K-MOOC
한국형온라인공개강좌

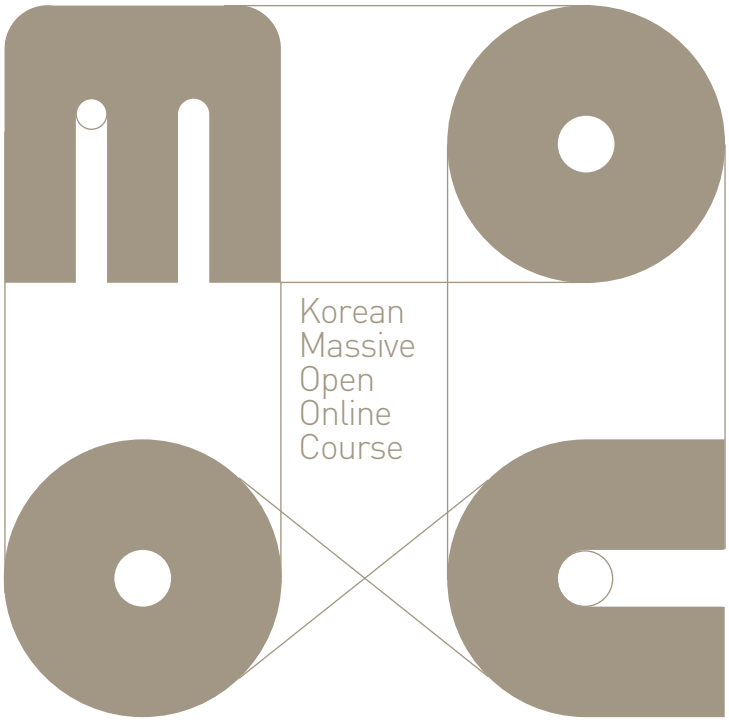
연구자료 OR 2019-2

2018 이슈페이퍼

K-MOOC 학습데이터 분석 및 활용방향 탐색

An Analysis of K-MOOC Learners' Data and an Investigation of Its Future Applications

연구책임자: 김주호 교수 (한국과학기술원)
참여연구원: 김서영 (한국과학기술원, 박사과정)
권순우 (한국과학기술원/중앙대학교, 학사과정)
신동훈 (서울대학교, 학사과정)



국가평생교육진흥원
National Institute for Lifelong Education

K-MOOC 학습데이터 분석 및 활용방향 탐색

An Analysis of K-MOOC Learners' Data and an Investigation of Its Future Applications

연구책임자: 김주호 교수 (한국과학기술원)

참여연구원: 김서영 (한국과학기술원, 박사과정)

권순우 (한국과학기술원/중앙대학교, 학사과정)

신동훈 (서울대학교, 학사과정)

연구 요약

MOOC(Massive Open Online Course)는 평생학습, 열린 교육, 그리고 상호유대라는 가치 아래 널리 활용되고 있다. 하지만 MOOC의 비대면 교육이라는 특징으로 인해 교수자는 학습자의 학습 패턴을 기존의 오프라인 교육에 대비하여 파악하기 힘들다. 이러한 단점을 극복하기 위해 학습자가 MOOC 플랫폼을 활용할 때 축적되는 로그를 분석하는 여러 시도가 있었다. 하지만 K-MOOC 플랫폼의 경우 증가하는 강좌와 학습자 수에 비해 학습자 로그 데이터 분석이 부족한 상태이다. 따라서 본 연구에서는 K-MOOC의 2018년 학습자 로그를 바탕으로 (1) 전체 데이터의 전반적인 이해, (2) 전체 데이터를 기반으로 한 지표 분석, 그리고 (3) 한 개의 강좌를 대상으로 사례 분석을 진행하였다. 학습자, 교수자, 연구자 및 기관은 본 연구에서 진행된 학습자 로그 데이터 분석 결과를 활용하여 추후 더 나은 K-MOOC 플랫폼과 강좌를 설계할 수 있다.

차례

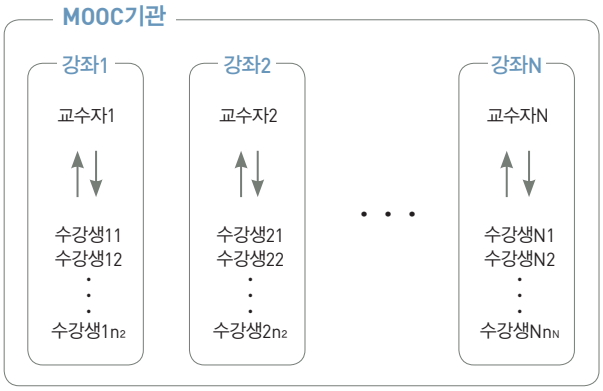
1 연구 문제

1.1 연구의 배경

2008년 Prince Edward Island 대학교의 Dave Cormier에 의해 명명된 MOOC(Massive Open Online Course)는, 평생학습, 열린 교육, 그리고 상호유대라는 가치 아래 널리 확산되고 있다. MOOC는 2010년대 초반부터 Stanford University(Stanford Online), MIT(MITx), Harvard University(HarvardX) 등 대학을 중심으로 제공되었고, Udacity, Coursera 등 영리 목적의 MOOC 제공업체도 생겨나 다양한 강좌를 제공하고 있다.

또한, MOOC는 비단 한 국가의 현상으로 머무르는 것이 아닌, 전세계적인 트렌드가 되고 있다. 많은 국가에서 국가, 대학, 또는 기업을 중심으로 MOOC 플랫폼을 개발하였고, 현재 중국 칭화대학교의 XuetangX, 일본의 J-MOOC, 영국의 FutureLearn 등 다양한 플랫폼이 개발되어 각국에서 MOOC 강좌들을 제공하고 있다.

우리나라의 경우, 2015년 10월 국가주도로 한국형 온라인 공개강좌 'K-MOOC(www.kmooc.kr)'이 시범 개통된 이래로 현재까지 많은 학습자들이 이용하고 있다. 2019년 1월 현재 92개 대학에서 인문, 사회, 공학, 자연과학 등 다양한 분야에 걸친 520여개 강좌를 개설하여 제공하고 있는데. 이에 2018년 12월을 기준으로 가입자 36만 명, 수강신청 78만 명을 달성하는 등 학습자들의 관심이 지속적으로 증가하는 추세이다.



[그림 1] K-MOOC 플랫폼의 구성 요소

이와 같이 거대하고(Massive), 수강인원의 제한이 없는(Open), 온라인을 통한(Online) MOOC는 [그림 1]과 같은 체계를 가지고 있다. 즉, 강좌의 단위는 교수자와 다수의 수강생으로 이루어져 있으며, 단일 수강생이 여러 강좌를 수강하는 것이 가능하다. 또한 오픈소스로 공개된 edX 기반으로 된 MOOC 플랫폼(K-MOOC, MITx, HarvardX, XuetangX 등)에는 학습자 정보, 강좌 정보, 학습자가 특정 강좌를 들을 때의 event 로그 등이 저장되는데, 이러한 방대한 로그데이터를 이용하면 수강자의 행동을 분석할 수 있다. 이를 통해 MOOC 플랫폼과 강좌를 개선할 수 있기 때문에 MOOC에서는 양적 연구의 필요성이 대두되고 있다. 이미 HarvardX, MITx, XuetangX 등의 해외 유명 MOOC 기관에서는 학습자 로그 데이터를 기반으로 한 여러 양적 분석을 하고 있다 (Ho et al., 2014). 특히 ACM SIGKDD가 후원하는 데이터마이닝 콘테스트인 KDD CUP 2015에서는, 중국 칭화대의 MOOC 플랫폼인 XuetangX의 학습자 데이터를 제시하고 학습자의 수강취소 패턴을 분석 및 예측하는 것을 주제로 내건 바 있다 ("KDD Cup 2015", 2015). 이는 방대한 학습자 데이터를 이용하여 MOOC 서비스를 개선할 수 있다는 것을 암시한다.

