학, 구글신은 모든 것을 알고 있다, 버스트 : 인간의 행동 속에 숨겨진 법칙)를 입력하는 경우 이공계생 뿐 아니라 다양한 분야의 친구를 추천해준다.

위의 두 예시에서 사용된 단순 쿼리 이외에도 사용자-도서 복합 쿼리를 사용하여 관심도 있는 사람을 추천해줄 수 있다(그림 14). 예를 들어 독서 모임의 리더가 있고, 리더가 추천하는 독서 목록을 동시에 가지고 있는 경우 이 두 가지를 모두 고려하여 리더의 독서 취향 및 리더가 새로 읽고자 하는 책에 모두 관심 있는 사용자를 추천하는 것을 수행할 수 있다.



	user_id	group	college	department
68	777333222139989	학부생	사회과학대학	지리학과
67	777333222139570	학부생	생활과학대학	의류학과
35	777333222106214	학부생	음악대학	국악과
17	777333222090067	학부생	자연과학대학	물리·천문학부
47	777333222116147	학부생	사회과학대학	경제학부
39	777333222113734	학부생	사회과학대학	경제학부
14	777333222082821	학부생	약학대학	약학대학 제약학과(2+4)
48	777333222116365	학부생	자연과학대학	지구환경과학부
59	777333222124625	학부생	공과대학	기계항공공학부
24	777333222091716	학부생	공과대학	산업공학과

	user_id	group	college	department
24	777333222105915	학부생	공과대학	재료공학부
11	777333222054864	대학원생	공과대학	기계항공공학부
22	777333222100740	대학원생	공과대학	컴퓨터공학부
9	777333222054299	대학원생	자연과학대학	생명과학부
29	777333222119151	직원	융합과학기술대학원	None
3	777333222017680	학부생	사회과학대학	경제학부
34	777333222141349	대학원생	행정대학원	행정학과
35	777333222240784	대학원생	자연과학대학	생명과학부
33	777333222138349	학부생	공과대학	조선해양공학과
31	777333222129246	학부생	사범대학	교육학과

[그림 14] 친구 추천 예시 3: 복잡계 및 통계물리 서적(세상물정의 물리학)과 미술 서적(곰브리치 서양미술사)의 도서 조합을 입력하는 경우와 동양사학과 대학원생을 함께 입력하는 경우의 추천 도서 예시 비교. 아래 결과의 경우 위의 단순 추천과 다른 사람을 추천하는 것을 알 수 있다.

제안이유

코로나19 이후 비대면이 일상이 되며 도서관 서비스에 대한 주된 수요가 단순한 도서 대출에서 다양한 정보 서비스로 개편되고 있다. 기존에는 동아리나 학생회 등 다양한 경로로 본인과 유사한 취향을 가진 사람을 찾는 방법이 있었으나, 코로나19 이후 비대면 시대에 입학한 2020학번 이후의 학생들은 사회적 거리두기 등으로 위의 활동이 거의 불가능하였다. 도서관에서 취향에 맞는 도서를 추천하고, 더 나아가 동일 취향을 공유하는 독서 모임 등을 구성한다면 비대면 시대의 새로운 친구 찾기 경로가 될 수 있다. 독서 모임의 경우 서로 독립적으로 도서를 읽은 이후 서로 의

견을 교환하는 짧은 모임이 이어지는 형태로, 취향이 잘 맞는 친구를 구한다면 대면/비대면에 무관하게 높은 친밀도를 형성할 수 있다는 장점이 있다. 위에서 제시한 추천 시스템 등을 활용하면 대면/비대면 독서 모임 등을 구성하여 도서관이 단순히 도서를 대여하는 수동적 역할 뿐 아니라 비대면 시기에 교내의 결속력을 다지는 적극적인 매개체 역할을 할 수 있을 것으로 보인다.

향후계획

- **도서관 홈페이지의 추천 도서 서비스 개편에 활용**
- **메일링 리스트 등을 통한 사용자 추천 도서 발송**
- **1학년 논술 등 기초 과목 등과 연계하여 학생의 취향에 맞는 도서 추천으로 도서관 이용률 및 만족도 향상**
- **단순 대출량 기반이 아닌 고도화된 서비스를 통해 도서관의 도서 구매시 기반 데이터로 활용: 학과별 관심 도서 분야 등을 추출하여 도서 구매시 선 반영**

PART. II
연구과제

서울대학교 중앙도서관
도서 대출 빅데이터 분석

연구과제
연구내용
연구결과
향후계획

서울대학교 중앙도서관 도서 대출 빅데이터 분석

서울대학교 대학혁신센터 데이터통합관리부
사회학과 교수 **이준환**

연구과제

연구의 필요성 및 목적

- 도서관 대출 빅데이터를 이용하여 효율적인 도서구입을 위한 계획 수립을 지원하여 실제 이용자들의 만족도를 제고할 필요가 있음
- 빅데이터를 이용하여 도서 주제별 장서 현황과 이용도를 분석하고 도서관의 주 이용자인 학생 이용자들의 주제별, 도서연령별 장서 이용 현황을 분석하여 서울대학교 중앙도서관의 효율적인 도서구매계획 수립을 지원하기 위한 데이터 활용방안 모색
- 이를 통해 궁극적으로는 장서구성의 질적 수준을 높이고 이용자들의 만족도를 제고할 수 있을 것으로 기대함

연구의 필요성 및 목적

- 서울대학교 중앙도서관 소장자료 및 대출이력 데이터 이용
- 서울대학교 도서관 분류체계 기준 (대분류) (v_book_refine 데이터 이용)
- 분석대상: 학생이용자(학부생 및 대학원생 전체)
  - 28,645명 (2016-2021년)
- 전공계열: 학교 전공계열 구분 기준에 따라 인문사회계열, 자연과학계열, 공학계열, 예체능계열, 의학계열 5개 그룹으로 구분 (진행과정, 단과대, 학과 정보 활용)
  (계열 구분에 있어 간호대학은 이용도서 특성상 의학계열로 분류함)
  - 인문사회계열 (14,001명), 자연과학계열(6,050명), 공학계열(5,847명), 예체능계열(1,463명), 의학계열(1,284명)

분석 내용

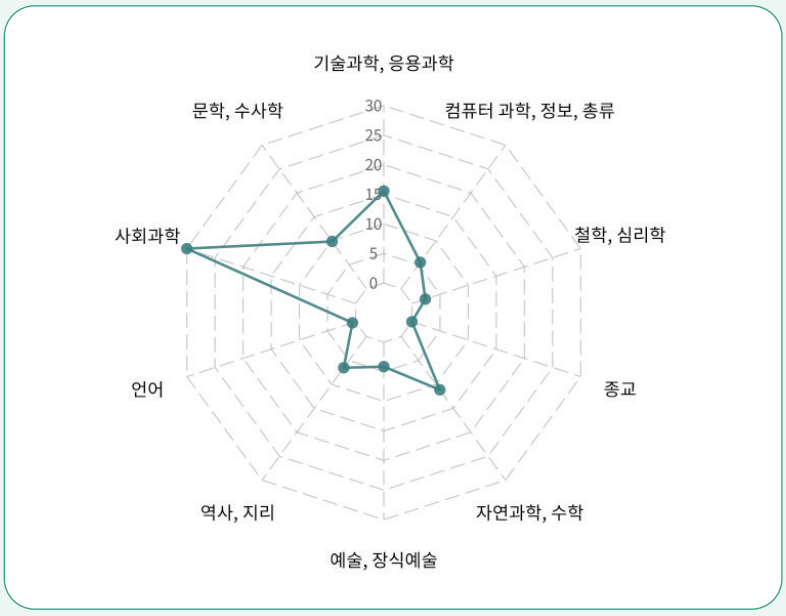
- 중앙도서관 주제별 장서 이용도 분석
- 학생 전공계열에 따른 주제별 이용도 분석
- 학생 전공계열에 따른 도서연령별 자료대출 현황 분석

연구내용

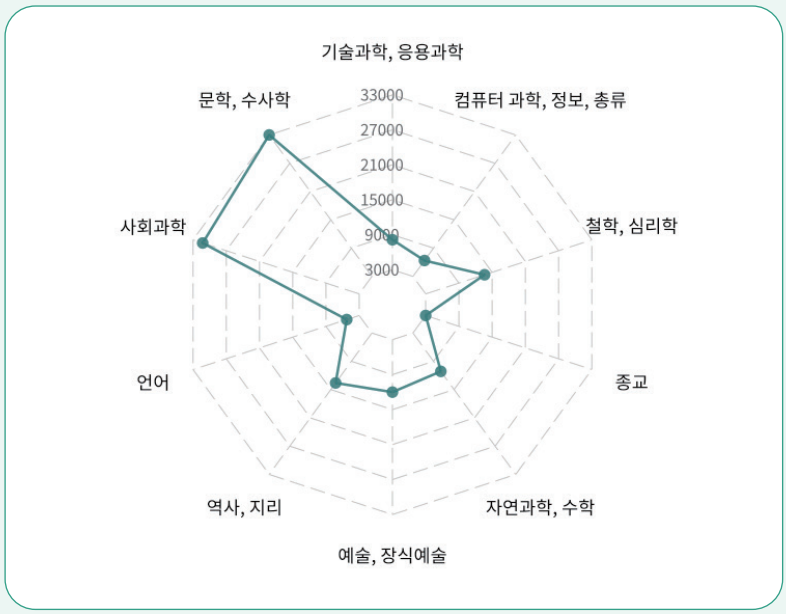
주제별 장서 이용도 분석

- 주제별 장서보유현황

- 2016-2021년 기간 주제별 장서보유현황(6년 평균)은 사회과학(28.50%)과 기술과학 및 응용과학(16.03%)이 가장 많았으며, 언어(3.10%)와 종교(2.61%)가 가장 낮은 구성을 보임
- 한편 학생이용자들의 주제별 대출분포도와 비교하여 살펴보면 사회과학 외에도 문학 · 수사학, 철학 · 심리학 등에 대한 대출 수요가 높은 것으로 드러나 보유 중인 장서와 학생이용자들이 자주 이용하는 장서 간에는 다소 차이가 있는 것으로 보임



[그림 1] 주제별 장서분포도(6년 평균)



[그림 2] 주제별 대출분포도(6년 평균)

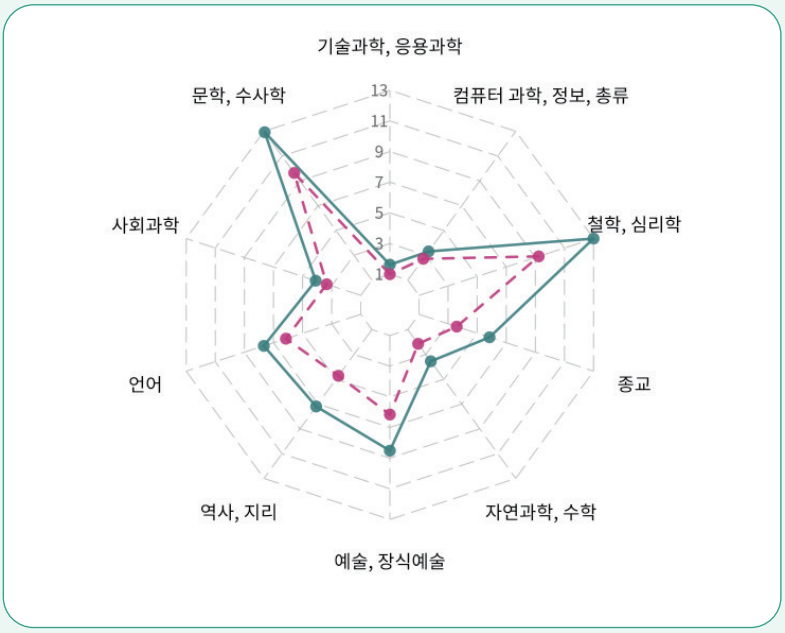
〈표 1〉 중앙도서관 보유 장서 현황 (2016-2021년, 각 연도별 8월 기준)

구분	2016년	2017년	2018년	2019년	2020년	2021년
	N(%)	N(%)	N(%)	N(%)	N(%)	N(%)
컴퓨터 과학, 정보, 총류	177,563 (7.45)	179,515 (7.40)	182,544 (7.41)	184,342 (7.37)	184,969 (7.32)	186,978 (7.23)
철학, 심리학	109,954 (4.62)	112,453 (4.64)	114,791 (4.66)	117,188 (4.69)	118,995 (4.71)	121,739 (4.71)
종교	109,954 (4.62)	62,650 (2.58)	64,158 (2.60)	65,699 (2.63)	66,688 (2.64)	68,067 (2.63)
사회과학	674,509 (28.31)	690,551 (28.47)	703,457 (28.54)	715,359 (28.61)	721,817 (28.57)	736,707 (28.49)
언어	73,510 (3.09)	74,828 (3.08)	75,821 (3.08)	77,851 (3.11)	78,833 (3.12)	79,926 (3.09)
자연과학, 수학	295,841 (12.42)	299,420 (12.34)	302,058 (12.25)	303,688 (12.14)	305,830 (12.11)	310,351 (12.00)
기술과학, 응용과학	388,576 (16.31)	391,797 (16.15)	394,239 (15.99)	396,930 (15.87)	398,532 (15.78)	415,401 (16.07)
예술, 장식예술	144,611 (6.07)	148,680 (6.13)	151,733 (6.16)	154,345 (6.17)	157,716 (6.24)	161,486 (6.25)
문학, 수사학	263,486 (11.06)	268,795 (11.08)	273,827 (11.11)	278,507 (11.14)	282,380 (11.18)	287,833 (11.13)
역사, 지리	193,199 (8.11)	197,123 (8.13)	202,450 (8.21)	206,734 (8.27)	210,308 (8.33)	217,010 (8.39)
합계	2,382,331 (100)	2,425,812 (100.00)	2,465,078 (100.00)	2,500,643 (100.00)	2,526,068 (100.00)	2,585,498 (100.00)

주) N: 장서수(권), %: 비율

주제별 장서회전율

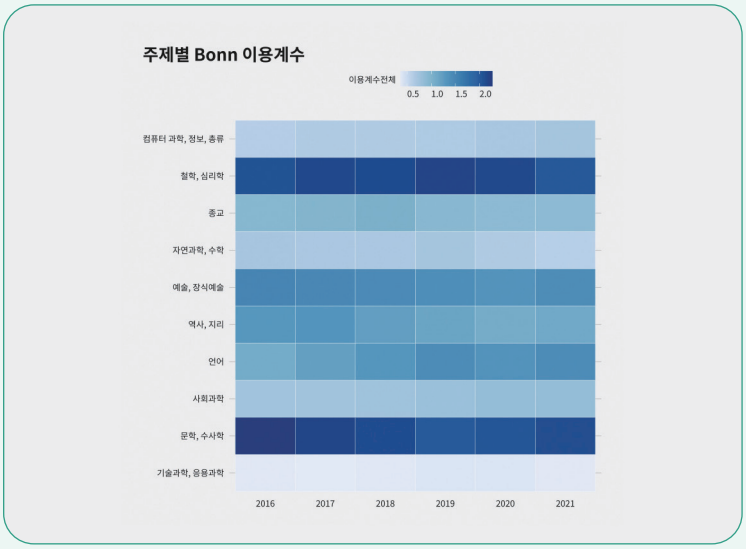
- 장서구성의 적합성을 평가하는 방법으로 장서회전율이 사용됨(이순영 · 이수상, 2021)
- 장서회전율 = 주제별 도서 대출건수/주제별 보유 도서 수(값이 클수록 도서 대출수가 많음을 의미)
- 코로나19 이전과 이후 시점인 2018년도와 2021년을 대상으로 주제별 장서회전율을 비교분석한 결과 주제별 장서회전 비율은 비슷한 경향을 보이는 반편 2021년의 경우 회전율이 증가한 것으로 나타남



[그림 3] 주제별 장서회전율 분포도
(자주색: 2018년, 청록색: 2021년)

주제별 이용계수

- 이용계수에 의한 주제별 이용정도를 살펴보기 위해 Bonn의 이용계수(Use Factor)를 적용하여 이용정도를 환산(양지안 · 남영준, 2016)
  - 이용계수(Use Factor) = [(특정 주제자료의 총 대출수 ÷ 도서관 총 대출수) × 100] ÷ [(특정 주제 분야의 장서수 ÷ 도서관의 총 장서수) × 100]
  - 이용계수가 1보다 크면 활발히 이용되는 장서로 평가(양지안 · 남영주, 2016)
- 이용계수에 의한 주제별 이용정도를 살펴보면 문학 및 수사학(2.13), 철학 및 심리학(2.12)으로 가장 많은 대출이 발생, 반면 가장 낮은 수치를 보인 주제는 컴퓨터 과학, 정보 및 총류(0.66)와 기술과학 및 응용과학(0.38)으로 나타남

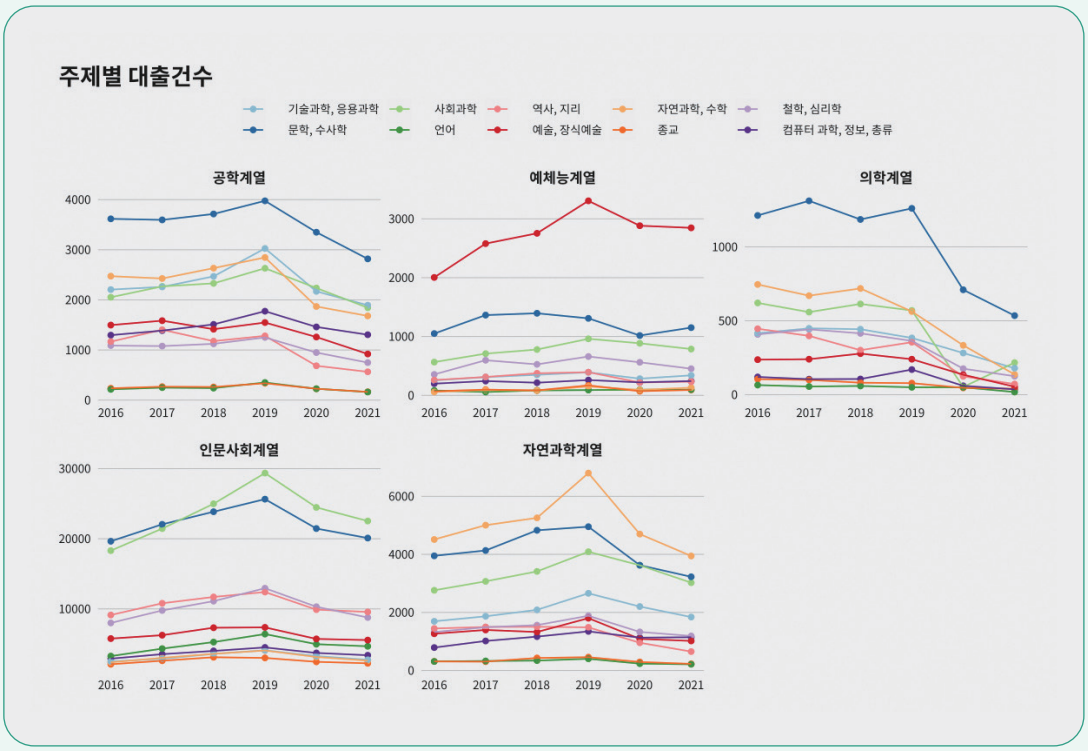


[그림 4] 연도별 주제별 이용계수

학생 전공계열에 따른 주제별 이용도 분석

- 주제별 대출건수

○ 예체능계열, 인문사회계열, 자연과학계열의 경우 전공 관련 주제 도서 대출 건수가 높게 나타나는 반면 공학계열과 의학계열의 경우 문학 및 수사학 도서를 가장 많이 대출하는 경향이 있는 것으로 보임