하지만 우리나라에서는 아직까지 이러한 학습 데이터를 이용한 연구가 부족한 상태이며, K-MOOC 플랫폼에서 학습자들의 패턴을 분석하여 개선하려는 시도는 현재까지 없는 상태이다. 이에 플랫폼의 지속적이고 성공적인 운영을 위해선 학습자 로그 데이터를 기반으로 한 연구가 필수적이다.

1.2 기본 방향

교수자와 학습자가 면대면으로 의사소통하는 일반적인 오프라인 교육 환경에서 교수자는 학습자들과의 피드백을 주고받는 과정을 통해 학습자의 행동을 분석하고 적절한 교육 방식을 택하게 된다. 하지만, MOOC와 같은 온라인 환경에서는 비대면이라는 특성 때문에 이러한 상호작용이 불가능하고, 이에 교수자는 다른 방식으로 학습자의 상태를 모니터링하고 교수법을 개선해야 한다. 온라인 교육기관은 학습자에 관한 다양한 종류의 로그 데이터를 축적하게 되는데, 이를 이용하여 학습자를 분석하는 개념인 EDM이 한 가지 방안으로 대두되고 있다 (Romero & Ventura, 2007; Siemens, George & Ryan, 2012).

EDM(Educational Data Mining, 교육 데이터 마이닝)이란, 교육 환경에서 발생하는 대규모의 데이터를 탐색하고 이를 이용하여 수강생들과 그들의 환경을 분석하는 것을 일컫는다("Educational Data Mining", n.d.). EDM에서 연구자들은, 학습자들이 축적한 대규모의 데이터를 데이터마이닝, 머신러닝, 그리고 통계학 등의 기법을 이용하여 분석하고, 이를 통해 유의미한 결론을 도출한다.

EDM의 목표는 크게 다음과 같은 4가지로 정리할 수 있다 (Baker & Yacef, 2009).

A. 학습자의 차후 학습 행동 예측

지능형 교육 시스템(Intelligent tutoring system)의 학습자 모델링 기법을 이용하여, 수강생 개개인의 패턴을 분석하여 맞춤 교육을 제공할 수 있다.

B. 학습자 모형 발견 및 개선

EDM을 활용하여 1번에서 사용한 모델을 새로 발견 혹은 개선할 수 있다.

C. 교육 시스템의 효과에 대한 분석

교육 플랫폼이 어떻게 수강생에게 영향을 미치는지 EDM을 이용하여 확인할 수 있다.

D. 학습과 학습자에 대한 과학적 사실 분석 및 개선

학습자 모델과 플랫폼, 그리고 이론을 통합하여 학습 자체와 학습자에 대한 사실을 분석 및 개선할 수 있다.

또한 EDM은 아래와 같은 집단에 영향을 줄 수 있는데, 각각의 집단에서 EDM을 어떻게 적용할 수 있는지 알아보면 다음과 같다.

A. 학습자

EDM을 통해, 학습자는 자신의 패턴을 확인하고 학습자 모델을 바탕으로 제공되는 맞춤형 교육을 통해 보다 효율적으로 학습할 수 있다.

B. 교수자

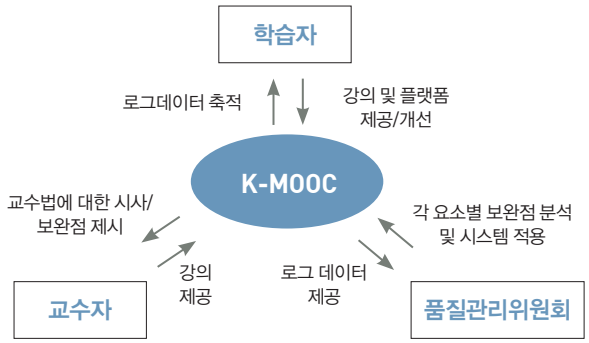
교수자는 자신의 교수법을 향상시키기 위해 EDM을 활용할 수 있다(Romero & Ventura, 2007). 예를 들어, 교수자들은 학습자의 데이터 변화를 이용하여 자신의 어조 및 속도를 파악하여 더 알맞게 개선할 수 있고, 학습자들의 전체적인 변화를 파악하여 커리큘럼 자체를 보완할 수도 있을 것이다(Bienkowski, Feng & Means, 2012).

C. 연구자 및 기관

연구자 및 기관은 EDM을 교육의 '효율성'에 집중한다(Romero & Ventura, 2010). EDM을 통해, 연구자는 수강생들이 보다 효율적으로 학습하기 위한 방법을 모색하고, 기관에서는 이렇게 해서 만들어진 자료들을 적재적소에 배치하는 것을 목표로 한다.

위와 같이 재래식 교육과 대비되는 비대면 교육의 단점을 보완한다는 EDM의 특성 덕분에(Ruipérez-

Valiente et al., 2017), MOOC 관련 연구에 EDM을 적용하는 사례가 늘어나고 있다. 또한, 빅데이터의 특성상 데이터를 보다 객관적이고 다양한 방면으로 분석하여 학습자의 패턴을 확인할 수 있기 때문에, 교육 데이터의 활용은 많은 실용적인 관찰을 돕는다(Guo et al., 2014). 실례로 MOOC에 데이터 마이닝을 적용하여 수강생의 수강취소 비율과 영상 내 재생 기록을 분석하는 연구가 있는 등(Kim et al., 2014), 최근 MOOC의 경험적 데이터를 이용한 연구가 많은 관심을 끌고 있다(Papamitsiou & Economides, 2014).



[그림 2] K-MOOC 내 세 요소의 상호작용

정리하자면, EDM을 MOOC 플랫폼에 적용한다면 다량의 데이터를 바탕으로 [그림 2]와 같이 각 이해관계자(학습자, 교수자, 품질관리 위원회)의 효율적인 의사결정을 지원할 수 있다.

1.3 연구의 목적과 범위

본 연구의 목적은 2018년도 한 해 분량의 비디오 이벤트 로그 데이터를 바탕으로 정량적, 실증적 분석을 통한 통찰을 제시하는 바에 있다.

이러한 통찰을 바탕으로, 온라인 강좌의 한계점을 극복하기 위한 제언 등 K-MOOC의 보다 효율적인 운영을 위한 추후 개선안을 제시할 수 있다(Ruipérez-Valiente et al., 2017). 이번 연구를 통해, K-MOOC의 데이터를 이용해 플랫폼의 학습자들을 보다 실증적으로 분석하고, 이를 바탕으로 좀 더 나은 의사결정을 도울 수 있는 것을 목표로 한다.