[그림 5] 전공계열별 주제별 대출건수

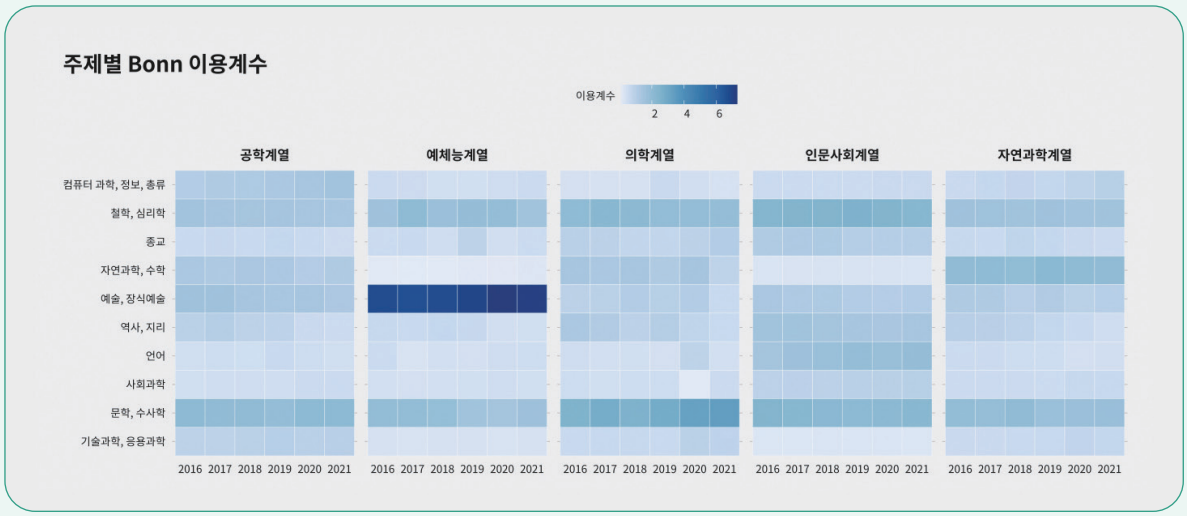
- 주제별이용계수

○ 이용계수 분석 결과 가장 이용정도가 활발한 주제로 인문사회계열은 철학 및 심리학(2.40), 자연과학계열은 자연과학 및 수학(2.02), 공학계열은 문학 및 수사학(2.01), 예체능계열은 예술 및 장식예술 (6.92), 의학계열은 문학 및 수사학(2.87)인 것으로 확인됨

○ 예체능계열의 경우 전공 관련 주제인 예술 및 장식예술 장서 이용도가 매우 활발한 것으로 나타남

〈표 2〉 주제별 계열별 이용계수 평균 (2016-2021년)

구분	인문사회계열	자연과학계열	공학계열	예체능계열	의학계열
컴퓨터 과학, 정보, 총류	0.5544	0.7416	1.2631	0.4935	0.3960
철학, 심리학	2.4038	1.5410	1.4066	1.7546	2.0073
종교	1.1027	0.6296	0.5989	0.5743	0.8777
사회과학	0.9139	0.5785	0.4973	0.4282	0.4320
언어	1.7147	0.4822	0.4871	0.4341	0.4956
자연과학, 수학	0.2836	2.0224	1.1963	0.1273	1.2300
기술과학, 응용과학	0.2197	0.6343	0.9239	0.3143	0.6966
예술, 장식예술	1.1435	1.0460	1.4030	6.9165	0.9241
문학, 수사학	2.2183	1.8252	2.0091	1.7252	2.8690
역사, 지리	1.4302	0.7480	0.7908	0.5697	0.9479



[그림 4] 연도별 주제별 이용계수

- 계열별 1인당 장서 대출현황

○ 계열별 1인당 대출건수 산출

- 계열별 1인당 대출건수 = 계열별 총 대출건수 ÷ 계열별 재적생 수(재학생 및 휴학생)
- 계열별 인원수의 경우 연도별 4월 기준 고등교육통계자료를 이용하였음

○ 계열별 전체 재적생 수를 고려했을 때 1인당 대출건수가 가장 많은 계열은 인문사회계열인 반면 가장 적은 계열은 의학계열인 것으로 확인됨

〈표 3〉 계열별 1인당 대출건수

구분	인문사회계열	자연과학계열	공학계열	예체능계열	의학계열
2016	5.66	2.23	2.03	2.39	1.74
2017	6.66	2.46	2.12	3.09	1.75
2018	7.56	2.70	2.19	3.19	1.67
2019	8.52	3.18	2.50	3.78	1.63
2020	7.00	2.33	1.87	3.05	0.79
2021	6.44	1.97	1.52	3.04	0.57

- 계열별 주제편중도 분석

○ 계열별 주제 편중도 산출

- 최대 주제편중도 = (계열별 최대 대출 주제 분야 대출건수 ÷ 계열별 총 대출건수) × 100
- 최소 주제편중도 = (계열별 최소 대출 주제 분야 대출건수 ÷ 계열별 총 대출건수) × 100
- 주제 편차정도 = 최대 주제편중도 ÷ 최소 주제편중도
- 주제 편차정도가 크면 주제 분야와 비주제 분야 사이의 이용정도의 차이가 크다는 의미인 반면, 해당 값의 수치가 작으면 다양한 주제를 균형적으로 이용한다는 것을 의미함(양지안·남영준, 2016)

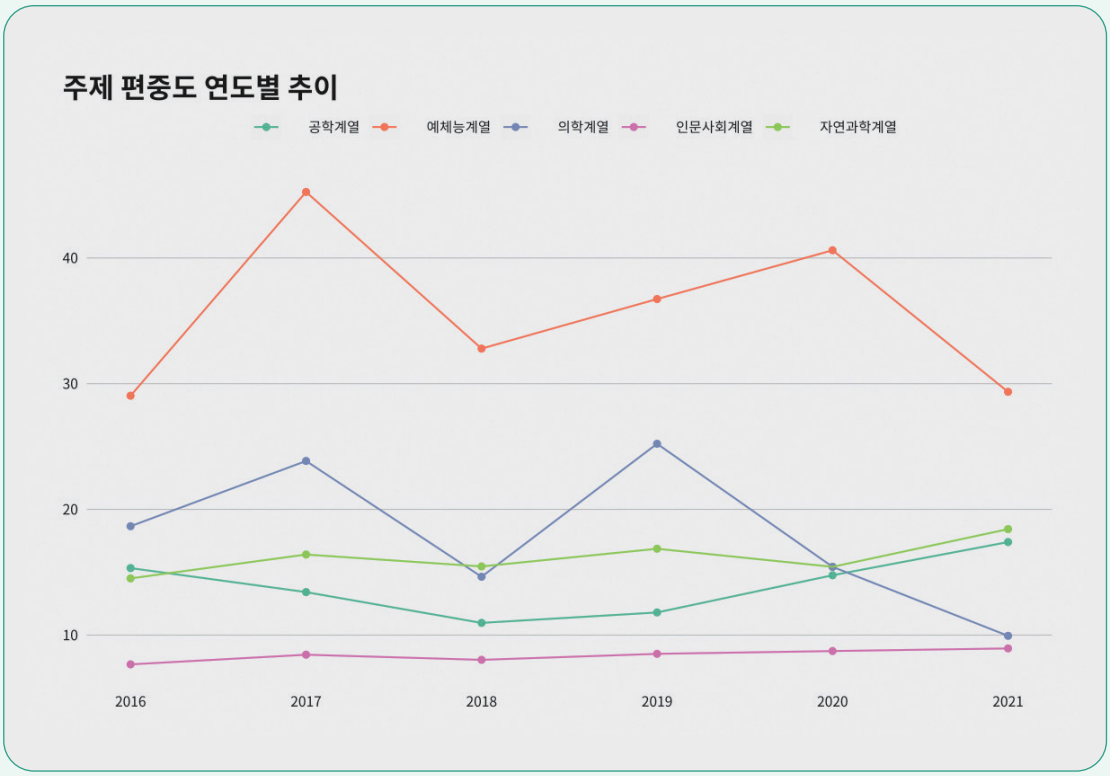
〈표 3〉 계열별 1인당 대출건수

구분	인문사회계열		자연과학계열		공학계열		예체능계열		의학계열	
2016	문학, 수사학	24.72	자연과학, 수학	24.58	문학, 수사학	22.82	예술, 장식예술	40.91	문학, 수사학	27.71
2017	문학, 수사학	25.42	자연과학, 수학	24.89	문학, 수사학	21.76	예술, 장식예술	40.56	문학, 수사학	30.30
2018	사회과학	25.35	자연과학, 수학	23.99	자연과학, 수학	15.60	예술, 장식예술	41.45	문학, 수사학	28.22
2019	문학, 수사학	23.35	자연과학, 수학	26.28	문학, 수사학	20.90	예술, 장식예술	42.97	문학, 수사학	31.25
2020	문학, 수사학	23.99	문학, 수사학	18.91	문학, 수사학	23.21	예술, 장식예술	45.54	문학, 수사학	36.06
2021	문학, 수사학	24.44	자연과학, 수학	23.93	문학, 수사학	23.29	예술, 장식예술	44.67	기술과학, 응용과학	12.67

〈표 3〉 계열별 1인당 대출건수

구분	인문사회계열		자연과학계열		공학계열		예체능계열		의학계열	
2016	문학, 수사학	3.22	언어	1.69	종교	1.49	종교	1.41	사회과학	1.49
2017	종교	3.01	종교	1.52	종교	1.62	언어	0.90	언어	1.27
2018	종교	3.16	언어	1.55	언어	1.42	종교	1.26	종교	1.93
2019	종교	2.74	언어	1.56	종교	1.77	언어	1.17	언어	1.24
2020	종교	2.75	언어	1.23	언어 & 종교	1.57	종교	1.12	종교	2.34
2021	종교	2.73	언어	1.30	언어	1.34	언어	1.52	언어	1.27

○ 도서 이용행태에 있어 인문사회계열이 이용편차가 가장 작게 나타나 주제별로 균형 있게 이용하는 것으로 보임. 한편 주제 편차정도가 가장 크게 나타난 계열은 예체능계열로 타 주제 대비 전공 관련 주제인 예술 및 장식 예술 도서에 대한 선호도가 매우 높은 것으로 드러남



[그림 7] 전공계열별 주제 편중도 변화 추이 (2016-2021년)

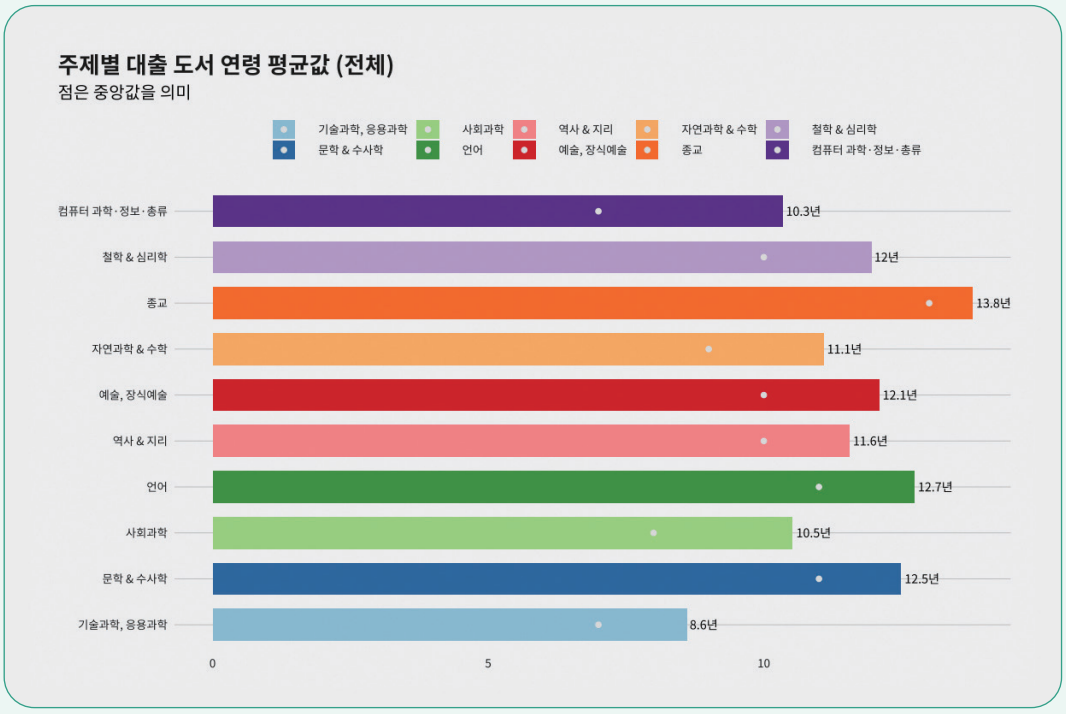
학생 전공계열에 따른 도서연령별 자료대출 현황 분석

- 학생 전공계열에 따른 도서연령별 자료대출 현황 분석

○ 도서연령 = 대출연도 - 출판연도

○ 주제에 따라 이용되는 자료의 최신성에 차이가 있는 것으로 나타남

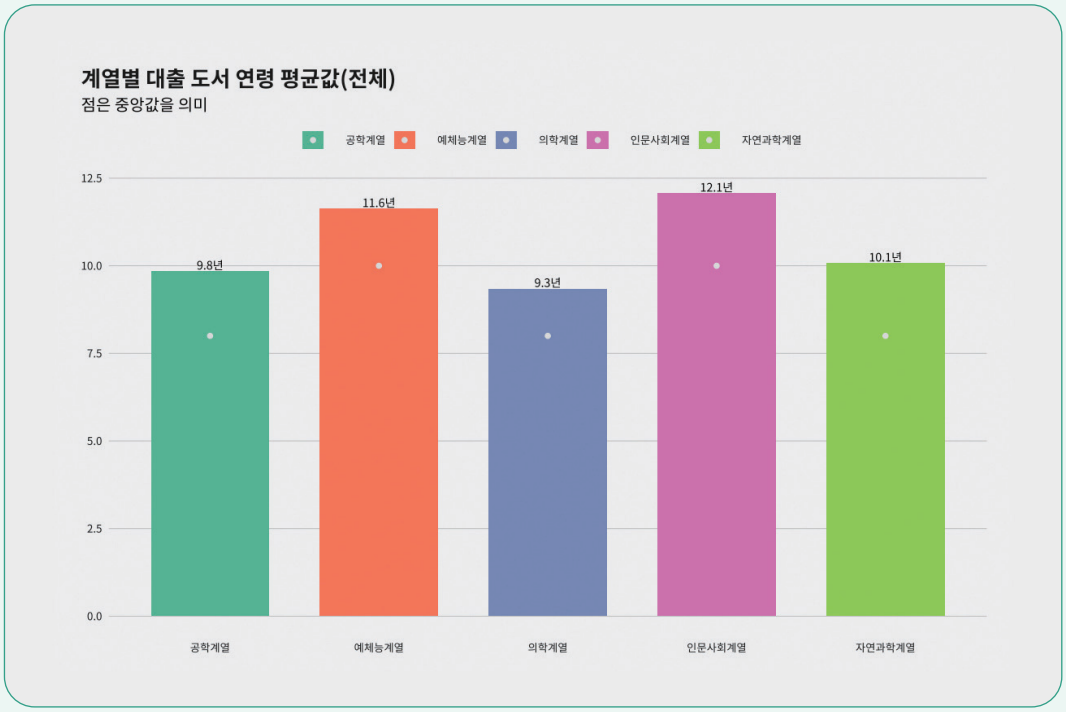
○ 평균보다 작은 값을 가진 분야에 해당하는 주제는 컴퓨터과학·정보·총류, 자연과학·수학, 사회과학, 기술과학·응용과학로 확인됨