2 연구 방법

2.1 연구 도입

2018년 교육부가 발표한 한국형 온라인 공개강좌(K-MOOC) 운영계획에 따르면, i) 학습자, ii) 교수자, 그리고 iii) 품질관리위원회를 기준으로 플랫폼 개선 계획을 설정하였다. 본 연구를 통해서 학습자들의 K-MOOC 이용 현황 및 학습 패턴을 분석하고, 분석 결과를 활용하여 학습자, 교수자, 품질관리위원회에 각각 개선이 필요한 부분을 연구한 후, 플랫폼을 개선한다면 개선의 효과를 극대화할 수 있으리라 생각된다.

2.2 연구 절차

〈 표 1 〉 본 연구의 연구단계

	연구 내용	연구 방법
1단계	MOOC 데이터 분석 관련 선행연구 조사	문헌 분석
2단계	데이터 분석 파이프라인 구축	-
3단계	K-MOOC 데이터 전처리	양적연구
4단계	전처리된 데이터 분석	양적연구

(1) MOOC 데이터 분석 관련 선행연구 조사

K-MOOC 데이터를 활용한 분석연구에 앞서 MOOC 데이터를 활용한 선행 연구에 대한 조사를 진행했다. 선행 연구들이 해결하고자 했던 문제, 문제해결을 위해 사용한 방법, 그리고 얻은 결과와 한계를 집중적으로 살펴보았다. 이를 통해 본 연구를 통해 해결하고자 하는 문제를 구체화 시켰다.

(2) 데이터 분석 파이프라인 구축

주어진 K-MOOC 데이터를 살펴보고 원하는 분석을 위해 필요한 데이터 분석 파이프라인을 구축했다.

이를 위해 최신 데이터 분석 파이프라인들과 그 사례를 조사하고, 선행연구들이 채택했던 방법들도 살펴본 후 본 연구에 알맞은 과정으로 파이프라인을 구축하였다.

(3) K-MOOC 데이터 전처리

구축된 데이터 분석 파이프라인을 따라 필요한 데이터 전처리 과정을 거쳤다. 데이터 전처리 과정은 오류가 있는 데이터를 처리하는 데이터 클리닝 과정과 분석을 용이하게 하기 위해 데이터 형식을 변환하는 데이터 형식 변환 과정으로 구성된다.

(4) 전처리된 데이터 분석

전처리를 마친 데이터에 대해 다양한 분석을 진행하였다. 이 과정에서 시각화를 통해 다양한 시사점을 발견하였는데, 이러한 시각화 등의 기법을 활용하면 K-MOOC 서비스의 개선점을 확인할 수 있다(Zhang et al., 2016). 또한, 수강생들을 분류하여 분석하는 것은 수강 패턴을 이해하는 데에 상당한 도움이 되기 때문에(Kizilcec et al., 2013), 다양한 기준으로 수강생을 분류하여 분석하였다. 추후 본 분석에서는 강좌(course)를 구성하는 동영상 강의(lecture)에 한하여 '강의'라 일컫는다.

2.3 데이터 처리

(1) 데이터 분석 파이프 라인

[그림 3] 데이터 분석 파이프 라인

(a) 데이터 저장소

대용량의 데이터를 효과적으로 관리하기 위해, 본 연구에서는 Amazon 社의 웹 서비스인 AWS(Amazon Web Service) S3(Simple Storage Service)를 데이터 저장소로 사용하였다. AWS의 명령줄 인터페이스(CLI)를 활용하면 S3에서 local machine으로 또 그 반대 방향으로 데이터를 안전하고 쉽게 보낼 수 있기 때문에, 연구에서는 이를 적극 활용하여 분석을 진행하였다. 연구자의 분석 기기(local machine의 사

양이 좋지 않거나, 데이터 전처리의 시간을 줄이기 원하는 경우 local machine 대신 AWS의 EC2(Elastic Compute Cloud)를 사용할 수 있다.

2018년 한 해 동안에 수집된 이벤트 로그 데이터의 총 크기는 약 240GB 였다. 해당 데이터 내에는 잘못된 형식으로 저장된 데이터, 특정 속성 값이 들어있지 않은 데이터 등 오류가 있는 데이터도 포함되어 있었으며, 오류를 포함한 데이터는 데이터 전처리 과정을 통해 제외되고 분석에 사용될 데이터가 다시 데이터 저장소에 저장되었다(총 31,127,341개의 이벤트 로그).

(b) 데이터 전처리

S3에 저장된 데이터를 분석하는 작업에 앞서, 데이터 전처리(pre-processing) 과정이 필요하다. 데이터에 존재하는 문제점을 수정하고 이를 분석하는 데에 용이한 형식으로 변환하는 과정이 필요하기 때문이다. 본 연구에서는 AWS의 Athena를 이용해 빅데이터에 SQL(Structured Query Language) 쿼리와 Python을 활용하여 데이터 전처리를 진행하였다.

데이터 전처리 과정은 데이터 클리닝과 데이터 형식 변환으로 나누어 진행하였다. 먼저 데이터 클리닝이란, 수집된 데이터에 저장 형식과 관련된 오류가 있는 경우 오류를 수정하고, 수정할 수 없는 경우 데이터를 삭제하는 것을 의미한다. 많은 연구를 통해 다양한 전처리 과정을 거친 edX 등의 타 플랫폼과는 다르게, K-MOOC 관련 연구에서 데이터를 활용하여 진행한 것은 거의 전무하다. 이에 K-MOOC에서 전처리하는 예시 또한 없었으므로, 본 연구에서는 K-MOOC 학습자 로그 데이터를 직접 전처리하여 작업하는 과정을 추가했다. 전처리 과정에서 많은 데이터 오류를 발견했는데, 각 오류 유형 별로 일관된 형태로 데이터에 적절한 처리를 해주었다. 오류 데이터의 처리는 분석의 결과에 영향을 미칠 수 있기 때문에 대표적인 오류의 유형과 수정사항을 소개한다.

〈표 2〉 데이터 클리닝 예시

원본 데이터	오류	수정사항
{"id":"28d3c0181a9f40a9a8abf0a9b5b8ac8c", "code":"html5","currentTime"%}	이벤트 발생 시간이 저장되지 않음	데이터 삭제
{"context": {"course_id": "", "org_id": '', ..}	1. 강의 아이디, 기관 아이디가 저장되지 않음 2. 따옴표가 잘못 저장	데이터 삭제
{"id":"b9dca5fe292842ed9e753b0d07f61c9e", "code":"html5","currentTime":213.9997634r, "code":"html5","currentTime":213.9997634}	1. 같은 속성 값이 두번 들어가 있음 2. 현재 시간 값에 문자가 포함	문자가 포함되지 않은 현재 시간을 저장
{"id":"1d5739c580dd43679e6b4703cf30b3c6", "code":"html5","old_time":137.45cf30b3c6", "code":"html5","old_time":137.456,"new_time" :137.456,"type":"onCaptionSeek"}	1. 같은 속성 값이 두번 들어가 있음 2. seek 이전 위치 값에 문자가 포함	문자가 포함되지 않은 seek 이전 위치를 저장

위와 같이 데이터 클리닝 과정에서 다양한 문제점을 발견할 수 있었는데, 먼저 연구에 영향을 미칠 수 있는 중요한 데이터 오류들은 다음과 같다.

1. 이벤트 발생 시간이 저장되지 않거나 이벤트 발생 시간에 문자가 포함되는 오류(event_type이 play video, pause video 일 때)
2. seek 이전 혹은 seek 이후의 위치에 문자가 포함되는 오류(event_type이 seek_video 일 때)
3. 중요한 속성 값들(course_id, lecture_id, org_id, user_id) 중에 특정 값이 저장되지 않는 오류
4. 문자열 안에서 따옴표가 잘못 들어가 있는 상태로 저장된 오류

위 항목 중, 3, 4 번과 같이 데이터의 정보를 훼손시키지 않고 수정이 가능한 경우는 데이터 형식을 수정하였고, 1, 2 번과 같이 사용할 수 없는 데이터는 제거하여 진행하였다.

(C) 데이터 형식 변환

데이터 클리닝 이후, 데이터 형식 변환 과정을 진행하였다. 학습자의 이용 목적에 맞게 데이터를 수정하는 것을 데이터 형식 변환이라고 하는데, 데이터 형식 변환은 다음과 같은 이유로 필수적이다.

1. 보다 분석할 항목에 적합하게 형식을 설정할 수 있다
2. 분석에 사용되는 CPU, 메모리 등 리소스를 절약할 수 있기 때문에, 비용 측면에서 용이하다

이에 본 연구에서는 필요한 항목을 확인한 후 해당 항목에 맞게 로그를 추출하였다. 추출된 항목은 다음과 같다.

〈표 3〉 추출되어 저장되는 데이터 항목

항목	설명	항목	설명
course_id	강좌 ID	event_type	이벤트 종류
lecture_id	강의 ID	current_time	이벤트가 발생한 실제 시각
user_id	학습자를 식별할 수 있는 암호화된 ID	video_time	이벤트가 발생한 비디오내 시각
org_id	강의를 제공하는 기관 ID	new	event_type에 따른 세부사항
session	학습자의 session 코드	old	event_type에 따른 세부사항
language	학습자가 설정한 언어		

〈표 4〉 추출된 항목들을 가지는 데이터 예시

course_id	lecture_id	user_id	org_id	session
course-v1:UOSk+ACE_UOS03+2018_1	fa94e200d91b407581a858d6a29976e2/4d12039dd70f4be5b3e7ed4c05b3a067	e81088a3820f6a2f99899842d26ddcb0e7bd91a98280beb5c364334795027882	UOSk	2993ac58578aec413016d8708aafca69

current_time	video_time	event_type	new	old	language
2018-04-15T04:17:02.218056+00:00	424.5926002	seek_video	487	424.5926002	ko-KR

원본 데이터에 강의 아이디(lecture_id)가 명시적으로 저장되어 있지 않은 경우가 있었는데, 이 경우 데이터 원본에 있는 referer 항목에서 강의 아이디에 해당하는 부분을 추출해 저장하였다.

한편, 본 연구에서 활용하고자 하지만 원본 로그 데이터에서 바로 얻을 수 없는 항목들이 있었다. 해당 항목은 다음과 같다.

1. 학습자의 비디오 재생 속도
2. 캡션 및 스크립트가 학습자에게 보여지는 여부

K-MOOC 시스템에서는 강의를 듣고 있는 학습자의 로그 데이터를 저장할 때 위와 같이 현재 비디오 재생 속도, 캡션과 스크립트 사용 여부에 대해 저장하지 않기 때문에, 이는 기존 데이터를 정렬하여 확인해야 한다. 따라서, 해당 항목을 추출하기 위해 데이터를 각 학습자별로 그룹화한 다음, 그룹화된 데이터를 이벤트가 발생한 실제 시각(current_time)기준으로 정렬하였다. 또한, 비디오 재생 속도와 캡션과 스크립트 사용여부에 대한 정보를 남기는 event_type (Speed_change_video, Video_show_cc_menu, Video_hide_cc_menu, Show_transcript, Hide_transcript)이 발생한 시점을 나열하였다. 각 시점을 기준으로 이전과 이후의 비디오 재생 속도와 캡션과 스크립트 사용여부에 대한 정보를 알아낼 수 있었다.

〈표 5〉 구간 형식으로 변환된 데이터 스키마

항목	설명	항목	설명
course_id	강좌 ID	event_type_from	interval 시작시의 event_type
lecture_id	강의 ID	event_typt_to	interval 종료시의 event_type
user_id	학습자를 식별할 수 있는 암호화된 ID	current_time	이벤트가 발생한 실제 시각
session	학습자의 session 코드	speed	interval 동안 비디오 재생 속도
video_time_from	interval 시작시의 비디오 시간	cc	cc의 show/hide 여부
video_time_to	interval 종료시의 비디오 시간	transcript	transcript의 show/hide 여부

또한, 해당 연구에서는 강사의 말하기 속도 변화와 학습자의 비디오 재생 속도를 비교하고자 하였다. 이와 같은 분석을 진행하고자하여 이벤트 발생 시에 저장된 데이터들을 구간(이벤트 발생 시점부터 다음 이벤트 발생시점까지) 하나의 구간으로 나타내고, 구간을 하나의 데이터로 변환하는 작업을 진행했다. 위와 같은 과정을 통해 최종적으로 얻어진 데이터 형식은 [표5]와 같다.

(a) 데이터 분석

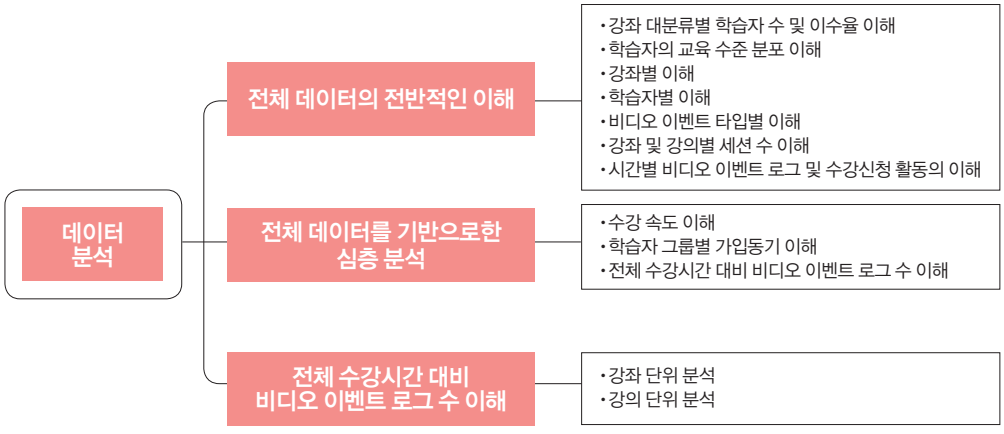
위와 같이 데이터 전처리 과정을 거친 후, 정제된 데이터를 활용하여 크게는 3가지 항목으로 분류되는 여러 질문 및 가설들을 확인해보는 작업을 진행하였다. 데이터 분석에는 Python, R등을 활용하였다.

(b) 데이터 시각화

수강자들이 분석을 통해 얻은 결과로부터 통찰을 얻는 것을 돕기 위해, 수강자들에게 제공할 수 있는 간단한 시각화를 진행하였다. 데이터 시각화의 결과는 데이터 분석 항목에 있는 그림을 통해 확인할 수 있다.

2.4 주요 분석 활동

본 연구의 제 4단계, 즉 전처리된 데이터의 분석은 다음의 세 가지 측면으로 구성된다. ([그림 4] 참조): (1) 전체 데이터의 전반적인 이해, (2) 전체 데이터를 기반으로 한 심층 분석, (3) 단일 강좌에 대한 사례 분석. 각각의 측면에서 검증하고자한 바와 그 중요성을 아래에 제시한다.