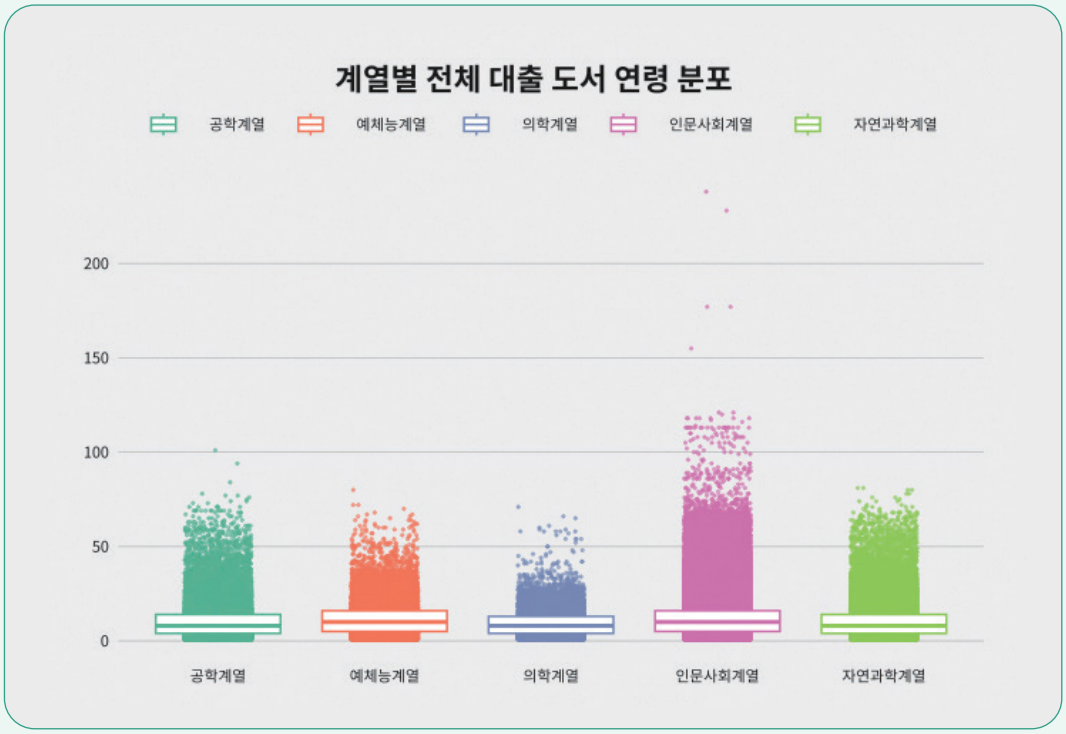
[그림 8] 주제별 대출 도서 연령 평균값

- 도서연령에 따른 자료대출 현황 분석

○ 의학계열과 공학계열은 비교적 최신 도서를 대출하는 경향을 보이는 반면 인문사회계열의 경우 그러한 경향이 가장 약한 것으로 나타남



[그림 9] 계열별 대출 도서 연령 평균값

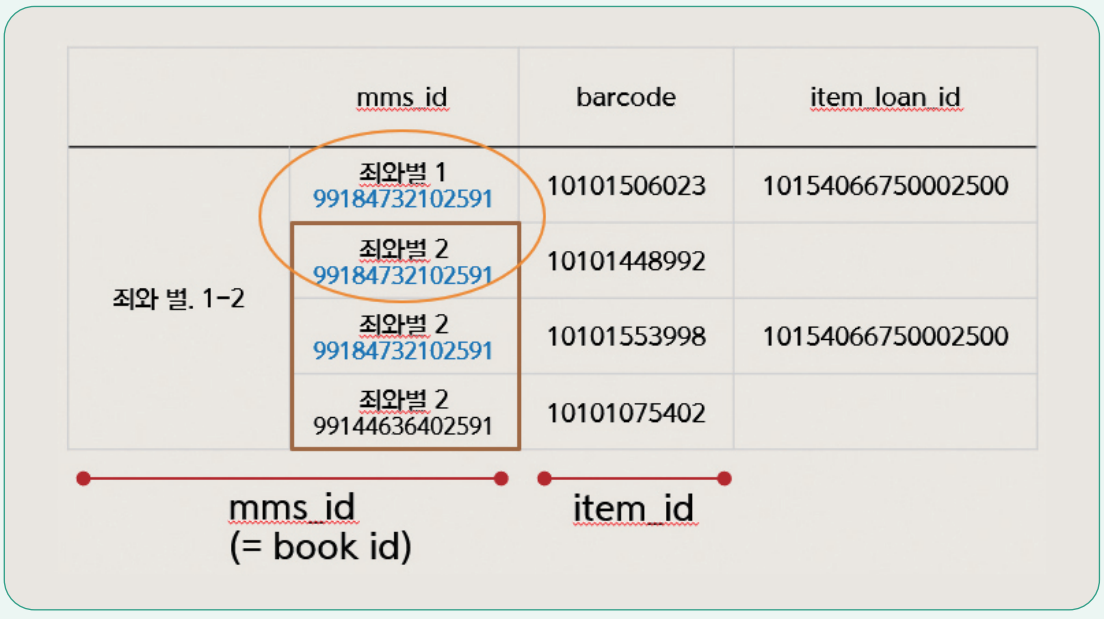


[그림 10] 전공계열별 대출 도서 연령 분포

연구결과

대출데이터 분류체계 개선 필요

- 현재 데이터는 여러 권수로 구성된 단행본의 경우 권수가 다른 동일 제목의 단행본의 경우 동일한 id(mms_id)를 갖는 방식으로 설정되어 있음
- 한편 바코드 정보(item_id, barcode)를 이용할 경우 각각의 단행본 서적별로 고유의 값을 가지고 있어 동일한 제목, 동일한 권수의 서적 여부를 구분할 수 없는 구조임
- 단행본 권수(volume number)별 별도 고유 번호를 부여하거나 단행본 제목에 책 권수를 반영하여 동일한 제목, 동일한 권수의 개별 서적들을 동일한 단행본으로 분류할 수 있도록 새로운 분류체계 마련이 필요할 것으로 보임
- 이를 이용하여 향후 "핵심장서" 분석에 활용하여 핵심장서만을 대상으로 대출 패턴을 분석하여 이용이 활발한 자료와 이용되지 않는 자료를 구별할 필요가 있음 (트루스웰 80/20 법칙에 따른 핵심장서 분류)
- 핵심장서 대출 패턴 분석을 통해 대출이 빈번한 핵심장서에 대한 구매를 통해 이용자 만족도 제고 방안을 모색할 수 있을 것으로 기대됨



[그림 11] 현재 데이터 문제 사례

- 또한 동일 단행본에 대해 도서 주제에 따른 분류가 다른 경우가 있어 개선이 필요할 것으로 보임. 가령 (그림 12)와 같이 동일한 도서임에도 불구하고 상이하게 분류가 되는 한계가 존재함

publishe	title	author_y	call_numbe	section_desc	division	categor
아카넷	국가에 관한 6	Bodin, Jean,	081 H1932da	컴퓨터 과학·정보·총류	일반 전집	한국어총서
아카넷	국가에 관한 6	Bodin, Jean,	081 H1932da	컴퓨터 과학·정보·총류	일반 전집	한국어총서
아카넷	국가에 관한 6	Bodin, Jean,	081 H1932da	컴퓨터 과학·정보·총류	일반 전집	한국어총서
아카넷	국가에 관한 6	Bodin, Jean,	081 H1932da	컴퓨터 과학·정보·총류	일반 전집	한국어총서
아카넷	국가에 관한 6	Bodin, Jean,	081 H1932da	컴퓨터 과학·정보·총류	일반 전집	한국어총서
아카넷	국가에 관한 6	Bodin, Jean,	081 H1932da	컴퓨터 과학·정보·총류	일반 전집	한국어총서
아카넷	국가에 관한 6	Bodin, Jean,	081 H1932da	컴퓨터 과학·정보·총류	일반 전집	한국어총서
아카넷	국가에 관한 6	Bodin, Jean,	320.1 B632s 2	사회과학	정치학	정치학
아카넷	국가에 관한 6	Bodin, Jean,	320.1 B632s 2	사회과학	정치학	정치학
아카넷	국가에 관한 6	Bodin, Jean,	320.1 B632s 2	사회과학	정치학	정치학
아카넷	국가에 관한 6	Bodin, Jean,	320.1 B632s 2	사회과학	정치학	정치학
아카넷	국가에 관한 6	Bodin, Jean,	320.1 B632s 2	사회과학	정치학	정치학
아카넷	국가에 관한 6	Bodin, Jean,	320.1 B632s 2	사회과학	정치학	정치학
아카넷	국가에 관한 6	Bodin, Jean,	081 H1932da	컴퓨터 과학·정보·총류	일반 전집	한국어총서
아카넷	국가에 관한 6	Bodin, Jean,	081 H1932da	컴퓨터 과학·정보·총류	일반 전집	한국어총서

[그림 12] 현재 데이터 주제별 분류 문제 사례

⚙ 출판연도 데이터 관리 개선 필요

- 도서 연령 산출 시 출판년도와 대출연도에 대한 정보를 이용하였음. 그러나 출판연도에 대한 정보가 없거나 잘못 기록된 경우가 있어 개선이 필요할 것으로 보임. 출판연도에 대한 정보가 없을 경우 출판일자에 대한 정보를 활용하였으나, 출판일자 데이터를 정제하는데 추가적인 노력이 요구되었음
- 출판경과년수에 따른 대출 패턴을 보다 정확히 분석하기 위해 대출연도에 대한 데이터 관리 개선이 필요할 것으로 판단됨

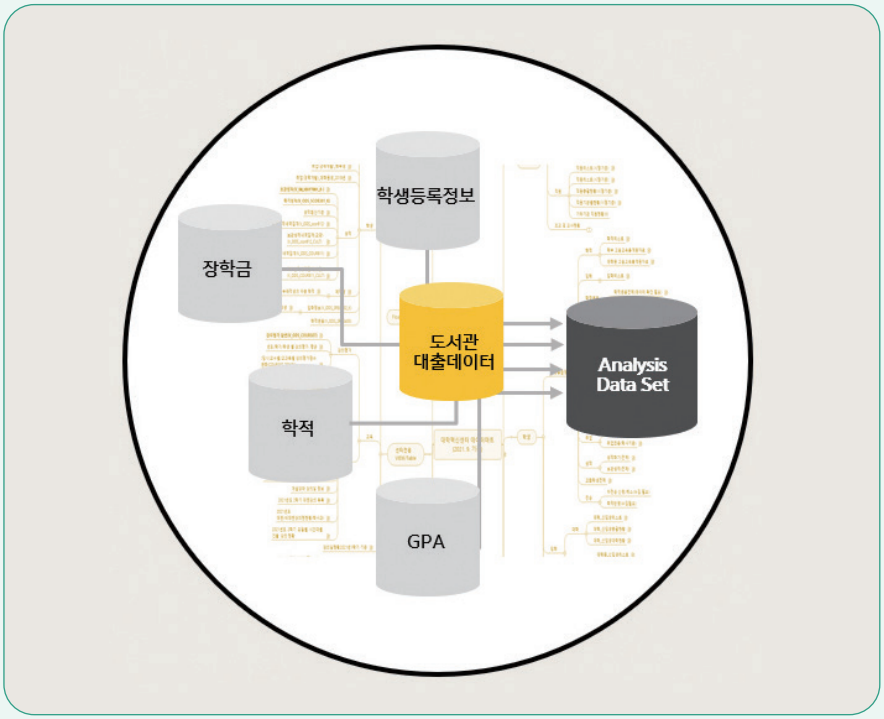
⚙ 정보화본부가 관리하는 학내 데이터와의 연계방안 모색 필요

- 정보화본부는 학내DB에 축적되는 다양한 데이터를 관리하고 있으며 데이터마트를 통해 원본데이터를 제공하고 있음. 데이터마트에는 학적, 교원, 직원, 졸업, 등록, 장학, 수업, 강의평가, 성적, 국제협력 외에도 건물 단위 고정IP 및 무선랜(WIFI) 접속정보 등에 관한 데이터가 관리됨
- 대학혁신센터는 데이터마트의 데이터를 활용하여 학내 정책 수립을 지원하기 위한 다양한 분석을 수행하고 있는 바, 향후 도서관과의 긴밀한 협력을 통해 도서관에서 관리하는 다양한 빅데이터 분석을 통해 의미있는 분석을 수행할 수 있을 것으로 기대됨

※ 대학혁신센터는 2021년 데이터분석과제로 장학복지과와의 협업을 통해 '학내 밀집도 분석을 통한 효율적인 코로나19 방역대책 마련'을 주제로 고정IP 및 와이파이 빅데이터를 이용하여 서울대 건물 단위 유동인구를 분석하였음. 해당 분석의 일환으로 관정관과 중앙도서관의 시간대별 유동인구 분석을 수행한 바 있음

○ 당 분석을 수행는데 있어 정보화본부가 관리하는 데이터마트(DM)로부터 학생 전공계열에 대한 정보를 담고 있는 학부 및 대학원 고등교육통계원자료를 활용하였음. 이를 vuser_refine 데이터에 포함된 신분, 소속대학 및 소속학과 정보와 매칭시켜 학생이용자들의 전공계열에 관한 정보를 이용하여 분석을 수행하였음

○ 향후 도서관 빅데이터를 정보화본부의 데이터마트와 연계할 수 있는 구체적인 방안에 대한 논의가 필요할 것으로 보임. 구성원별 ID의 비식별화 방식을 통일할 경우 데이터 연계를 보다 용이하게 할 수 있을 것으로 판단되며, 이와 같은 노력을 통해 도서관 빅데이터를 이용한 다양한 분석 수행이 가능할 것으로 기대됨



[그림 13] 도서관 빅데이터와 정보화본부 DM 데이터 간의 연계 예시

향후계획

핵심장서 분석

○ 도서관 단행본의 이용률 분석을 통해 핵심장서를 산출할 수 있으며, 가장 대표적인 방법으로 Trueswell(1969)의 80/20법칙(약 20%의 장서가 전체 대출의 약 80%를 차지)이 있음

○ Trueswell은 단행본의 대출횟수가 높은 장서는 장서활용도가 상대적으로 높다는 것을 복수의 연구를 통해 밝혀내었음(양지안, 2017)

○ 서울대학교의 경우에도 대출횟수를 이용하여 전체 대출건수의 과반(50%) 또는 그 이상(80%)을 차지하는 대출

도서 법칙을 확인하여 연도별 대출 패턴을 상세히 살펴보면 보다 의미있는 시사점 도출이 가능할 것으로 보임

- 이와 같은 분석을 수행하기 위해서는 상기에서 제안한 바와 같이 장서 권수별 고유 번호를 부여하는 등의 대출 데이터 분류체계 개선이 선행될 필요가 있음
- 향후 핵심장서 대출 패턴 분석을 통해 이용이 활발한 자료와 이용되지 않는 자료를 중심으로 효율적 장서 구성 및 구매 방안 마련이 가능할 것으로 사료됨

장서 대출 패턴 관련 대시보드 구축

- 빅데이터에 가치를 부여하고 이를 통해 사용자들에게 인사이트를 제공하는 데 있어 데이터 시각화의 중요성이 높아지고 있음
- 빅데이터 분석의 핵심은 얼마나 많은 양의 데이터를 사용했는지가 아니라, 수집된 빅데이터의 분석 결과를 쉽게 이해할 수 있도록 하는 시각화를 통한 유용성에 있음
- 도서관 빅데이터의 다양한 활용이 가능할 수 있도록 기반과 체계를 강화하고, 이를 토대로 다수의 시각화 차트를 종합하여 실시간으로 보여줄 수 있는 데이터 시각화 대시보드의 구축을 통해 적시에 인사이트를 도출할 수 있을 것으로 기대됨

참고문헌

양지안, 남영준(2016). 대학도서관 단행본 대출이력통계를 통한 집중장서에 관한 연구. 한국문헌정보학회지, 50(3): 429-435.

양지안 (2017). 대학도서관 대출데이터분석을 통한 장서 이용행태 및 특성에 관한 연구. 한국도서관·정보학회지, 48(2), 263-293.

이순영, 이수상(2021). 부산지역 공공도서관의 빅데이터 분석 연구: 도서관 정보나루 장서/대출데이터를 중심으로. 한국문헌정보학회지, 55(4): 89-114.

PART. III
연구과제

서울대학교 도서관 데이터 분석

연구배경
연구내용
연구결과
향후계획

서울대학교 도서관 데이터 분석

KAIST 인문사회융합과학대학
김란우

연구배경

본 연구는 학생들이 도서관을 이용하면서 느끼는 두 가지 불편한 점(pain point)을 개선하고자 하는데 있다. 관악 캠퍼스는 너무 넓고, 접근성이 좋지 않다. 그리고 장서들이 다양한 단과대학교 도서관에 흩어져있다. 도서관 이용자들이 다양한 학내도서관을 사용할 때는 크게 두 가지 패턴이 있을 것이다.

1) 단과대학 도서관에서만 대여할 수 있는 책이 있다.
2) 관악 캠퍼스가 넓기 때문에, 멀리 위치한 도서관에 방문하기는 어렵다. 책을 빌리겠다는 의지는 상대적으로 commitment가 낮기 때문에, 지리적 이유로 포기하기 쉬울 수 있다.

두 가지 패턴을 종합하면, 책을 대여하여 읽음으로써 생겨날 수 있는 특정 분야에 대한 관심이 도서관의 지리적인 위치가 가지는 한계로 인하여 없어질 수 있다. 이는 서울대학교의 단과대학 체제를 공고히 하고 융합학문의 가능성을 줄일 수 있다는 한계점을 내포한다.

도서관 서비스는 기본적으로 내가 빌리는 책을 매개로 그 책을 대여한 다른 누군가와 관계가 생기는 2-mode network의 성격을 띤다. 따라서 책을 대출하려할 때 해당 책이 이미 대출중인 경우, 상대방에 대한 정보를 사용하지 않으면 해당 도서가 언제 반납될지 예측할 수 없다. 대기 신청을 하였을 때 알람 문자가 오지만, 이는 도서관을 이용할지 여부를 결정하는데는 크게 효율적이지 않다.

도서 반납 기간에 영향을 주는 요인은 다음과 같다.

○ 개인의 특성이 도서 반납 일자를 정하는데 영향을 미친다 : 책을 오래 읽는 사람과 그렇지 않은 사람 간에 차이가 있음

○ 학교에서의 소속, 지위 등에 따라 도서 대여 기간이 다르다. 학부생의 도서 대여 기간이 교수에 비해 짧다

○ 책의 특성이 도서를 반납하는데 영향을 미친다. 책의 언어, 페이지 수, 분야 등의 특성에 따라 책이 얼마나 빨리 읽히는지 예측할 수 있다

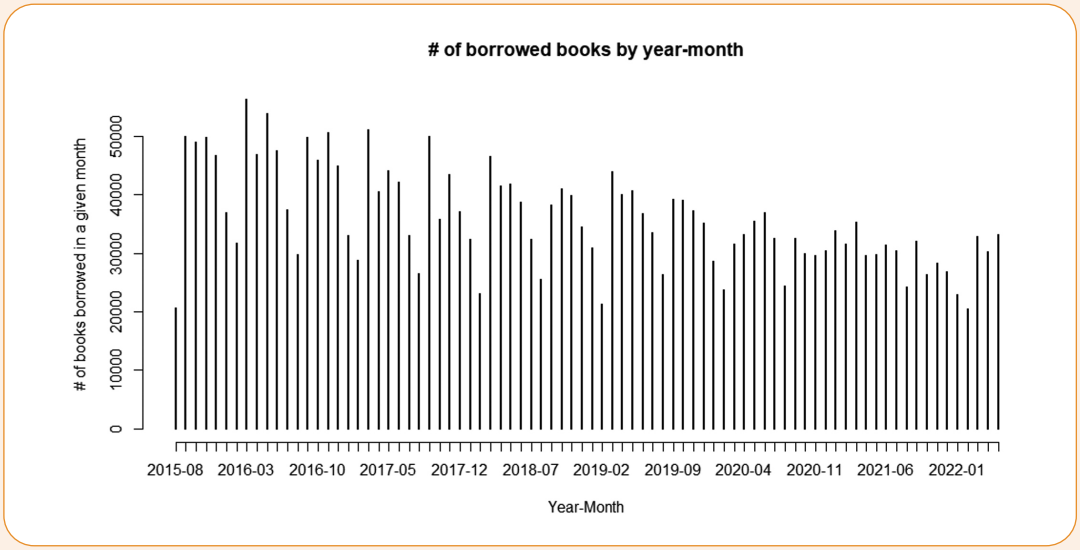
○ 시기의 특성이 도서를 반납하는데 영향을 미친다. 학기가 끝나는 시기에 대규모로 도서들이 반납될 것을 예측할 수 있다

이와 같은 변수를 고려하여, prediction machine learning model을 만들고, 원하는 책이 대출중일 때, 예상 반납시간을 제시할 수 있다.

연구내용

도서관 접근성은 필요한가?