[그림 4] 분석 측면별 상세 분석 내용

(1) 전체 데이터의 전반적인 이해

(a) 강좌 대분류별 학습자 수 및 이수율 이해

강좌 대분류별 학습자 수 및 이수율의 차이를 이해함으로써 학습자층의 관심 분야를 알고 적은 이수율을 갖는 분야의 강좌의 문제점 파악으로 이어질 수 있다.

(b) 학습자의 교육 수준 분포 이해

학습자의 교육 수준 분포를 이해함으로써 추후 K-MOOC 제공 강좌의 난이도 및 타겟층을 설정하는데 도움이 될 수 있다.

(c) 강좌별 이해

각 강좌별 비디오 이벤트 로그 분포 분석을 통해 학습자의 관심의 분포를 이해할 수 있다.

(d) 학습자별 이해

각 학습자별 비디오 이벤트 로그 분포 분석을 통해 학습자의 학습 정도 분포를 알 수 있다.

(e) 비디오 이벤트 타입별 이해

각 비디오 이벤트 타입별 비디오 이벤트 로그 분포 분석을 통해 학습자가 학습을 진행하면서 강의와의 인터랙션 패턴을 이해할 수 있다.

(f) 강좌 및 강의별 세션 수 이해

각 강좌 및 강의별 학습자 개개인의 세션 수 분포의 이해를 통해 학습자의 강좌 및 강의 학습 패턴을 이해할 수 있다.

(g) 시간별 비디오 이벤트 로그 및 수강신청 활동의 이해

비디오 이벤트 로그가 발생한 시간 및 수강신청 활동이 발생한 시간의 월별, 요일별, 시간대별 분석을 통해 학습자의 실제 강의 학습 패턴 및 학습 준비 패턴을 이해할 수 있다. 이러한 분석을 바탕으로 추후에 학습자의 동기 유발 및 관심 지속을 위한 적절한 조치를 취할 수 있다.

(2) 전체 데이터를 기반으로 한 심층 분석

(a) 수강 속도 이해

K-MOOC 플랫폼은 수강자가 처음 강의를 수강 할 때 기본적으로 강의를 1.0배속으로 제공한다. 하지만

학습자가 수강 속도를 맞춤식으로 변경할 수 있기 때문에, 이러한 수강 속도 로그는 학습자들의 학습 패턴을 분석하는 데에 유용한 정보가 될 수 있다. 또한 이러한 수강 속도의 분포를 강좌 종류별의 다각적인 면에서 분석하여 이를 바탕으로 교수자에게 피드백이 제공하거나 추후 비디오 인터페이스를 개선시킬 수 있다.

(b) 학습자 그룹별 가입 동기 이해

학습자는 K-MOOC 홈페이지에 회원가입을 할 때, 선택적으로 가입 동기를 서술형으로 적을 수 있다. 이러한 가입 동기는 학습 동기를 내포하므로 매우 중요한 지표가 된다. 이러한 학습 동기를 연령별, 학력별 학습자 그룹별로 이해할 경우, 추후 K-MOOC 소개 시 학습자 그룹별 맞춤식 홍보를 할 수 있다. 또한, 이러한 학습 동기가 추후 학습자의 수강패턴에 미치는 영향을 분석하여 학습 수강 성과 차이의 원인을 유추해볼 수 있다.

(c) 전체 수강 시간 대비 비디오 이벤트 로그 개수 이해

학습자가 강의를 수강하는 동안 발생한 강의와의 인터랙션 정도는 학습자의 학습 정도와 관련된 중요한 지표가 될 수 있다. 예컨대, 단시간 내에 많은 비디오 이벤트 로그가 발생하는 강좌의 경우, 강좌의 난이도, 강좌 내 강사의 말 속도 등 다양한 원인과 이을 수 있다.

(3) 단일 강좌에 대한 사례 분석

(a) 분석 동기

단일 강좌에 대한 사례 분석을 하게 된 이유를 언급하였다.

(b) 강좌 선택

하나의 분석 사례로 선택한 강좌의 정보와 선택한 이유 및 타당성에 대해 언급하였다. 이를 통해 위 사례 분석이 특수한 한 사례를 분석한 것이 아님을 보인다.

(c) 강좌 단위 분석

선택된 강좌(Course)에 포함된 강의들을 커리큘럼 순으로 배치하고 학습자들이 각 강의에서 어떤 이벤트 로그를 남겼는지, 어떤 경향성이 있는지를 분석한다. 이를 통해 학습자들이 강좌 전체에 참여하고 있는지, 강좌 내에서 학습자들이 어려움을 겪는 부분은 어디인지와 같은 학습자의 강좌를 수강하는 패턴을 분석할 수 있다.

(d) 강의 단위 분석

보다 자세히 들어가서 한 강의(lecture)안에서 학습자들은 어떤 이벤트 로그를 남겼는지 분석한다. 이를 통해 '강의 중간에 어떤 부분에서 학습자들이 어려움을 겪는가'와 같은 질문에 대한 답을 찾을 수 있다.

3 연구 결과

3.1 전체 데이터의 전반적인 이해

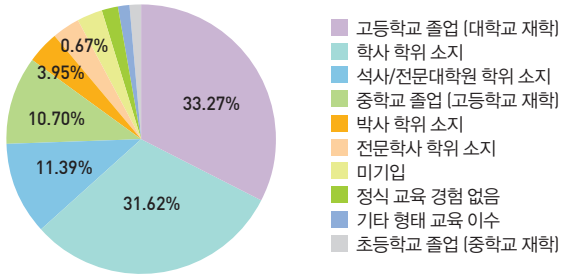
(1) 강좌의 대분류별 학습자 수 및 이수율 이해

각 강좌는 강좌 주제에 따라 7가지 대분류중 하나에 포함된다. 대분류별로 강좌 개수와 학습자의 수 및 강좌의 50% 이상을 수강한 학습자의 비율을 구하였다. [표6]을 통해 확인할 수 있듯이, 공학과 자연 분류의 경우 다른 분류들보다 강좌의 개수 대비 학습자 수가 많다. 이를 통해 공학과 자연의 경우 많은 학생들이 관심을 보이고 있음을 알 수 있다. 하지만 그에 비해 강좌를 끝까지 수강한 학습자의 비율이 낮다는 것 또한 알 수 있다. 따라서 보다 많은 공학, 자연 강의를 개설하고, 학습자가 강좌를 끝까지 수강하지 않는 원인을 찾아 강의를 개선한다면 많은 학습자들에게 도움이 될 수 있으리라 예상된다.

〈표 6〉 강좌 대분류별 정보

대분류	강좌 개수	학습자 수	강좌를 50% 이상 수강한 학습자의 비율 (%)	강좌를 이수한 학습자의 비율 (%)
사회	123	75,395	16.8	7.7
인문	119	44,872	19.9	8.3
공학	77	54,518	11.5	6.2
자연	56	42,501	15.9	7.4
예체능	36	13,835	24.9	9.6
의약	29	15,051	17.4	11.6
교육	22	13,187	21.5	8.8

(2) 학습자의 교육수준 분포 이해



[그림 5] 학습자의 교육수준 분석

본 연구에서는, 학습자들을 교육 수준을 기준으로 분류하고 분석해 보았다. [그림 5]에서 볼 수 있듯이, 학부생과 학부 졸업생에 해당하는 비율이 매우 큰 것을 알 수 있는데, 이는 많은 수의 강좌들이 학부/학부 졸업생들을 대상으로 한 콘텐츠를 제공하기 때문이라고 추측할 수 있다. 또한, 대학교가 K-MOOC 강좌를 연계하여 수업을 진행하는 경우도 많기 때문에, 학부생의 비율이 크게 나타났을 것이라 예측된다. 이는 학사 학위 소지자가 73%를 차지하였던 HarvardX 및 MITx에서의 4년치 학습자 데이터 분석과는 차이를 보인다 (Chuang et al., 2016). 이러한 차이는 학습자들의 강좌 활용 동기 등에서 기인할 수 있다.