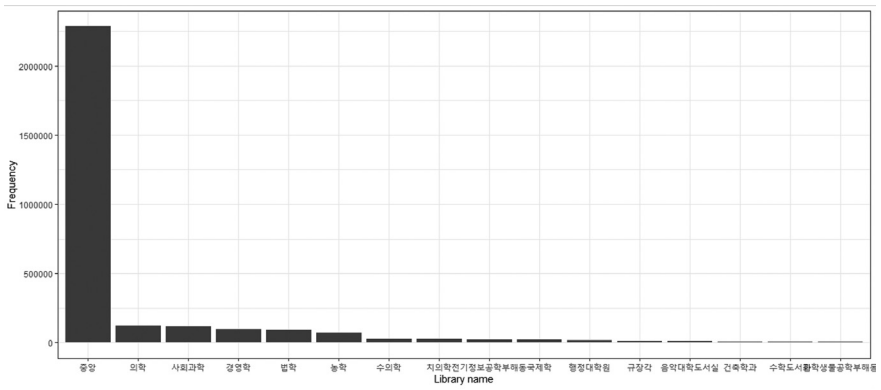
○ 본격적인 분석에 앞선 데이터 트렌드 및 문제점



위의 그림은 대출 데이터만을 통해 월별 대출 트렌드를 살펴본 것이다. 2015년부터 데이터가 시작하며, 전반적으

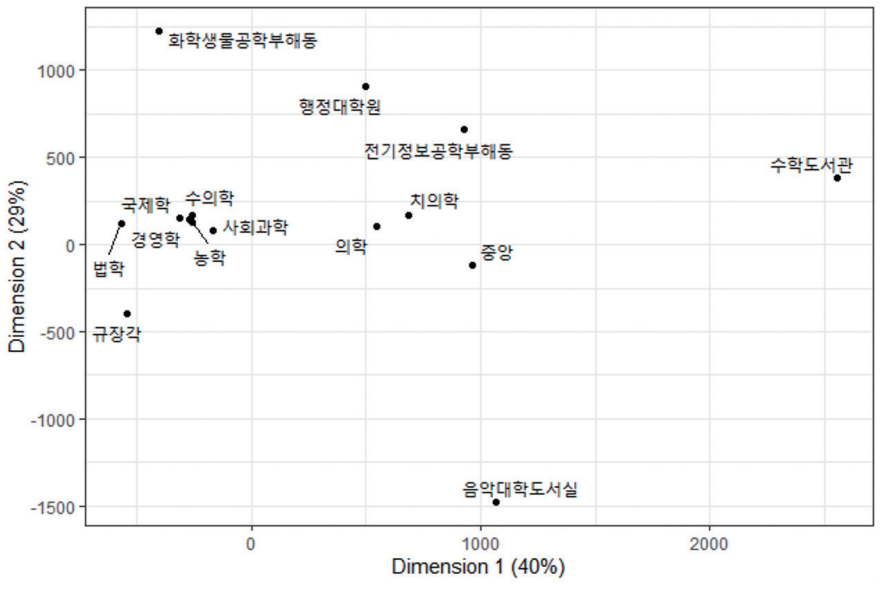
로 대출 횟수는 꾸준히 감소하는 것으로 나타난다. 그리고 뚜렷한 Seasonality가 있는 점을 찾아볼 수 있다.(학기 시작할 때 대량 대출). Seasonality도 Covid 기간 동안에는 희미해진 점을 찾아볼 수 있다. 하지만 같은 데이터를 대출자 정보(borrower)와 연동하여 데이터베이스에서 query를 진행했을 경우, 위의 그림과 다른 트렌드를 보여주고 있는 점을 찾아볼 수 있다. 즉, 대출 데이터의 초반 부분 (2015년~ 2018년)의 대출자 정보에 어딘가 구조적인 missing이 있는 점을 찾아볼 수 있다.

○ 중앙도서관에 대출 집중



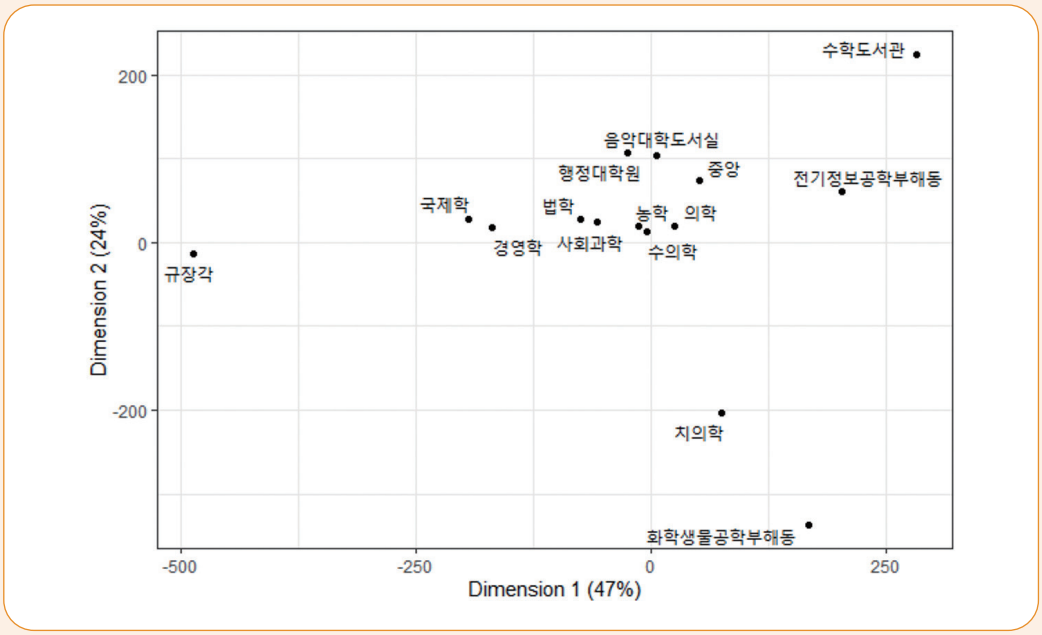
중앙도서관에 대부분의 도서 대출이 집중되어 있는 점을 찾아볼 수 있다.

○ 도서관-책의 2-mode network를 사용한 distance 분석



가장 이상적인 분석은 각 도서관이 소장하고 있는 장서 정보를 정확히 수집하여, 얼마나 장서가 겹치는지를 분석하는 것이겠으나, 해당 데이터가 현재는 없기 때문에 이용자의 도서 대출 데이터를 토대로 분석하였다. 두 도서관 사이가 가깝다는 것은 두 도서관에서 대출된 책들이 겹치는 경우가 많다는 것을 뜻한다. MDS 분석에서 2 dimension이 도서관들 사이의 distance matrix의 69%를 설명하는 것으로 나타났다. 특히, 중앙 도서관은 이 2 dimension에서도 중앙에 위치하는 것으로 나타났다. 국제학, 경영학, 수의학, 농학, 사회과학 도서관은 가깝게 cluster 되어 있는 것으로 나타났다. 음악대학 도서실, 수학도서관, 행정대학원 도서관의 대출 기록은 타 도서관들과 차이가 큰 것으로 보인다. 하지만 해당 분석은 2권 이상 겹치는 책들만을 대상으로 했다는 한계점이 있다. 즉, 예를 들어 규장각에서 100권의 책은 규장각에만 있는 책들이었다면, 해당 분석에서 이 데이터는 반영되지 않았다.

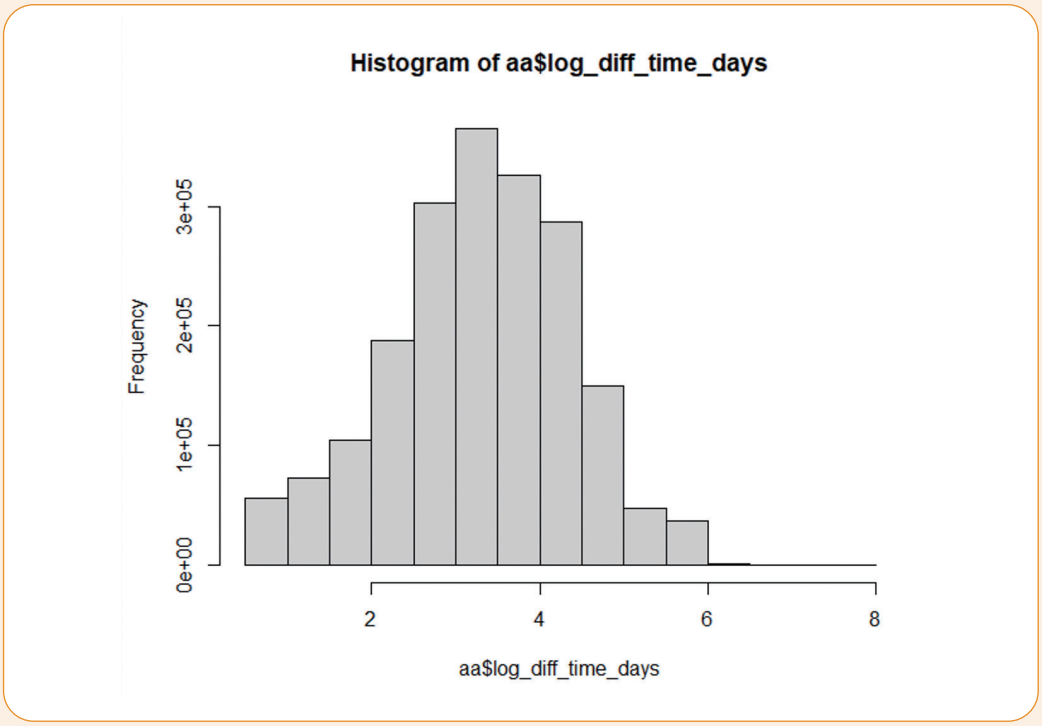
○ 대출자-책의 2-mode network를 사용한 distance 분석



이번 분석에서는 책이 아닌 대출자를 중심으로 도서관 사이의 distance를 계산하였다. MDS 분석에서 2 dimension이 도서관들 사이의 distance matrix의 71%를 설명하는 것으로 나타났다. 첫 번째 dimension에서 가장 거리가 먼 곳은 규장각과 전기정보공학부해동/수학도서관인 것으로 나타났다. 두 번째 dimension에서 가장 거리가 먼 곳은 수학도서관과 화학생물공학부해동이었다. 그 외의 대부분 도서관들은 중앙에 모여있는 것을 확인할 수 있다. 도서관-책의 클러스터링 결과와 비교하여 보았을 때, 대출자를 기반으로 한 분석이 책을 기반으로 한 분석보다 더 중앙에 밀집해 있는 모습을 확인할 수 있다. 즉, 장서는 도서관에 따라 구비하고 있는 책에 특성이 있고 차이가 있지만, 사람들은 필요하다면 어느 도서관이든 방문한다는 점을 보여준다.

도서 대출 예측 모델

○ 종속변수: 도서 대출 기간. Right-skewed 분포이기 때문에 log값을 취함. log를 취한 이후에는 정규 분포와 매우 흡사함.



○ 독립변수: 출생연도, 대출자 그룹(강사, 교수, 학부생 등), 소속 단과대학, 학과, 성별, 학위과정, 입학연도, 대출연도, 현재 재적 상황(수료, 수학가능, 재적, 졸업 등), 대출 도서관. NA가 많기 때문에 이들을 삭제하기 보다는 하나의 그룹으로 생성하여, 모든 변수는 범주형 변수로 변형되었다.

○ 모델: 종속변수가 정규분포를 따르기 때문에 가장 간단한 OLS 회귀를 사용하였다.

○ 결과: Adjusted R-squared가 23%로 약 23%의 variance가 위의 변수로 설명되는 것으로 나타났다. 몇몇 결과는 아래와 같다.

```
> summary(m1)

Call:
lm(formula = log_diff_time_days ~ birth_year_category + user_group +
    college + gender + degree.y + admission_year_category + loan_year +
    department + grade_detail + library_code, data = aa)

Residuals:
    Min      1Q  Median      3Q     Max
-4.7617 -0.5436  0.0970  0.6971  5.0422

Coefficients: (119 not defined because of singularities)
                                                    Estimate Std. Error  t value Pr(>|t|)
(Intercept)                                       -1.126e+02  8.933e-01 -126.042  < 2e-16 ***
birth_year_categoryUnder1980                      -1.446e-01  4.568e-03  -31.659  < 2e-16 ***
birth_year_categoryUnder1990                      -1.985e-01  5.282e-03  -37.582  < 2e-16 ***
birth_year_categoryUnder2000                      -3.315e-01  5.643e-03  -58.744  < 2e-16 ***
birth_year_categoryUnder2010                      -5.158e-01  6.914e-03  -74.611  < 2e-16 ***
user_group강사                                     3.803e-01  1.116e-01    3.408 0.000654 ***
user_group교수                                     9.002e-01  1.116e-01    8.063 7.43e-16 ***
user_group교환학생                                 7.935e-02  3.582e-01    0.222 0.824676
user_group기금교수                                 1.351e+00  1.160e-01   11.641  < 2e-16 ***
user_group도서관졸업중(졸업생)                     1.037e-01  1.116e-01    0.929 0.352838
user_group동문회원(대출)                           1.279e-01  1.116e-01    1.147 0.251524
user_group동문회원(회원제)                         2.681e-01  1.311e-01    2.044 0.040909 *
user_group명예교수                                 6.190e-01  1.130e-01    5.480 4.26e-08 ***
user_group명예회원(교수) (구 일용직)               1.086e+00  1.169e-01    9.296  < 2e-16 ***
user_group박사과정                                 4.026e-01  1.115e-01    3.610 0.000306 ***
user_group법인직원                                 3.689e-01  1.129e-01    3.266 0.001090 **
user_group병원직원                                 6.610e-01  1.515e-01    4.362 1.29e-05 ***
user_group부교수                                   1.142e+00  1.117e-01   10.223  < 2e-16 ***
user_group부속학교 교사                            1.473e+00  5.564e-01    2.648 0.008091 **
user_group비전임교수                               7.797e-01  1.120e-01    6.962 3.35e-12 ***
user_group석박사통합과정                           6.466e-01  1.474e-01    4.388 1.15e-05 ***
user_group석사과정                                 2.814e-01  1.115e-01    2.524 0.011617 *
user_group연구생(대학원생)                         3.055e-01  1.115e-01    2.739 0.006171 **
user_group연구원(특,교,보)                         2.823e-01  1.118e-01    2.526 0.011525 *
user_group인턴                                     2.251e-01  1.111e-01    2.026 0.042764 *
user_group임상강사                                 4.116e-03  1.490e-01    0.028 0.977963
user_group임상교수                                 2.580e-01  2.569e-01    1.004 0.315339
user_group자체직원                                 4.357e-01  1.117e-01    3.902 9.53e-05 ***
user_group전공의                                   1.625e-01  1.140e-01    1.425 0.154106
user_group조교                                     3.197e-01  1.118e-01    2.859 0.004251 **
user_group조교수                                   8.165e-01  1.119e-01    7.298 2.93e-13 ***
user_group조빙교원                                 1.189e+00  1.254e-01    9.483  < 2e-16 ***
user_group특별회원(회원제)                        -2.455e-01  1.448e-01   -1.695 0.090079 .
user_group학부                                     5.177e-02  1.115e-01    0.464 0.642476
user_group휴학생(대학원)                           3.019e-01  1.117e-01    2.704 0.006861 **
user_group휴학생(학부)                             1.425e-02  1.116e-01    0.128 0.898404
collegeRNA 연구단                                 -2.329e-01  1.263e-01   -1.844 0.065185 .
college간호과학연구소                             -3.186e-02  8.412e-02   -0.379 0.704890
college간호대학                                   -2.357e-01  9.616e-02   -2.452 0.014225 *
college강상관계물질연구단                          4.261e-01  1.643e-01    2.593 0.009506 **
college건강사회정책실                             -6.650e-01  5.660e-02  -11.751  < 2e-16 ***
college건설환경종합연구소                          2.204e-02  3.896e-02    0.566 0.571554
college경력개발센터                               -7.935e-01  1.254e-01   -6.328 2.48e-10 ***
college경영대학                                   -5.652e-02  1.440e-02   -3.925 8.66e-05 ***
```

```
college자동화시스템공동연구소                      2.900e-01  9.633e-02    3.011 0.002606 **
college자연과학대학                               -1.089e-01  1.874e-02   -5.812 6.18e-09 ***
college자유전공학부                                2.602e-02  1.446e-02    1.799 0.071977 .
college재단법인기초전력공학공동연구소             -3.241e-01  1.668e-01   -1.943 0.052015 .
college재단법인서울대학교발전기금                 -1.426e-01  2.107e-01   -0.677 0.498566
college재정전략실                                 -2.506e-01  8.146e-02   -3.076 0.002096 **
college전력연구소                                 -5.314e-02  9.688e-02   -0.549 0.583340
college정밀기계설계공동연구소                      6.806e-02  3.999e-02    1.702 0.088783 .
college정보화본부                                  4.616e-01  2.981e-01    1.549 0.121427
college조형연구소                                  5.837e-01  1.643e-01    3.554 0.000380 ***
college종합약학연구소                             -5.201e-01  3.917e-02  -13.277  < 2e-16 ***
college중앙도서관                                 -3.830e-01  8.399e-02   -4.560 5.12e-06 ***
college중권·금융연구소                             6.640e-02  3.556e-01    0.187 0.851865
college차세대우주추진연구센터                     -2.235e-01  9.402e-01   -0.238 0.812132
college차세대융합기술연구원                       -4.634e-01  2.220e-01   -2.087 0.036851 *
college천연물과학연구소                            8.371e-03  2.284e-01    0.037 0.970765
college철학사상연구소                             -1.715e+00  2.082e-02  -82.361  < 2e-16 ***
college치과대학                                    9.617e-02  5.852e-02    1.643 0.100315
college치의학대학원                                6.921e-02  5.186e-02    1.334 0.182066
college치학연구소                                  6.588e-03  4.619e-02    0.143 0.886577
college컴퓨터연구소                               -3.360e-01  8.647e-02   -3.886 0.000102 ***
college통일평화연구원                              3.681e-02  4.241e-02    0.868 0.385366
college평생교육원                                  1.360e-01  7.229e-02    1.882 0.059829 .
college학생생활관                                  2.659e-01  9.724e-02    2.734 0.006253 **
college학생처                                     -2.352e-02  3.441e-02   -0.684 0.494282
college한국우주신기술연구소                       -4.718e-01  1.012e-01   -4.663 3.12e-06 ***
college한국행정연구소                              4.538e-01  5.202e-02    8.725  < 2e-16 ***
college해양시스템공학연구소                        8.981e-02  1.964e-01    0.457 0.647557
college해양연구소                                 -3.383e-01  1.011e-01   -3.346 0.000821 ***
college행정대학원                                 -3.373e-02  1.482e-02   -2.275 0.022878 *
college화학공정신기술연구소                        2.051e-01  9.108e-02    2.252 0.024348 *
college환경계획연구소                              3.740e-01  2.511e-02   14.895  < 2e-16 ***
college환경대학원                                  1.137e-01  7.045e-02    1.614 0.106550
college환경안전원                                  1.359e-01  2.616e-01    0.519 0.603509
gender여자                                         7.582e-02  1.627e-03   46.599  < 2e-16 ***
degree.y박사                                       2.249e-01  6.241e-03   36.033  < 2e-16 ***
degree.y불명                                       1.002e-02  2.210e-02    0.453 0.650309
degree.y석사                                       1.942e-01  3.056e-03   63.563  < 2e-16 ***
degree.y학사                                      -1.055e-01  2.106e-03  -50.097  < 2e-16 ***
admission_year_categoryUnder 2020                 -1.340e-01  3.285e-03  -40.772  < 2e-16 ***
admission_year_categoryUnder1980                  -5.266e-02  4.329e-02   -1.217 0.223749
admission_year_categoryUnder1990                  -2.413e-01  7.407e-03  -32.576  < 2e-16 ***
 [ reached getOption("max.print") -- omitted 439 rows ]
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9401 on 1724803 degrees of freedom
  (결측으로 인하여 210559개의 관측치가 삭제되었습니다 )
Multiple R-squared:  0.2268,    Adjusted R-squared:  0.2266
F-statistic:   975 on 519 and 1724803 DF,  p-value: < 2.2e-16
```

연구결과

제안 분석 모델 1 - 지리적인 이유로 대출되지 않는 건수 측정

첫 번째 데이터 분석 결과는 책은 비교적 분관별로 특성화되어 있으나, 대출자는 이에 크게 구애받지 않고 이동하는 것을 확인할 수 있다. 하지만 위의 데이터 분석 결과는 두 군데 이상의 도서관을 방문한 대출자에 한하여 분석한 결과이며, 대부분의 대출자는 중앙도서관 한 군데만 이용하는 것으로 추정된다. 서울대학교 관악캠퍼스의 까다로운 지리적 특성을 고려할 때, 이는 개인의 선호라기보다 강제된 측면이 있다. 분관의 책이 중앙도서관에 진열되어 있었다면, 얼마나 더 다양한 책이 대출될 수 있었을지에 대한 분석 모델이 필요하다. 다른 연구자들 및 기업으로부터 도출된 도서대여 추천 시스템을 사용하여, 분관의 책과 중앙도서관의 책 사이에 propensity score matching을 진행할 필요가 있다. 그래서 propensity score matching이 된 책들이 중앙도서관에 있지 않다는 이유로 얼마나 덜 대출되는지를 분석한다면, 지리적인 이유로 줄어드는 대출의 정도를 파악할 수 있을 것이다.