(3) 강좌별 이해

학습자의 전체 비디오 이벤트 로그 데이터를 분석한 결과, 비디오 이벤트 로그가 존재하는 기관의 수는 92개였다. 비디오 이벤트 로그가 존재하는 강좌는 1161개가 있었으며, 각 강좌마다 남겨진 비디오 이벤트 로그 개수, 비디오 이벤트가 존재하는 강의의 개수, 비디오 이벤트를 남긴 학습자의 수는 [표 7]과 같다. [표 7]에서 확인할 수 있듯이, 강좌별 이벤트 로그의 개수 및 학습자의 수의 표준편차가 매우 컸다(이벤트 로그 개수: 54013.3개, 학습자 수: 210.3명). 이는 강좌별 강의 개수 및 총 강의 시간의 편차에서 기인될 수 있겠지만, 한편으로는 인기 강좌와 비인기 강좌사이에 큰 편차가 존재하는 것을 의미한다.

〈표 7〉 전체 비디오 로그 이벤트 분석을 통한 강좌별 이벤트 로그, 강의, 학습자 개수

	이벤트 로그 개수	강의 개수	학습자 수
평균	29,610	34	158
중앙값	9,606	33	81
최대값	462,128	120	2,540
최소값	1	1	1
표준편차	54,013.3	19.5	210.3

(4) 학습자별 이해

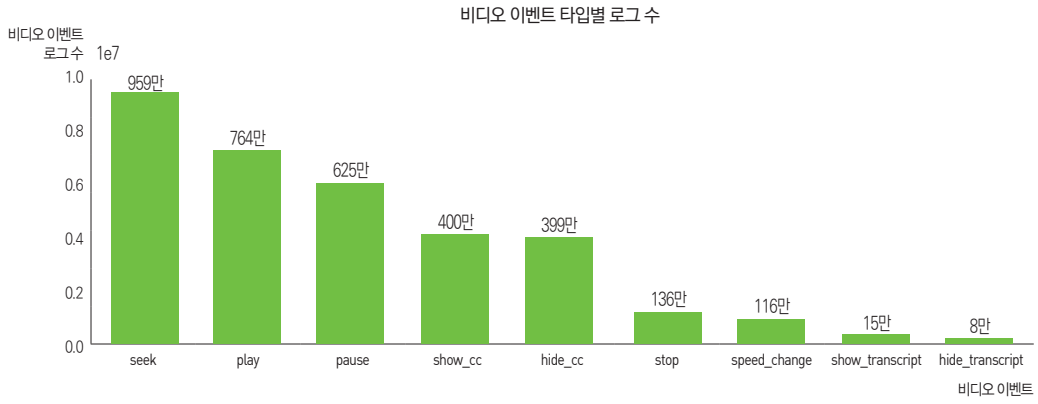
학습자의 로그 데이터를 분석한 결과 학습자의 수는 총 9445명 이었으며, 각 학습자마다 남긴 비디오 이벤트 로그 개수, 비디오 이벤트 로그가 존재하는 강좌 수, 비디오 이벤트 로그가 존재하는 강의 개수는 [표 8]에서 나타내었다. [표 8]과 같이, 학습자별 비디오 이벤트 로그의 개수 및 강의 수의 표준편차가 매우 컸다(이벤트 로그 개수: 1160.9개, 강의 수: 47.2개). 이를 통해 학습자의 학습정도의 편차가 큼을 알 수 있다.

〈표 8〉 전체 비디오 로그 이벤트 분석을 통한 학습자별 이벤트 로그, 강좌, 강의 개수

	이벤트 로그 개수	강의 개수	학습자 수
평균	363	1.93	18.6
중앙값	93	1	7
최대값	107480	165	3533
최소값	1	1	1
표준편차	1160.9	3.14	47.2

(5) 비디오 이벤트 타입별 이해

각 비디오 이벤트 타입별 학습자가 남기는 비디오 이벤트 로그 분포는 [그림 6]과 같다. [그림 6]과 같이 학습자 로그 중 28%는 비디오 내 탐색(seek), 22%는 재생(play), 18%는 일시 정지로, 세 가지의 인터랙션을 주로 활용하였다. 속도 변화에 대한 로그는 3%에 불과하였는데, 이를 통해 주로 강의 속도는 설정 후 일정하게 유지하면서 강의를 듣는 것을 알 수 있다.



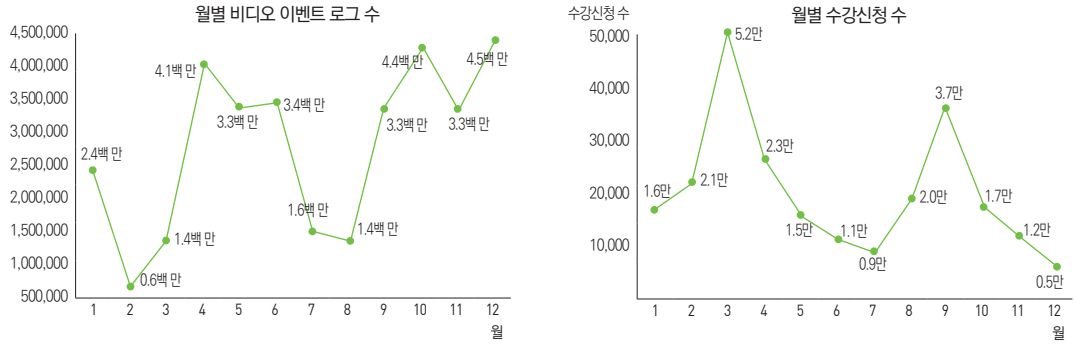
[그림 6] 비디오 이벤트 타입별 로그 수

(6) 강좌 및 강의별 세션 수

강좌 또는 강의 별 비디오 이벤트 로그의 서로 다른 세션 수를 보면, 학습자가 한 강좌 또는 강의를 몇 번이나 수강 및 참고하는지를 알 수 있다. 분석 결과, 총 10,415개의 비디오 이벤트가 존재하는 강의에 대해서 강의별 평균 학습자별 세션 수는 평균 1.22회(표준편차: 0.28, 최소: 1.0회, 최대: 8.75회)로, 대부분의 학습자는 한 강의를 한 번의 세션으로만 접속하는 것을 알 수 있었다. 이에 반해, 총 1,163개의 강좌에 대해서 강좌별 평균 학습자별 세션 수는 평균 3.12회(표준편차: 1.85, 최소: 1.0회, 최대: 35.17회)였다. 이는 강좌별 강의 수가 평균 64.55개(표준편차: 25.25, 최소: 10, 최대: 165, 강좌 수는 퀴즈, 기말평가 등을 포함. 강좌 내 카테고리의 개수는 평균 14.33개, 표준편차: 4.13, 최소: 3, 최대: 44)인 것에 비해 현격히 적으므로, 주로 학습자는 1회의 세션에서 여러 개의 강의를 한 번에 듣는 것을 알 수 있다. 이때 강의별 평균 학습자별 세션 수 및 강좌별 평균 학습자별 세션 수의 표준편차가 적은 것을 통해 이는 다양한 학습자를 불문하고 나타나는 패턴임을 알 수 있다.