제안 분석 모델 2 - 더 구체화된 대출 반납 시기 prediction algorithm 분석

두 번째 분석 결과에서는 오직 학교에서 제공하는 해당 대여자에 관한 정보만을 사용하여 분석하였으며, 개인의 특성 / 책의 특성 / 시기적 특성 등을 고려하지 않았음에도 반납할 날짜를 약 22% 예측할 수 있었다. 해당 정보를 더 추가하고 기계학습 모델을 더 견고하게 구성한다면, 책의 반납 날짜를 더 정확하게 예측할 수 있을 것이다.

단, 데이터상의 문제를 해결해야할 필요가 있다. 책 대출 데이터는 각 row가 1건의 대출 기록으로 구성되어 있다. 하지만 책을 대출한 대출자 정보의 경우 각 row가 1명의 대출자로 구성되어 있다. 문제는 대출자의 정보는 시간이 지나며 계속 업데이트 되지만, 현재 도서관 데이터의 정보는 2022년 현재 최종 업데이트 된 정보만 포함하고 있기 때문에, 정확히 책을 대출한 시점에 대출자가 어떤 상황이었는지 알기는 어렵다. 예를 들어, 2003년 입학하여 2007년 졸업한 학부생의 정보는 대출자 정보에서는 '졸업' 상태로 최종 기록되어 있다. 하지만 이 학부생이 2003~2007년 사이에 책을 대출하였다면, 이 학부생은 졸업생 신분으로 대출한 것이 아니라 재학생 신분으로 대출한 것이다. 그렇기 때문에 두 데이터베이스를 단순히 결합하여서는 안되며, 입학년도와 졸업년도를 사용하여 해당 시기의 상태를 파악할

필요가 있다. 그렇다고 하더라도 구체적으로 휴학생 신분이었는지 등의 여부는 알기 어렵다. 수업 데이터를 사용하여 수업을 수강한 학기였는지를 파악하면 휴학 중이었는지의 여부도 더 구체적으로 추정할 수 있을 것이다.

향후계획

도서관 circulation desk 운영

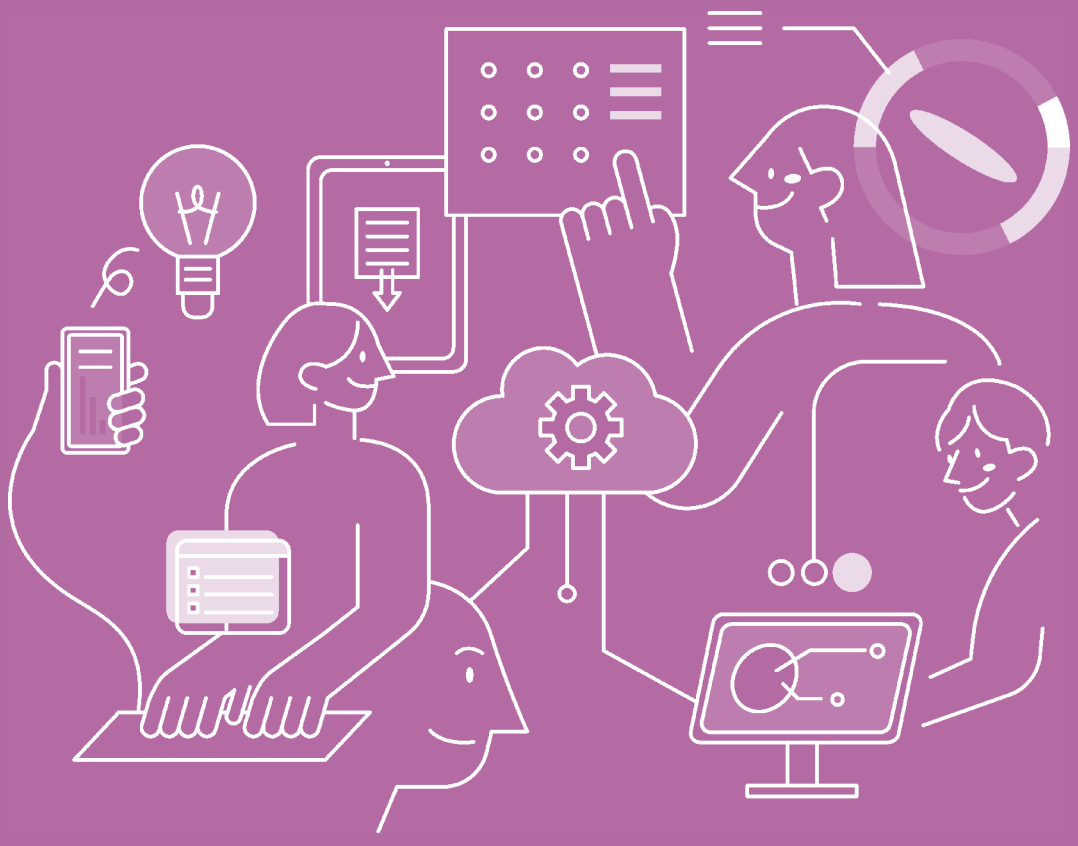
만약 분석모델 1의 결과에서 지리적인 이유로 대출되지 않는 건수가 상당량 이상으로 밝혀진다면, 도서관 차원에서 circulation desk 운영을 고려해볼 필요가 있다. 현재 반납은 어디에서나 가능하지만 대출은 어디에서나 가능하지는 않다. 미리 도서관 시스템을 통해 학교 어느 도서관에서나 대출하고자 하는 책을 신청하고, 정해진 한 군데의 장소에서 신청한 책을 받을 수 있는 circulation desk를 운영한다면 지리적인 이유로 대출을 포기했던 책을 대출받을 수 있도록 도와줄 수 있을 것이다.

대출된 도서의 반납 일자 예측 서비스 제공

분석모델 2의 예측 알고리즘 정확도가 상당 수준 이상이 된다면 도서관 서비스 개선에 활용할 수 있다. 도서관 홈페이지에서 검색한 책이 대출중인 경우, 사용자가 원할 때에 한하여 반납 날짜를 예측하는 서비스를 제공하는 방식(예: “이 책은 80%의 확률로 11월 17일에서 21일 사이에 반납될 예정입니다.”)이 유용할 것으로 생각된다.

본 데이터 중심 접근의 한계점

이용자 중심의 도서관 서비스를 만드는 것도 중요한 일이나, 이용자에게 필요한 모든 서비스를 제공하는 것이 도서관의 역할이 되어야 할까? 예를 들어, 해리포터 시리즈 대출을 10명이 예약해놓았다면, 도서관은 해리포터 시리즈를 10권 더 사는 것이 이용자중심의 도서관일까? 어떤 이용자 중심의 필요를 우선시해야할지 순위를 정하고, 도서관이 추구하고자 하는 원칙과 어긋나지 않는지 고민할 필요가 있다.



PART.Ⅳ
연구과제

다이나믹
반응형 탐색적 자료분석

연구배경
연구내용
연구결과
향후계획

다이나믹 반응형 탐색적 자료분석

서울대학교 자연과학대학
박건웅

연구배경

빅데이터의 이해

데이터의 크기, 성격, 정보 등을 고려하여 최적의 데이터 분석 방식을 선택하는 것은 사용자의 편의를 극대화하고 새로운 서비스를 창출할 수 있는 중요한 요소로 작용함. 도서관 대출 데이터에 대한 분석 결과는 기존 도서관 사용자의 이용 패턴을 이해하고, 정보 이용자에게 새로운 행동 양식을 유도하며 나아가 유의미한 정보 및 서비스를 제공할 수 있는 주요한 툴로 활용될 수 있을 것으로 기대됨.

빅데이터 분석은 분석 대상에 대한 최적의 데이터 분석 방법을 선정하고, 데이터 분석을 통해 가치 있는 요소 및 문제를 발굴하는데 의미가 있음. 최적의 데이터 분석 방법을 선정하기 위해서는 해당 데이터에 대한 심도 있는 이해가 선제적으로 필요하며, 데이터의 결합 방식, 크기 및 제공 정보 등을 기본적으로 분석하였음.

연구 방향

통계적 지식이 전혀 없는 사람도 도서관 이용 패턴을 손쉽게 이해할 수 있도록, 직관적으로 이해 가능한 정보 및 서비스 도출하는 것을 방향으로 설정. 특히, 서비스 제공자와 이용자 모두에게 흥미를 유발할 수 있는 탐색적 자료 분석인 (1) 다이내믹 시각화, (2) 반응형 시각화 개발에 집중

연구 방법

- 데이터 융합을 통한 새로운 정보 도출
  - 사람-학과, 사람-과목의 데이터 연계를 통한 새로운 정보를 도출
- 다이내믹/반응형 시각화 구현

- 데이터 시각화는 데이터를 이해하는 가장 효과적인 방법임. 특히 실적을 보여주고, 추세에 대해 커뮤니케이션하며, 새로운 전략의 영향력을 해석하는 등, 주요 행동-결과의 관계를 보여주고자 하는 많은 곳에서 사용되고 있음

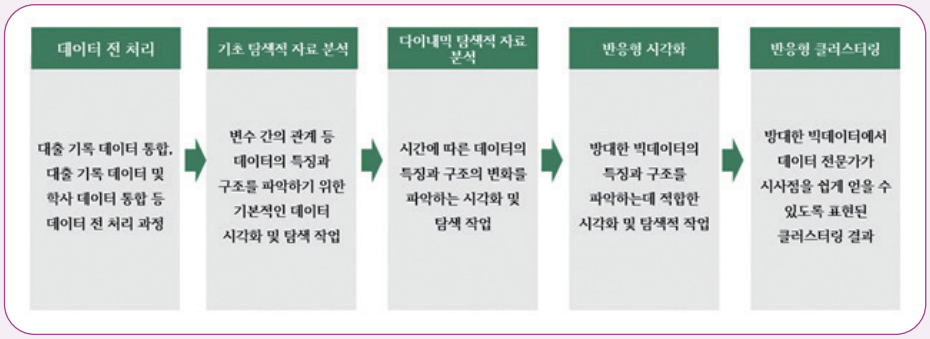
○ 도서관 데이터 시각화 구현 방향

- 다양한 기준으로 이용자들의 도서대출 순위를 보여주는 racing bar chart 구현
- 학과별 시간에 따른 문학과 비문학 도서대출 다이내믹/반응형 시각화 구현

○ 통계적 기계학습 기법을 이용한 데이터 분석(향후)

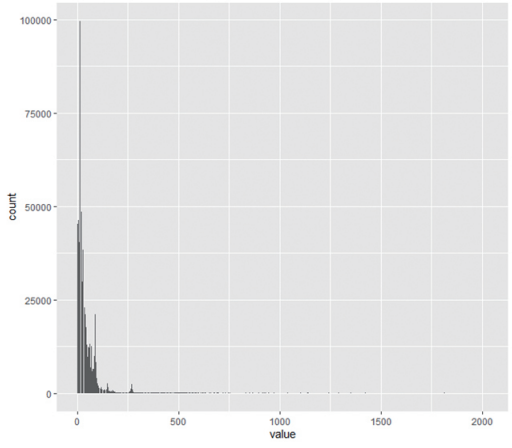
- 향후 연구에서 통계적 기계학습 기법 (예: 반응형 클러스터링)을 활용하여 보다 심도 있는 데이터 분석을 진행할 예정임

연구내용 및 결과



기초 탐색적 자료 분석 (Introductory EDA)

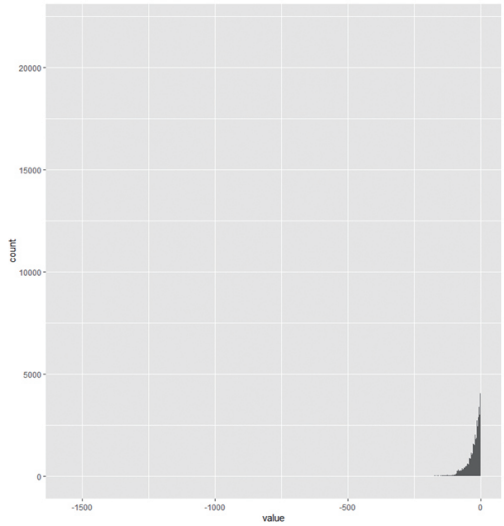
○ 반납날짜 - 대여날짜 그래프



days	count
0	535102
14	99514
7	71165
1	58913
28	58330
13	55535
42	48924
21	48525
15	47635
12	47009

- 0 이하 값들 제거 (음수 존재 288 cases)
- 대여 기간이 긴 책도 있어서 표로 나타내면 다음과 같음
- 14 days 기본 대여 기간, 이후 일주일 단위로 변동 보임

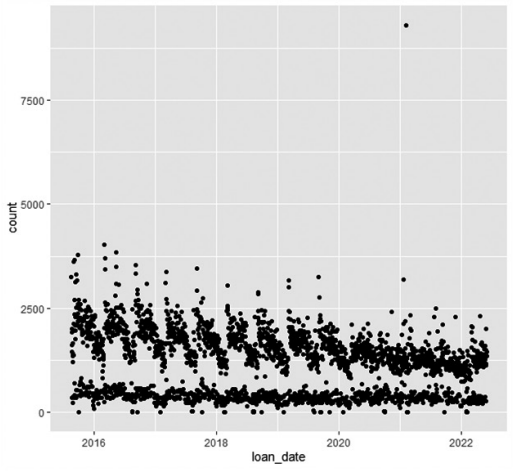
○ 예약 기간 그래프

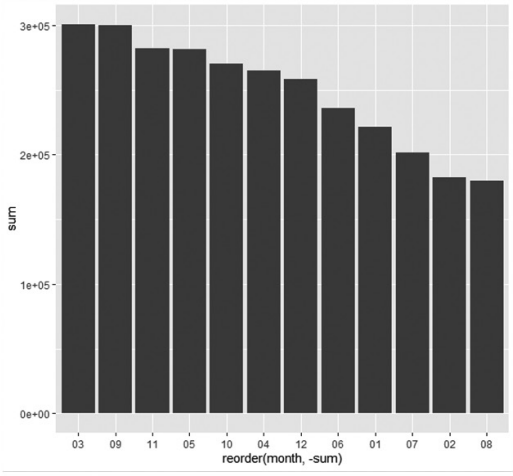


days	count
0	22009
-1	4063
-7	3547
-6	3382
-2	3330
-4	3044
-3	3020
-5	2924
-8	2723
-14	2605

- 당일 예약 제일 많음
- 대체로 예약한 지 일주일 만에 대여

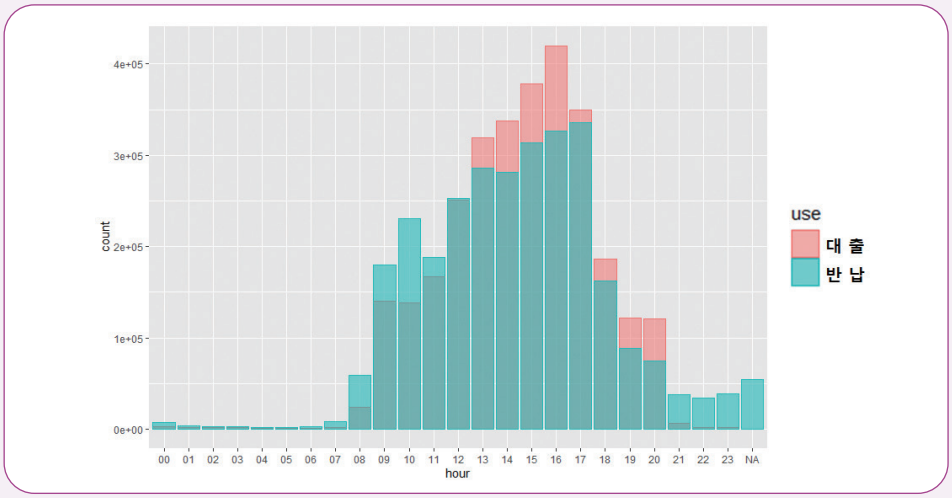
○ 날짜에 따른 도서 대여 건수





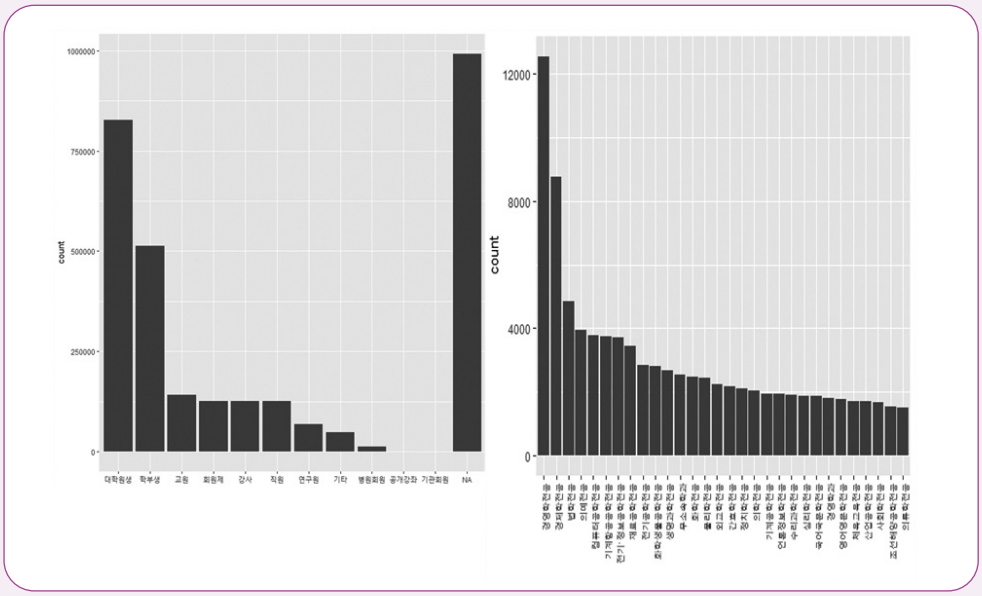
- 날짜에 따라 규칙적인 변동 보임
- 월별 나눠서 살펴보면 개강 달(3, 9), 중간고사 이후(11, 5), 남은 학기(10, 4) 순으로 감소

○ 시간에 따른 도서대출 및 반납



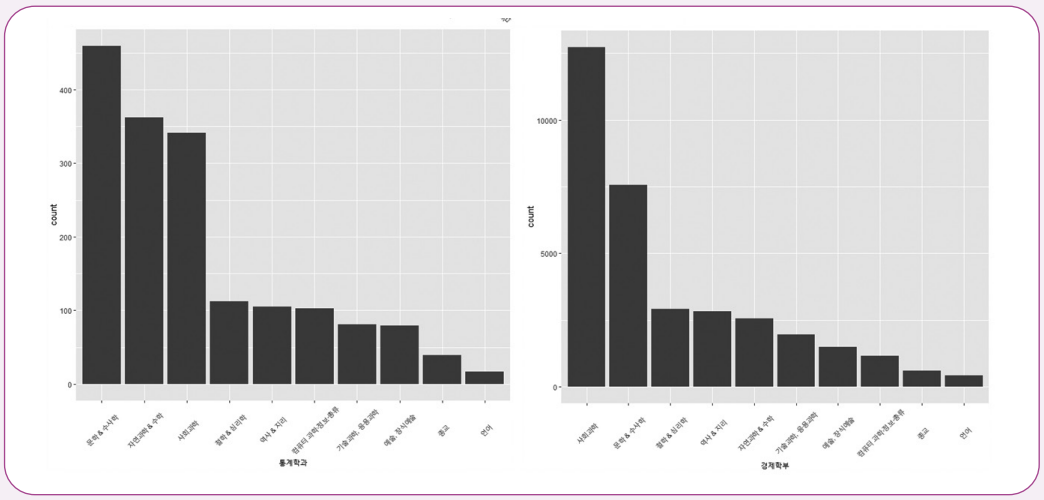
- 대출 시간은 오후에 몰려있고, 반납 시간은 비교적 고르게 퍼져 있음
- 대출과 반납 모두 8AM ~ 8PM에 이루어져 사람들이 학교에 많은 시간대에 주로 이루어지는 것을 알수 있음 하지만 늦은 시간(8PM ~ 11PM)에도 반납이 이루어 지는 것으로 보아 반납은 대출기한 도서대출에 영향 있음을 시사함

○ 사용자 소속에 따른 도서대출



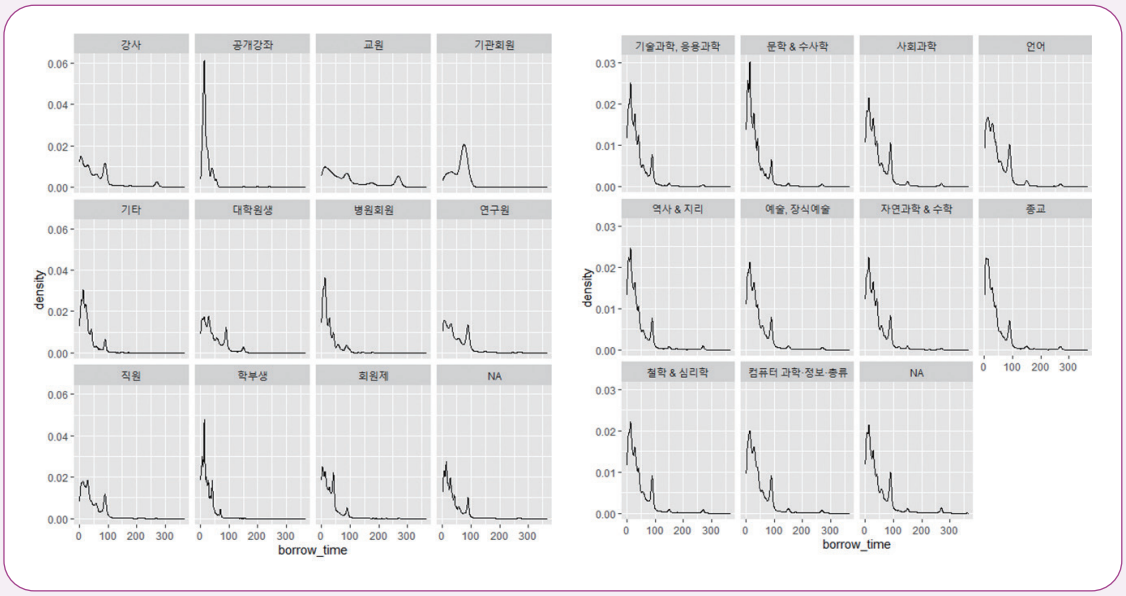
- 대학원생 > 학부생 > 교원 > 회원제 > 강사 > 직원 순으로 도서 이용
- 경영학전공 > 경제학전공 > 법학전공 > 의예 전공 > 컴퓨터공학 순으로 도서 이용
- 결측치 NA와 시간을 고려하지 않아 의미를 찾기 어려움

○ 사용자 소속별 대출 분야



- 통계학과: 문학&수사학 > 자연과학&수학 > 사회과학
- 경제학부: 사회과학 > 문학&수사학 > 철학&심리학
- 사용자 소속 그룹에 따른 책 추천을 달리할 필요성 확인

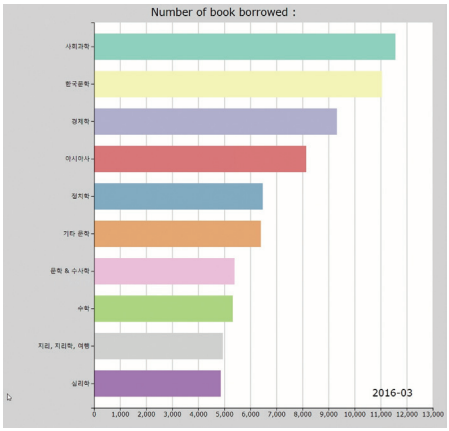
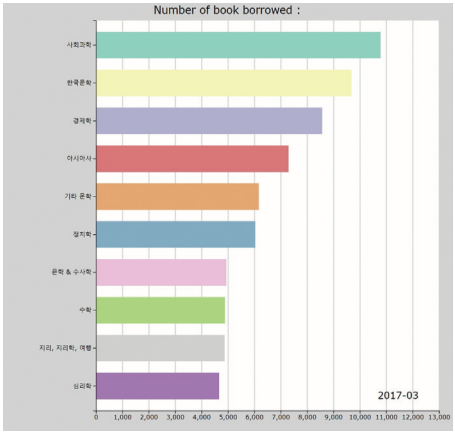
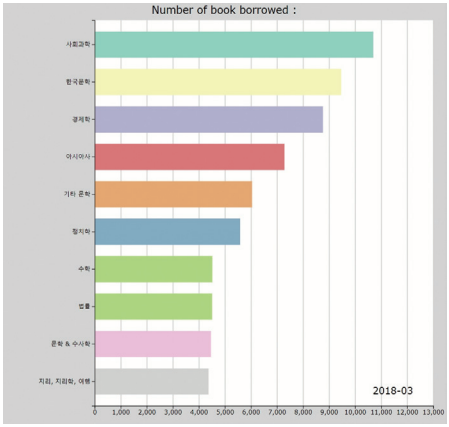
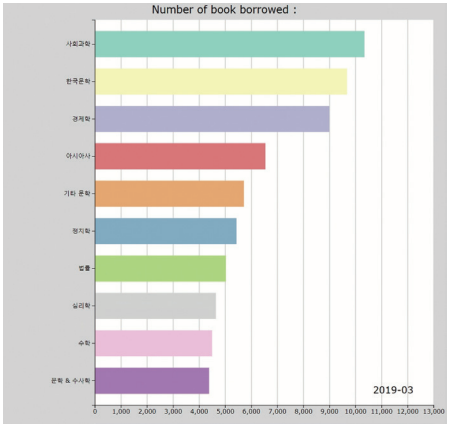
○ 이용자 및 도서 분야에 따른 도서 대여 기간

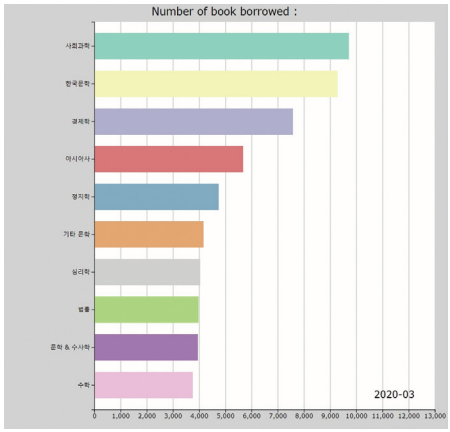
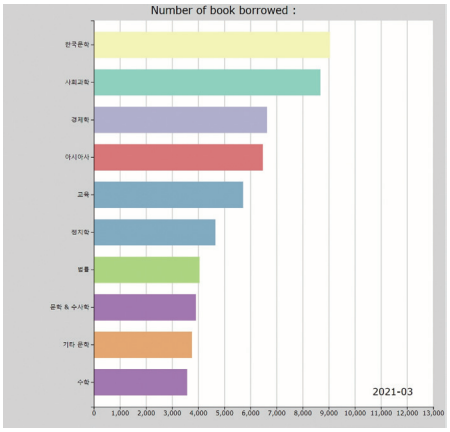


- 대여자 대부분은 책의 대여 기간이 기본적으로 주어지는 14일 대여기간 이내로 짧지만, 일부 강사, 교원과 연구원의 경우 대여연장을 통해 2달 이상씩 기간이 긴 경우가 있음
- 문학&수사학의 경우 책의 대여 기간이 역시 14일 이내로 대부분 짧지만, 언어의 경우 대여 기간이 2달 이상씩 긴 경우가 있음
- 하지만 대부분은 대여기간 패턴이 비슷하며 약 14일 기본 대여 기간 끝나는 날에 반납하는 경우가 많음
- 앞선 8PM - 11PM 반납시간의 패턴과 함께 고려할 때, 책을 실제로 읽는 시간이 짧더라도 여러 가지 이유로 대출 기한에 맞춰서 반납한다고 판단할 수 있음. 이를 기반으로 책 이용이 끝났을 때 바로 반납을 독려하는 제도가 필요하다고 판단됨

다이내믹 탐색적 자료 분석: Racing Bar Chart

○ 분야별 도서대출 순위

〈2016. 3. ~ 2016. 6.〉	〈2017. 3. ~ 2017. 6.〉
	
〈2018. 3. ~ 2018. 6.〉	〈2019. 3. ~ 2019. 6.〉
	

〈2020. 3. ~ 2020. 6.〉	〈2021. 3. ~ 2021. 6.〉
	

○ 2016년~2020년: 사회과학 > 한국문학 > 경제학 > 아시아사 순으로 대여량이 많음

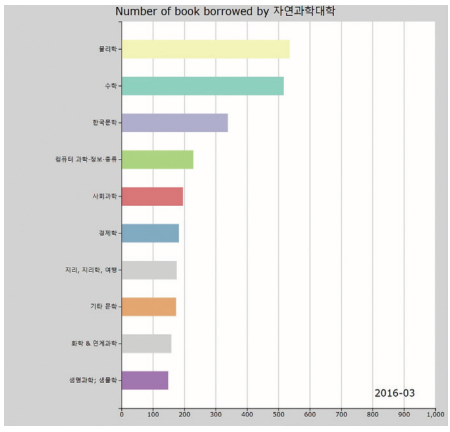
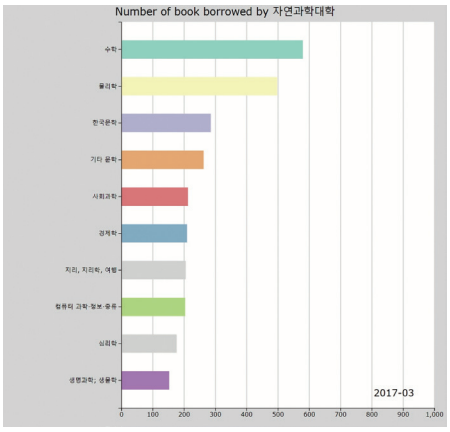
○ 2021년: 한국문학> 사회과학 > 경제학> 아시아사 순으로 대여량이 많아짐
즉, 1등과 2등이 역전됨

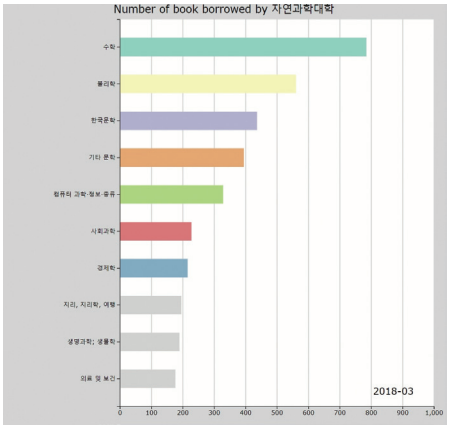
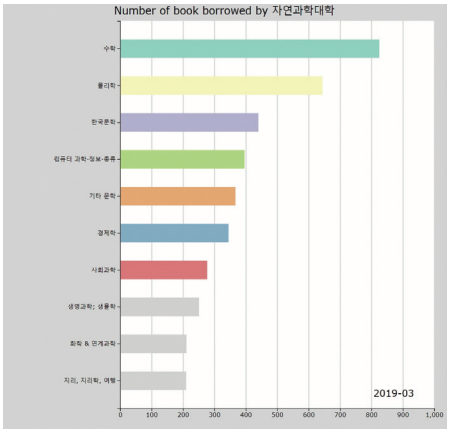
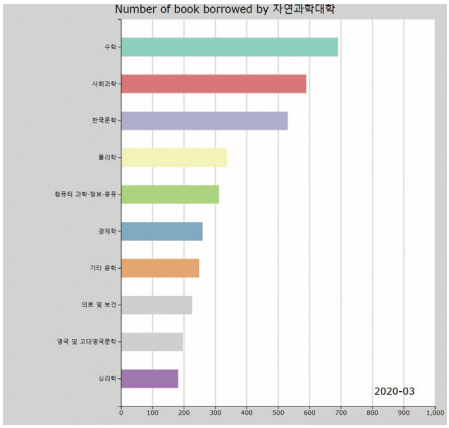
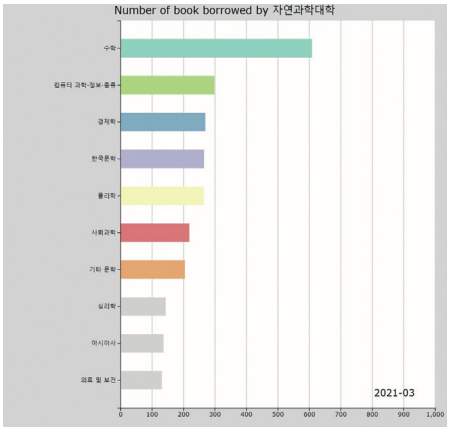
○ 기타문학과 수학은 시간이 갈수록 순위가 많이 떨어지는 경향이 있음

○ 교육과 법률은 시간이 갈수록 순위가 상승하는 경향이 있음

○ 코로나19로 인한 일시적인 영향인지 시대의 흐름인지 원인 파악이 필요

○ 자연과학대학 소속 이용자의 도서 분야별 순위

〈2016. 3. ~ 2016. 6.〉	〈2017. 3. ~ 2017. 6.〉
	

〈2018. 3. ~ 2018. 6.〉	〈2019. 3. ~ 2019. 6.〉
	
〈2020. 3. ~ 2020. 6.〉	〈2021. 3. ~ 2021. 6.〉
	

○ 2016년: 물리학 〉 수학 〉 한국문학 분야 도서대출 순위를 보임

○ 2017년~2019년: 수학 〉 물리학 〉 한국문학 분야 순으로 수학 관련 도서는 꾸준히 대출량이 늘어나지만, 물리학은 감소하는 경향을 보임

○ 2020년: 수학 〉 사회과학 〉 한국문학 순으로 앞서 설명한 경향이 더욱 심해짐

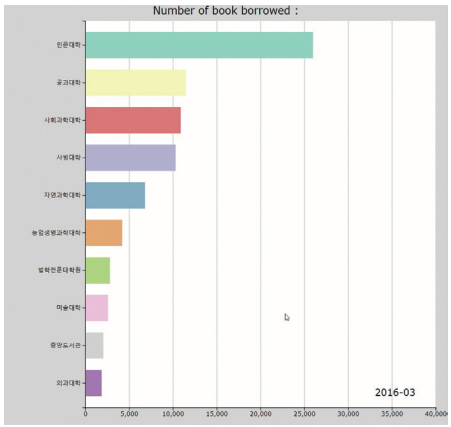
○ 사회과학 분야 도서대출량을 2020년 급격한 순위 상승을 보임

○ 컴퓨터 과학 분야 도서대출량은 2021년에 급격한 순위 상승이 있음

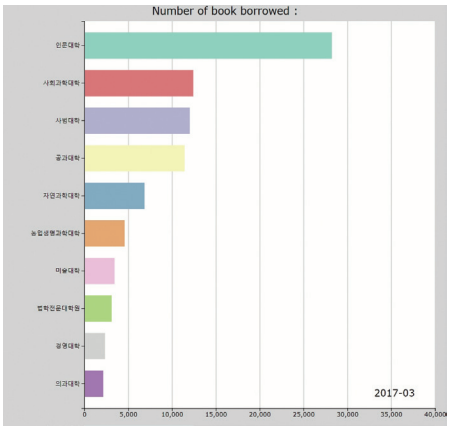
○ 현재는 3월~6월 경향을 분석했지만, 다른 시기 순위 및 대출량을 볼 필요성 있음