(7) 시간별 비디오 이벤트 로그 및 수강신청 활동의 이해

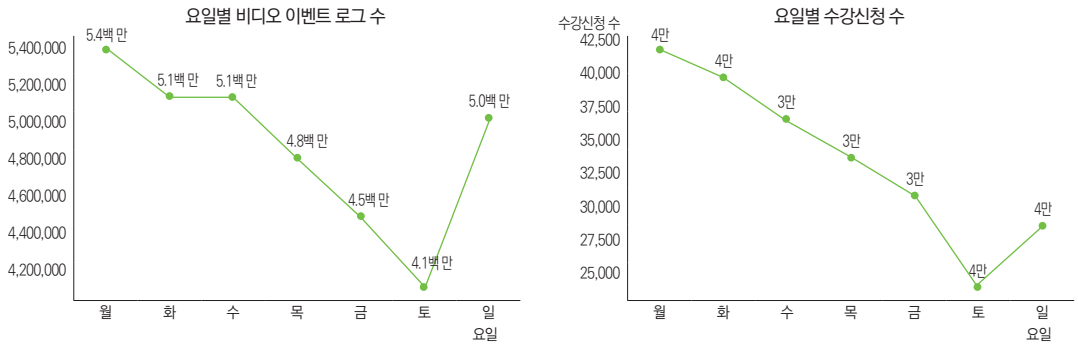
학습자의 비디오 이벤트 로그는 수강 신청을 하는 것과는 달리 실제 수업 수강 중에만 나타나는 지표이므로, 이러한 비디오 이벤트 로그의 수가 시간대별로 어떻게 다른지를 아는 것은 중요하다. 이를 위하여, 월별, 요일별, 시간대별로 비디오 이벤트 로그의 수가 어떻게 변동하는지를 살펴보았다. 이때, 실제 로그는 시간대가 +00시를 기준으로 기록되어 있으므로 수강생이 모두 한국에 있다는 가정하에, 시간대를 +09시를 기준으로 변환하여 계산하였다.



[그림 7] (좌) 월별 비디오 이벤트 로그 수, (우) 월별 수강신청 수

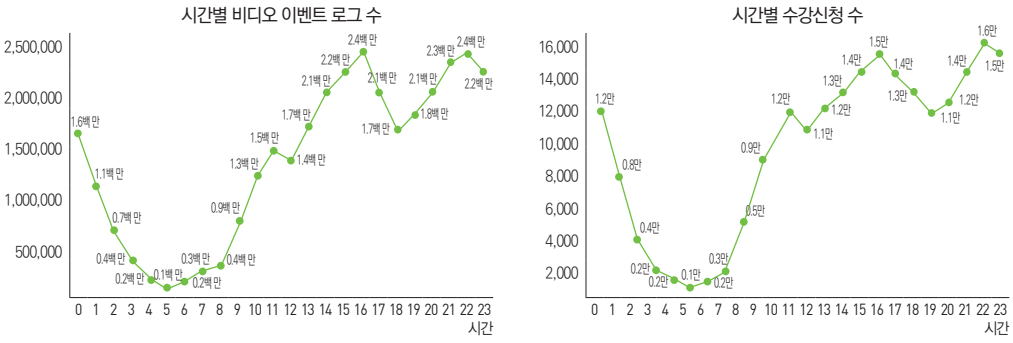
먼저, 월별 비디오 이벤트 로그 수의 변화는 [그림 7]에 나타나 있다. [그림 7]에서 볼 수 있듯이, 한국을 기준으로 하였을 때, 방학 기간인 1, 2, 7, 8월에는 이벤트 로그 수가 적음을 관찰할 수 있었다. 또한, 개강

후 바로 다음 달인 4월, 10월에는 이벤트 로그 수가 많음을 관찰할 수 있었다. 이는 3월 및 9월에 수강신청 수가 많은 것과 대조적이다.



[그림 8] (좌) 요일별 비디오 이벤트 로그 수, (우) 요일별 수강신청 수

다음으로, 요일별 비디오 이벤트 로그 수의 변화는 [그림 8]에 나타나 있다. [그림 8]에서 볼 수 있듯이, 한 주의 시작인 월요일에 가장 많은 비디오 이벤트가 발생하고, 토요일까지 그 수는 계속하여 감소하게 되며, 일요일에 다시 증가함을 알 수 있다. 이는 수강신청 수의 변동에서도 동일하게 관찰된다. 따라서, 주중 후반으로 갈수록 낮아지는 학습자의 관심을 고무시킬 적절한 조치가 필요하다.



[그림 9] (좌) 시간별 비디오 이벤트 로그 수, (우) 시간별 수강신청 수

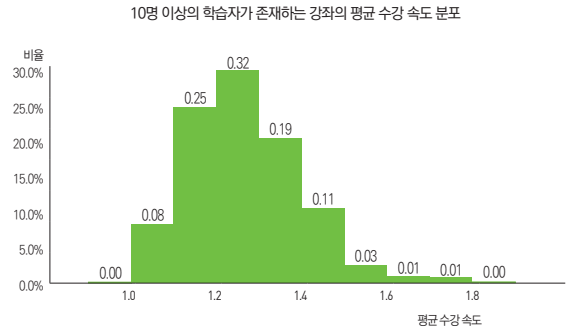
마지막으로, 시간별 비디오 이벤트 로그 수의 변화는 [그림 9]에 나타나 있다. [그림 9]에서 볼 수 있듯이 주로 늦은 오후 시간과 저녁 시간 후의 밤에 많은 수강 이벤트가 발생했음을 알 수 있다. 이 또한 수강신청 수의 변동에서도 동일하게 관찰된다.

3.2 전체 데이터의 전반적인 이해

(1) 수강 속도

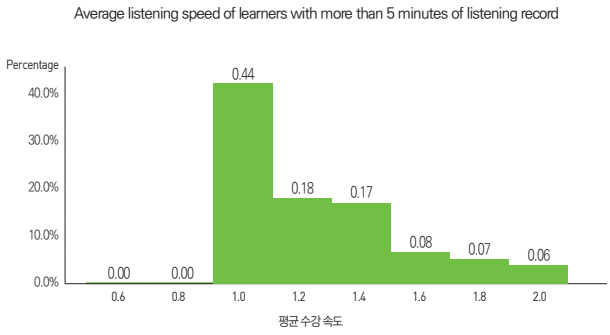
수강자는 한 강의 내에서도 구간별로 배속을 다르게 설정할 수 있고, 이러한 구간은 그 길이가 다를 수 있기 때문에 실제 수강 속도는 다음과 같이 정의하여 계산하였다.

본 연구에서는 적은 수의 학습자가 수강하는 경우 평균 속도가 유의미하지 않을 것이라고 가정하였기 때문에, 비디오 이벤트 로그 분석은 비디오 이벤트 로그가 존재하는 학습자가 최소 10명인 총 945개의 강좌에 대해서 진행하였다. 분석 결과, 강좌별 강의 수강 평균 속도의 분포는 [그림 10]과 같이 강좌별 수강 평균 속도는 1.26배속을 평균으로 하여 종 모양(bell-shape)을 형성하고 있다.