○ 단과대학별 도서대출량 순위

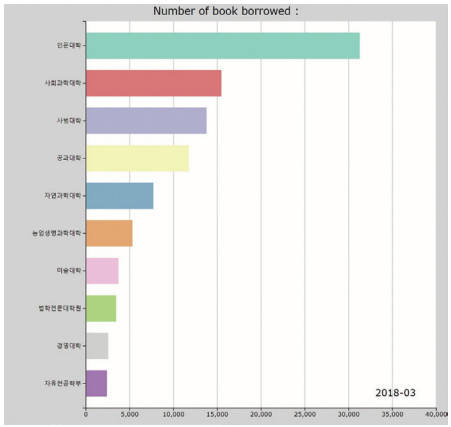
〈2016. 3. ~ 2016. 6.〉



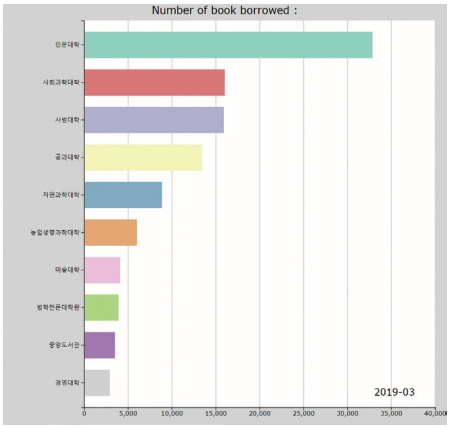
〈2017. 3. ~ 2017. 6.〉



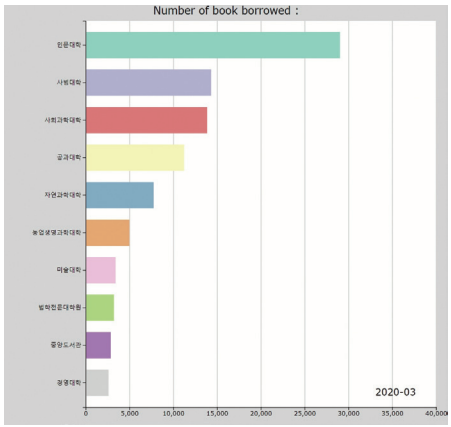
〈2018. 3. ~ 2018. 6.〉



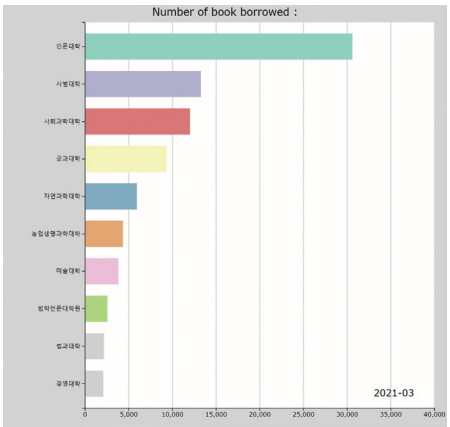
〈2019. 3. ~ 2019. 6.〉



〈2020. 3. ~ 2020. 6.〉

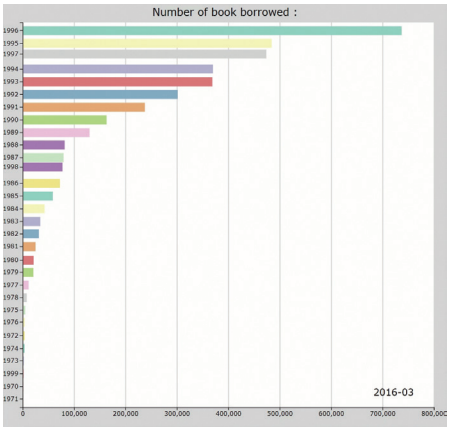
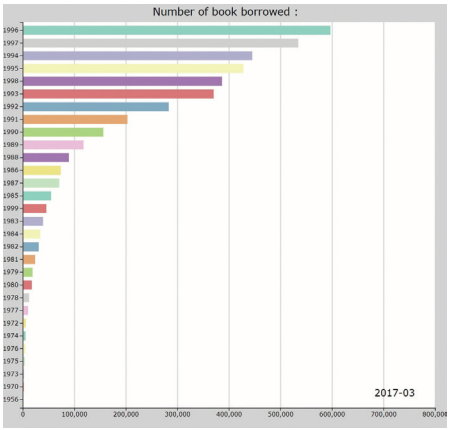
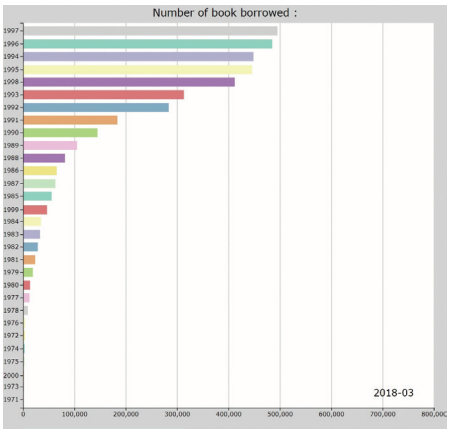
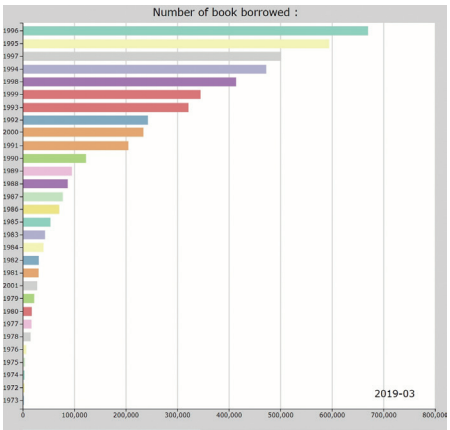


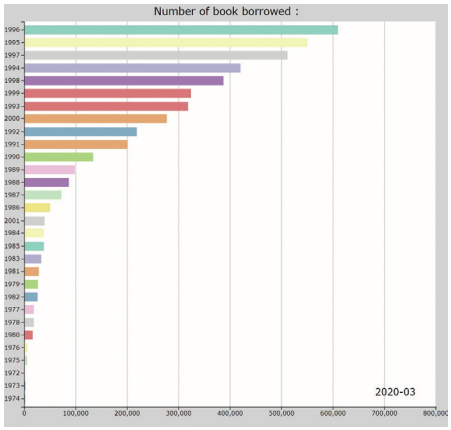
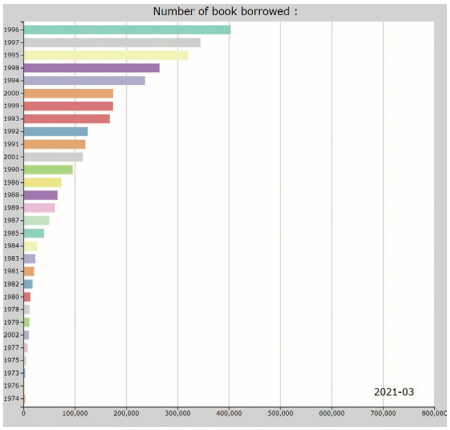
〈2021. 3. ~ 2021. 6.〉



○ 2016년: 인문대학> 공과대학 > 사회과학대학 > 사범대학
○ 2017~2019년: 인문대학 > 사회과학대학 > 사범대학 > 공과대학
○ 2020~2021년: 인문대학 > 사범대학 > 사회과학대학 > 공과대
○ 사범대학의 대출량이 꾸준히 상승하고 있으며, 공과대학은 꾸준한 감소세

○ 나이/출생연도별 순위

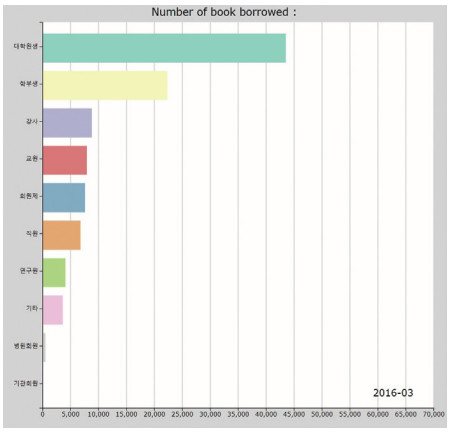
〈2016. 3. ~ 2016. 6.〉	〈2017. 3. ~ 2017. 6.〉
	
〈2018. 3. ~ 2018. 6.〉	〈2019. 3. ~ 2019. 6.〉
	

〈2020. 3. ~ 2020. 6.〉	〈2021. 3. ~ 2021. 6.〉
	

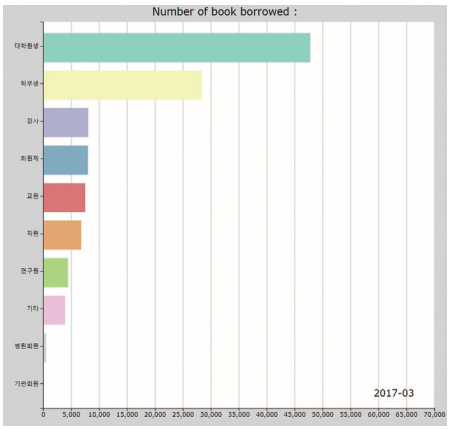
- 2016년: 1996년 > 1995년 > 1997년 > 1994년 > 1993년 > 1992년
- 2017년: 1996년 > 1997년 > 1994년 > 1995년 > 1998년 > 1993년
- 2018년: 1997년 > 1996년 > 1994년 > 1995년 > 1998년 > 1993년
- 2019년: 1996년 > 1995년 > 1997년 > 1994년 > 1998년 > 1999년
- 2020년: 1996년 > 1995년 > 1997년 > 1994년 > 1998년 > 1999년
- 2021년: 1996년 > 1997년 > 1995년 > 1998년 > 1994년 > 2000년
- 1996년생의 도서대출량이 제일 많음
- 1995년생의 경우 2016년 이후 도서대출량이 꾸준히 감소
- 1997년생의 경우 2016년 ~2018년까지 대출 순위가 꾸준히 상승 후 2019년부터는 상위 자리매김
- 1998년생의 경우 앞선 95, 96, 97년생의 패턴을 따르지 않고 5위권 자리매김
- 1999년생의 경우에도 1998년생과 마찬가지로 하위권 자리매김
- 1998, 1999년생의 패턴이 코로나19로 인한 일시적 경향인지, 장기적인 흐름인지는 살펴보아야 함

○ 그룹에 따른 총대출량 순위

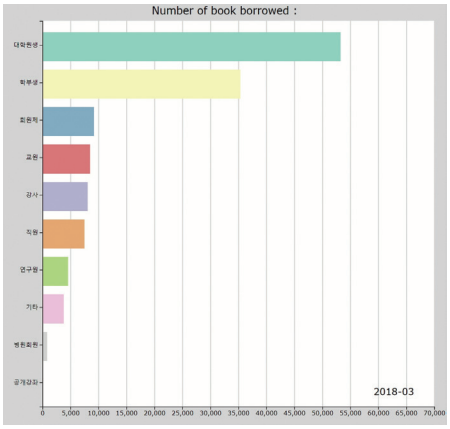
〈2016. 3. ~ 2016. 6.〉



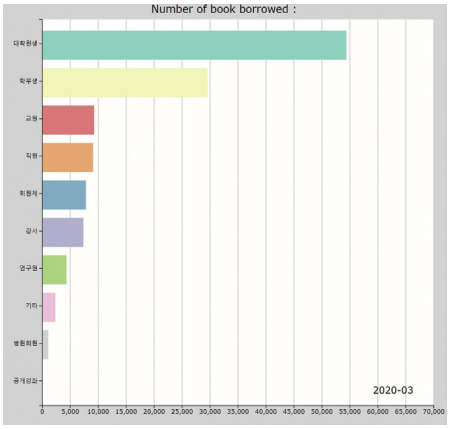
〈2017. 3. ~ 2017. 6.〉



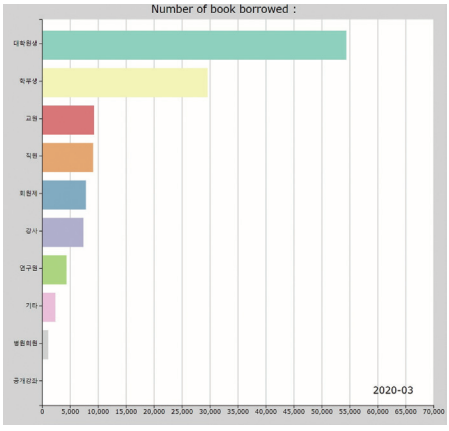
〈2018. 3. ~ 2018. 6.〉



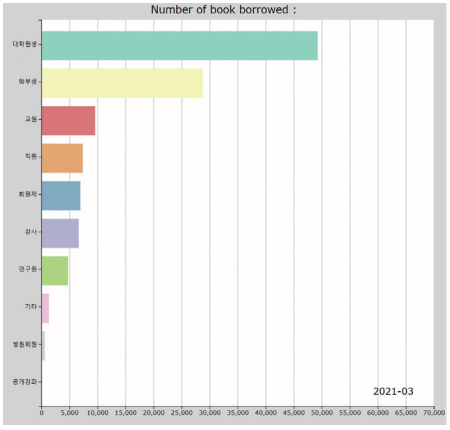
〈2019. 3. ~ 2019. 6.〉



〈2020. 3. ~ 2020. 6.〉

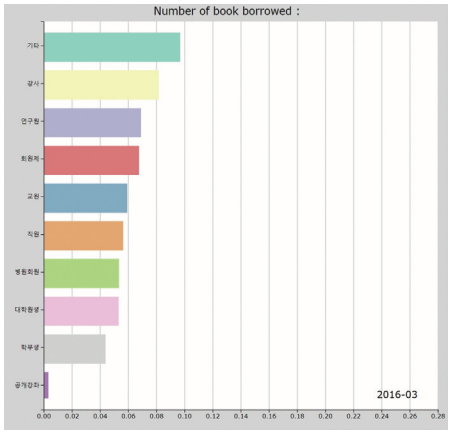
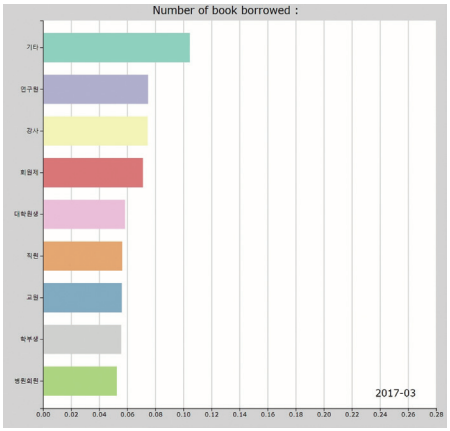
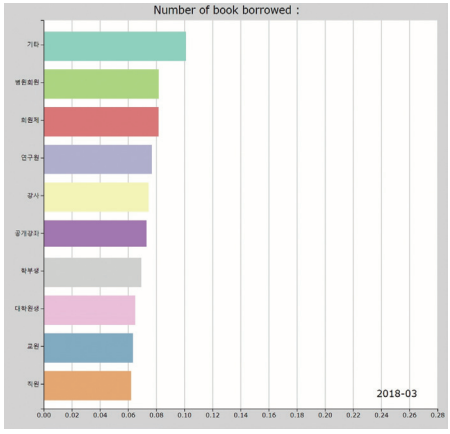
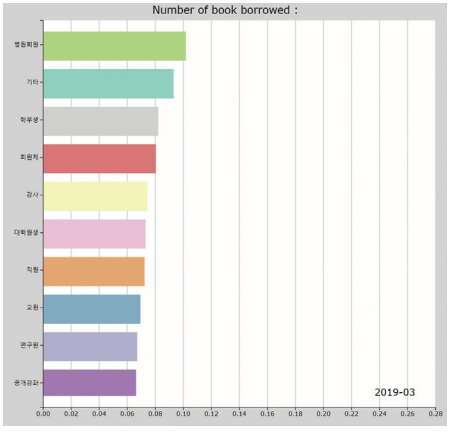


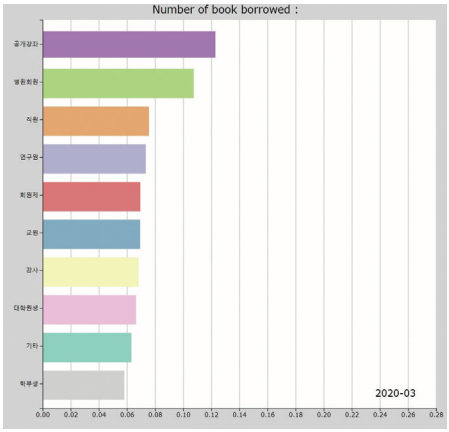
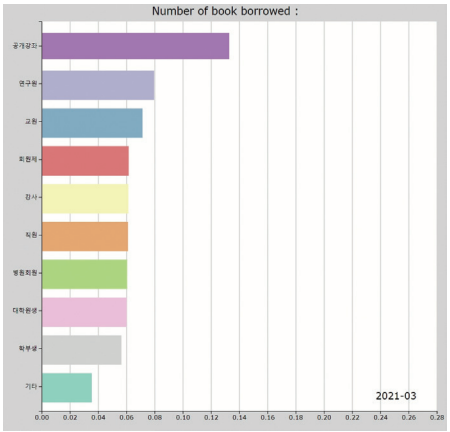
〈2021. 3. ~ 2021. 6.〉



○ 2016년: 대학원생 > 학부생 > 강사 > 교원 > 회원제
○ 2017년: 대학원생 > 학부생 > 강사 > 회원제 > 교원
○ 2018년: 대학원생 > 학부생 > 회원제 > 교원 > 강사
○ 2019년: 대학원생 > 학부생 > 교원 > 회원제 > 직원
○ 2020-21년: 대학원생 > 학부생 > 교원 > 직원 > 회원제
○ 대학원생과 학부생 1, 2위는 변화 없음
○ 강사의 대출 순위가 떨어지는 추세

○ 그룹에 따른 1인당 대출량 순위

〈2016. 3. ~ 2016. 6.〉	〈2017. 3. ~ 2017. 6.〉
	
〈2018. 3. ~ 2018. 6.〉	〈2019. 3. ~ 2019. 6.〉
	

〈2020. 3. ~ 2020. 6.〉	〈2021. 3. ~ 2021. 6.〉
	

○ 2016년: 기타> 강사> 연구원 > 회원제 > 교원 > 직원

○ 2017년: 기타> 연구원 > 강사 > 회원제 > 대학원생

○ 2018년: 기타> 병원회원 > 회원제 > 연구원 > 강사

○ 2019년: 병원회원 > 기타 > 학부생 > 회원제 > 강사

○ 2020년: 공개강좌 > 병원회원 > 직원 > 연구원 > 회원제

○ 2021년: 공개강좌 > 연구원 > 교원 > 회원제 > 강사

○ 1인당 도서대출은 1, 2위를 제외하고는 매년 비슷한 수준

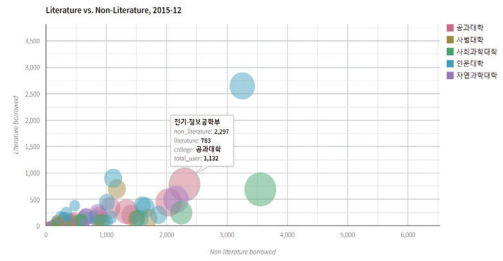
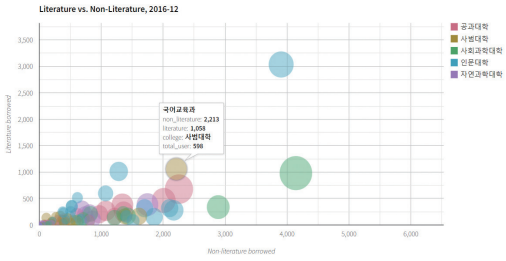
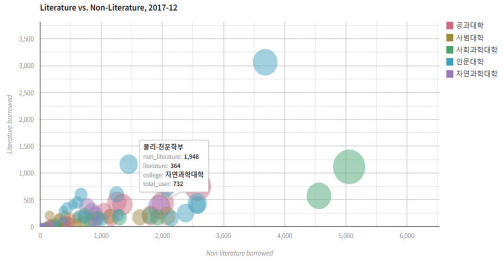
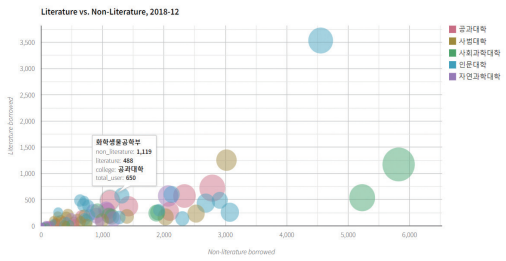
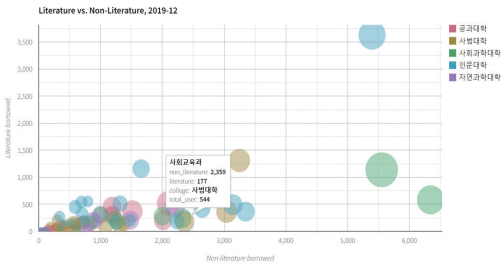
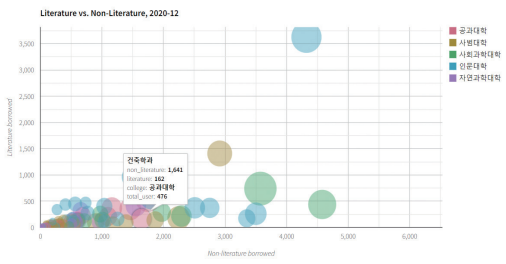
○ 기타회원은 수가 너무 적어서 신뢰할 수 없음

○ 2020년 이후 공개강좌 이용자들의 1인당 평균 대출량이 급증함

○ 학부생 회원의 경우 인원이 많아 총대출량은 많지만 1인당 평균 대출량은 적음

반응형 시각화(Google Chart)

○ 학과별 문학과 비문학 대여 관계: 대출량

〈2015. 6. ~ 2015. 12.〉	〈2016. 6. ~ 2016. 12.〉
	
〈2017. 6. ~ 2017. 12.〉	〈2018. 6. ~ 2018. 12.〉
	
〈2019. 6. ~ 2019. 12.〉	〈2020. 6. ~ 2020. 12.〉
	

○ 대출량 상위 5: 공과대학, 사범대학, 사회과학대학, 인문대학, 자연과학대학

○ 개별 과마다 도서관 이용 인원을 원의 크기로 표현하였으며 같은 단과대학은 같은 색으로 표현

○ 같은 단과대에 속해있더라도 도서대출량이 매우 다르며, 문학과 비문학 대출 비율도 매우 다름

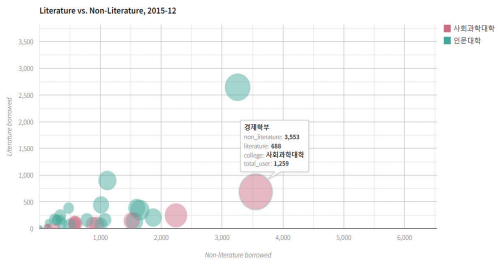
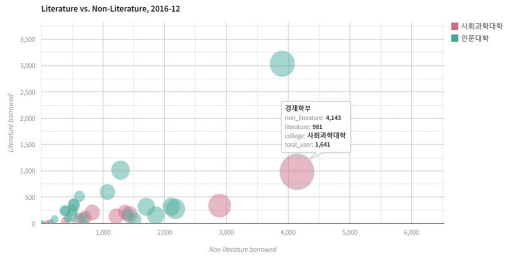
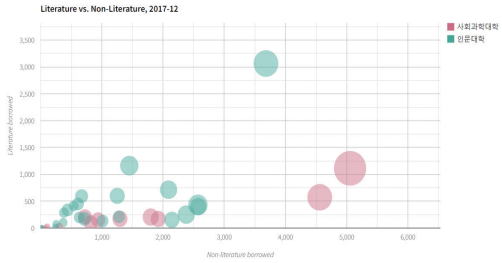
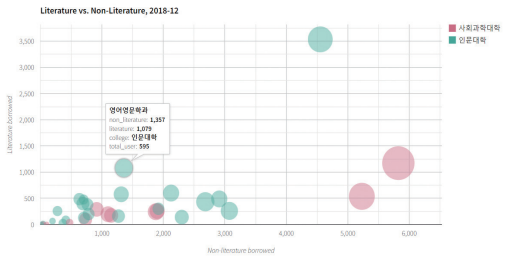
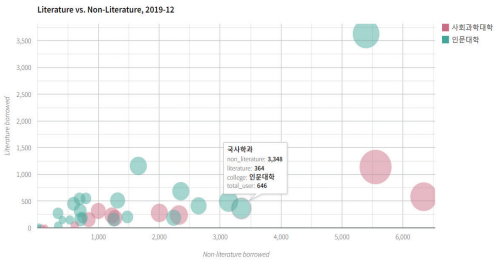
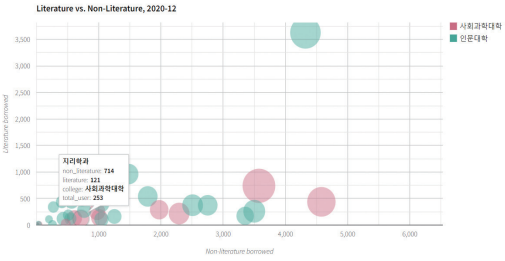
○ 대부분 비문학 도서대출량이 문학 도서대출량보다 현저히 많음

○ 인문대학의 경우 문학과 비문학 비율이 타 대학보다 유의미하게 높은 경향이 있음

○ 사회과학대학원의 경우 문학과 비문학 비율이 타 대학보다 유의미하게 낮은 경향이 있음

○ 시간이 지남에 따라 문학 도서대출량이 증가하였으나 문학/비문학 도서대출 비율은 일정하게 유지되는 경향이 있음

○ 학과별 문학과 비문학 대여 관계: 사회과학대학 vs 인문대학

〈2015. 6. ~ 2015. 12.〉	〈2016. 6. ~ 2016. 12.〉
	
〈2017. 6. ~ 2017. 12.〉	〈2018. 6. ~ 2018. 12.〉
	
〈2019. 6. ~ 2019. 12.〉	〈2020. 6. ~ 2020. 12.〉
	

○ 문학/비문학 도서대출 비율이 다르게 보이는 인문대학과 사회과학대학만 한정하여 분석

○ 인문대학의 경우 문학/비문학 도서대출 비율이 높은 학과와 사회과학대학과 비슷한 비율을 보이는 학과 두 개의 클러스터가 있는 것으로 보임

○ 비율이 높은 학과: 독어독문학과, 불어불문학과, 중어중문학과, 영어영문학과, 국어국문학과

○ 비율이 낮은 학과: 아시아언어문명학부, 서양사학과, 종교학과, 동양사학과, 미학과, 고고미술사학과 국사학

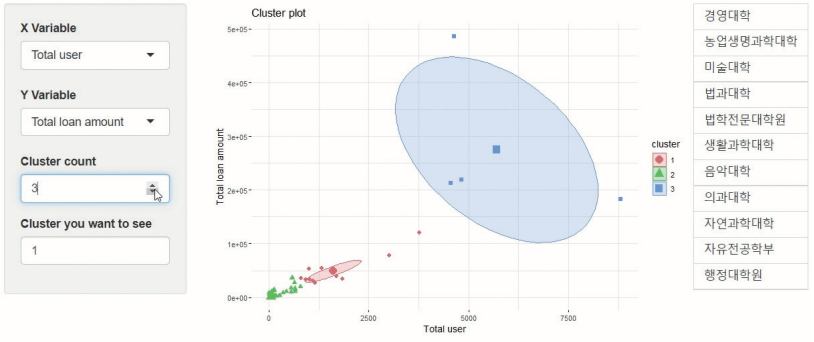
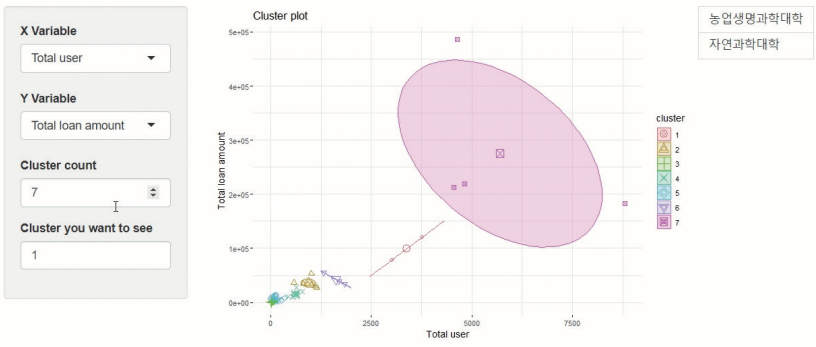
○ 코로나19 이후 문학 도서대출량보다 비문학 도서대출량이 더 많이 감소한 것을 확인유지되는 경향이 있음

반응형 클러스터링

- 두 가지 변수를 이용한 이용자 소속(e.g. 학과) 군집화 분석
- 방법: K-means 클러스터링
- 사용 변수

필드명	설명	비고
Total user	총 이용자 수	이용기준: 도서대출
Reuse rate	재이용률	Repeat user / Total user
Total loan amount	총 도서대출량	
Average age	총 이용자 나이의 평균	
Average rental period	평균 대출 기간	
College	소속대학	

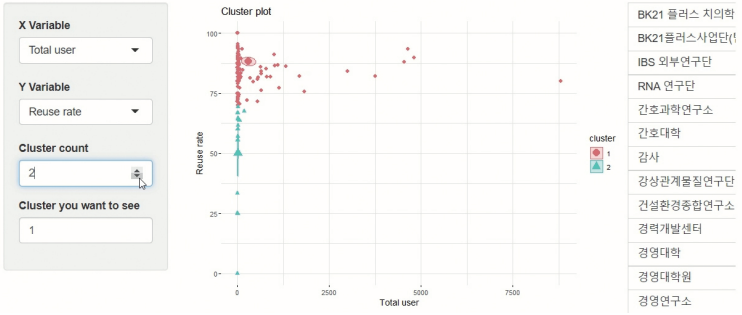
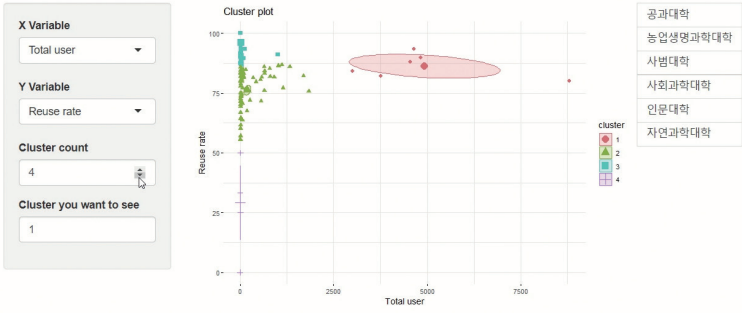
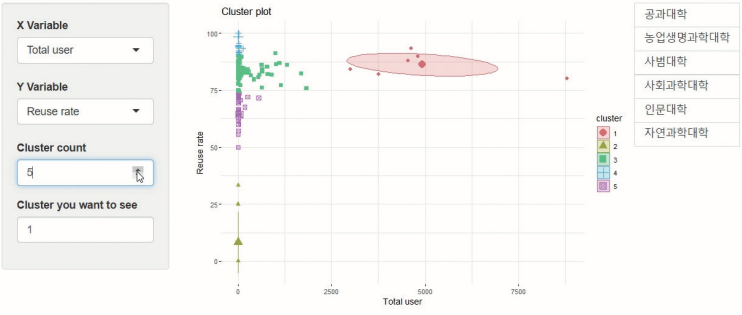
○ 총 이용자 수와 총 도서대출량 변수를 이용한 K-means 클러스터링 결과

클러스터	총이용자 수 vs 총 도서대출량
3	
7	

○ 총 이용자 수와 총 도서대출량이 강한 양의 상관관계를 가지고 있으며, 클러스터링 결과 두 변수 비슷한 수준의 값을 가진 소속대학들끼리 그룹화 함

○ 그룹의 수를 늘려도 비슷한 성격의 그룹만 확인됨. 일부 이 상치의 값을 가진 소속대학들도 관찰됨

○ 총 이용자 수와 재이용률을 이용한 K-means 클러스터링 결과

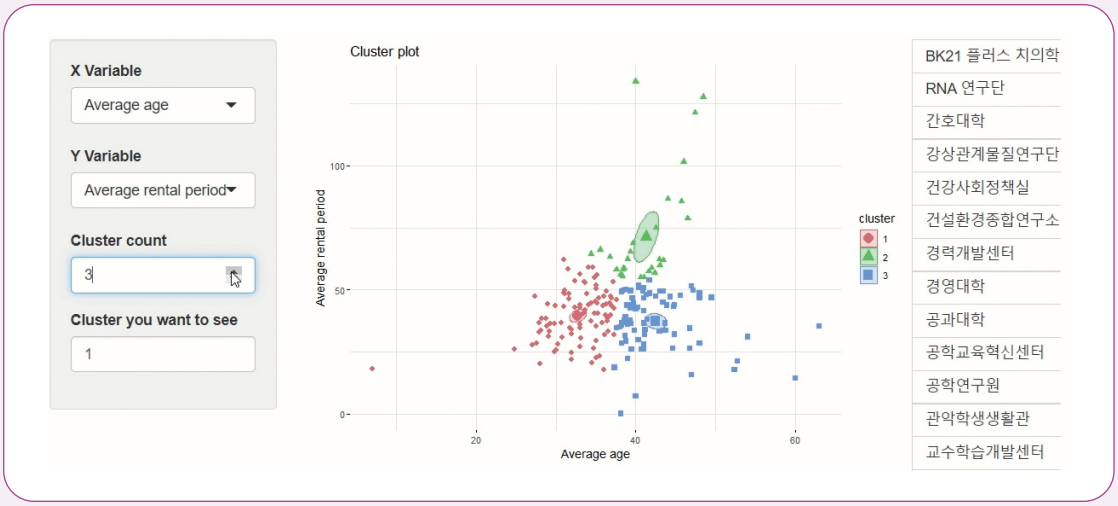
클러스터	총이용자 수 vs 재이용률
2	
3	
4	
5	

○ 총이용자 수가 많은 그룹은 재이용률이 75% 이상으로 높음

○ 하지만 총이용자 수가 작은 그룹의 경우 재이용률의 차이가 큼. 이 경우 소수의 인원이 재이용률에 큰 영향을 미칠 수 있어서 신뢰하기 어려움

○ K-means 클러스터링의 경우 그룹의 수를 바꾸어도 재이용률을 기준으로 그룹이 선택되는 경향이 있음

○ 이용자 나이 평균과 평균 대출 기간을 이용한 K-means 클러스터링 결과



○ 이용자 나이와 대출 기간은 큰 상관관계를 찾기는 어려움

○ 클러스터링 결과 나이가 적은 그룹 1, 나이는 많지만, 평균 대출 기간이 긴 그룹 2, 나이는 많지만, 평균 대출 기간이 짧은 그룹 3으로 나누어짐

향후계획

탐색적 자료 분석을 바탕으로 양질의 데이터 분석 방향 제공

- 시계열 데이터 분석 방법을 이용한 패턴 분석
- 분야별 책 대여량의 코로나19 영향력 파악

- 학과별 책 대여량의 코로나19 영향력 파악
- 패턴 분석을 이용하여 특정과의 도서대출량이 늘어난 혹은 감소한 원인 파악
- 다양한 기준으로 클러스터링 분석

⚙ 더욱 다양한 데이터 병합을 이용한 탐색적 자료 분석

- 현재는 문학과 비문학 간 총대출량의 상관관계를 단과대학/학과별로 살펴보고 있지만 이를 다른 변수들 고려 가능
- 허구와 실화 간 주제별 총대출량의 상관관계를 단과대학/학과별 분석
- 한국문학과 비 한국문학 총대출량의 상관관계를 단과대학/학과별 분석
- 문학과 비문학 간 총대출량의 상관관계를 나이별 분석
- 문학과 비문학 간 총대출량의 상관관계를 이용자 그룹별 분석

⚙ 잠재고객들의 관심을 유발하는 다양한 탐색적 자료 분석

- 빅데이터 기초통계 분석 사례: 배민트렌드 2022

⚙ 개인정보를 유출하지 않도록 데이터 마스킹 기법을 도입하여 도서관 데이터를 공개 데이터로 전환하고 연구에 자유롭게 활용할 수 있는 방법 모색

부록

연구에 활용한 데이터 테이블 스키마

제공 테이블 : 총 10개

테이블 자료구조 및 필드속성

부록
연구에 활용한 데이터 테이블 스키마

제공 테이블 : 총 10개

구분	필드명	설명
대출 데이터	loans	대출 트랜잭션에 대한 필수 정보
	users	대출자 정보
	detail	대출 트랜잭션에 대한 상세한 정보
학사 데이터	h_academic	학적데이터(성별, 학위과정, 입학일, 학년-현재기준)
	h_course	수강데이터
	h_course_name	수강데이터(과목명)
	h_course_department	수강데이터(개설학과)
	h_graduation	졸업일
	h_major	전공데이터
대출 전처리데이터	vbook_refine	loans의 도서 정보를 주제분야 기준으로 전처리한 테이블
	vbook_user	user의 신분 범주를 전처리한 테이블

※ 대출자 정보는 비식별처리하여 제공

테이블 자료구조 및 필드속성

1.loans

필드명	설명	비고
loan date	대출일	
borrower	대출자 ID	비식별화
item_loan_id	대출 트랜잭션별 고유 ID	
mms id	콘텐츠ID	
physical_item_id	아이템ID	
title	서명	
author	저자	

publisher	출판사	
isbn	국제표준도서 번호	
publication_date	발행일	
receiving date (calendar)	도서관에서 해당 자료를 입수한 날짜	
subjects	주제분야	

2.users

필드명	설명	비고
borrower	대출자 ID	비식별화
birth year	생년	
user group	신분	
college	소속대학	
department	소속학과	

3.detail

필드명	설명	비고
item_loan_id	대출 고유 ID	
mms_id	콘텐츠 ID	
item_id	아이템 ID	
barcode	바코드	
call_number	도서 청구 기호	
loan_date	대출일	
loan_time	대출시간	
due_datetime	반납기한	
location_code	서고위치	
new_due_date	갱신으로 연장된 반납 예정일	
new_due_time	갱신으로 연장된 반납 예정시간	
original_due_date	최초반납예정일	
original_due_time	최초반납예정시간	
publication_year	출판년	
recall_date	예약일	
recall_time	예약시간	
return_date	반납일	
return_time	반납시간	
title	제목	

campus_code	대출도서관 캠퍼스 위치	
library_name	대출도서관 명칭	
library_code	대출도서관 코드	
library_desc	대출도서관 설명	
circ_desk_code	대출데스크 코드	
circ_desk_desc	대출데스크 설명	

4.h_academic

필드명	설명	비고
borrower	대출자 ID	비식별화
gender	대출자 성별	
degree	학위명	
grade_detail	수학상태_상세	
grade	학년	
admission_date	입학일	

5.h_course

필드명	설명	비고
year	개설년도	
semester	개설학기	
semester_detail	개설학기구분 상세	
borrower	대출자	
code	교과목코드	
code_detail	교과목종류	
dept_kor_nm	개설학과	

6.h_course_name

필드명	설명	비고
sbjt_code	교과목코드	
sbjt_nm	교과목국문명	
sbjt_eng_nm	교과목영문명	
sbjt_smry_ctnt	교과목개요국문	
sbjt_smry_eng_ctnt	교과목개요영문	
hash_tag_kor_ctnt	교과목해시태그	

7.graduation

필드명	설명	비고
borrower	대출자ID	비식별화
degree	학위명	
graduation_date	졸업일	

8.h_major

필드명	설명	비고
borrower	대출자ID	비식별화
year	전공신청년도	
semester	전공신청학기	
category	전공구분	
major_name	전공명칭	

9.vbook_refine

필드명	설명	비고
mms_id	콘텐츠 ID	
title	제목	
author	저자명	
call_number	청구기호	
section	청구기호 분류 - 류	
division	청구기호 분류 - 강	
category	청구기호 분류 - 목	
section_desc	청구기호 분류 - 류 - 주제분야	
division_desc	청구기호 분류 - 강- 주제분야	
category_desc	청구기호 분류 - 목 - 주제분야	

10.user_refine

필드명	설명	비고
user_id	대출자 ID	비식별화
group	신분	신분을 간소화함
college	소속대학	
department	소속학과	

서울대학교 중앙도서관
도서대출 빅데이터 정책연구과제 결과보고서

발 행 처 서울대학교 중앙도서관
발 행 인 장덕진(서울대학교 중앙도서관장)
참여연구원 장덕진, 김란우, 김홍기, 박건웅, 윤진혁, 이준환
주 소 08826 서울시 관악구 관악로 1
전 화 (02) 880-5438
팩 스 (02) 880-2730
E-mail libit@snu.ac.kr
홈 페 이 지 http://library.snu.ac.kr
인 쇄 북팟