[그림 10] 강좌별 평균 수강 속도 분포

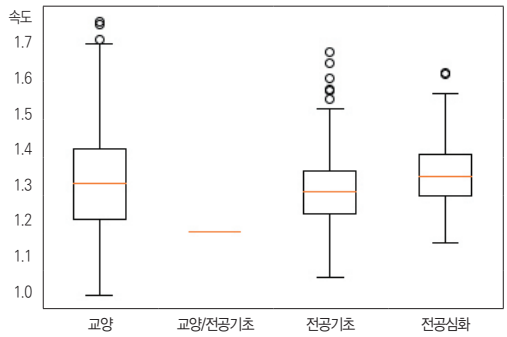
또한, 본 연구에서는 5분 미만을 시청하는 대부분 학습자의 경우 평균 속도가 유의미하지 않을 것이라고 가정하였다. 비디오 이벤트 로그 분석을 통해 총 수강 기록이 5분 이상이 되는 70,216명의 학습자에 한해서 나타내었는데, 해당 학습자별 강의 수강 평균 속도의 분포도는 [그림 11]과 같다.



[그림 11] 학습자별 강의를 듣는 평균 속도의 분포도

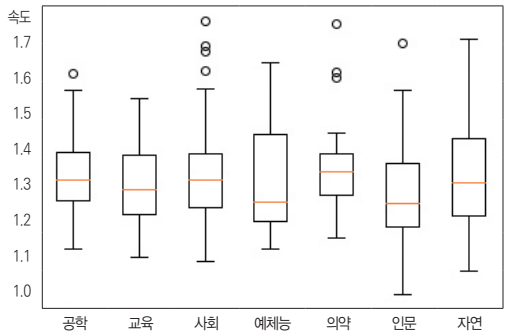
[그림 11]에서 알 수 있듯이, 1배속보다 느리게 듣는 학습자는 0.00%에 가까웠으며, 0.9~1.1배속으로 듣는 학습자가 44% 정도에 가까웠다. 그래프를 통해 학습자별 평균 속도의 분포는 종 모양의 분포를 하는 강좌별 평균 속도의 분포도와는 다른 분포를 함을 알 수 있다.

각 강좌는 교양, 교양/전공기초, 전공기초, 그리고 전공심화와 같이 총 4개의 강좌 난이도 중 하나로 분류된다. 따라서, 본 연구에서는 강좌 난이도별 수강 속도의 차이 유무를 살피고자 하였다. 각 강좌 난이도별 수강 속도 분포는 [그림 12]와 같은데, 정확한 검증을 위해 그룹 간 분산분석 (between group ANOVA)을 진행하였다. 분석 결과, 그룹 간 수강 속도는 유의미한 차이가 발견되지 않았다 ($F=1.5$, $p=0.23$, $\alpha=0.05$).



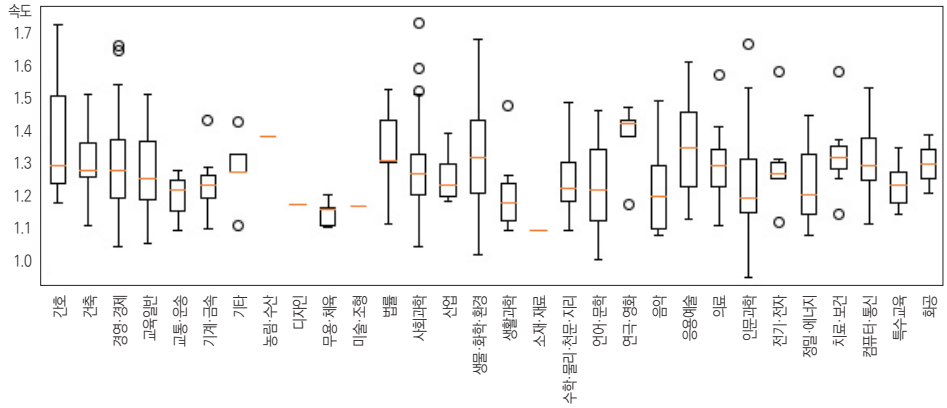
[그림 12] 강좌 난이도별 수강 속도 분포

또한 강좌는 공학, 교육, 사회, 예체능, 의약, 인문, 자연과 같이 총 7개의 분야로 대분류와 간호, 건축, 산업, 교육일반, 생활과학 등 총 30개의 상세 분야로 분류된다. 따라서, 강좌 난이도 이외에 강좌 분야별 수강 속도의 차이 유무 또한 살펴보았다. 각 강좌 대분야별 수강 속도 분포는 [그림 13]과 같이 도출되었는데, 정확한 검증을 위해 그룹 간 분산분석을 진행하였다. 그 결과, 유의미한 차이가 발견되었다($F=2.8$, $p=0.01$, $\alpha=0.05$). 따라서, Tukey HSD(FWER=0.05)로 사후검증(post-hoc test)을 진행해본 결과, '사회'와 '인문' 분야에서 유의미한 결과가 나왔다('사회' 강좌 평균: 1.32배속, '인문' 강좌 평균: 1.27배속). 이러한 유의미한 차이가 나타난 이유에 대해 추가적인 양적 및 질적 탐색을 통한 확인이 필요하다.



[그림 13] 강좌 대분야별 수강 속도 분포

이후, 강좌 중분야별 차이 또한 존재하는지를 살펴보았다. 각 강좌 중분야별 수강 속도 분포는 [그림 14]와 같다. 정확한 검증을 위해 그룹 간 분산분석을 진행하였고, 그 결과 유의미한 차이가 발견할 수 있었다 ($F=1.7$, $p=0.01$, $\alpha=0.05$). 이에 Tukey HSD(FWER=0.05)로 사후검증(post-hoc test)를 진행해본 결과, 총 435개의 중분류 쌍 간에 유의미한 차이가 발견되지 않았다.



[그림 14] 강좌 중분야별 수강 속도 분포

수강 속도와 관련하여 다양한 분석을 하였지만, 수강 속도만으로는 강좌 분석에 한계를 가진다. 이는 강좌마다, 강의마다, 그리고 강의 내에서도 구간마다 발화 속도가 다르기 때문에 똑같은 강좌 재생 속도더라도 그 의미가 다를 수 있기 때문이다. 따라서, 추후에 크롤링한 자막 데이터와 연결지어 시간당 글자 개수 지표, 시간당 단어 개수 지표 등을 이용하여 더 심화된 분석을 수행할 예정이다.

(2) 가입 동기

학습자가 K-MOOC 홈페이지에 회원가입을 할 때, 선택적으로 가입 동기를 서술형으로 적을 수 있다. 선택적 작성이기에 모든 학습자의 가입 동기 정보가 존재하지는 않지만, 총 361,195명의 학습자 중 215,813명이 가입 동기를 적어 분석의 의미가 있었다. 전체 학습자의 가입 동기에 대해 Python의 konlpy[1] 라이브러리를 적용하여 명사 품사만을 추출한 뒤, 해당 명사의 빈도수를 계산하여 [그림 15]와 같이 워드클라우드를 출력하였다. 워드클라우드를 통해 다양한 가입 동기 관련 단어를 살필 수 있었는데, 예를 들어 '관심'이 있어 가입을 하게 된 학습자가 많았다는 것을 나타내고 있고, '평생'교육을 위해 가입한 학습자, 학점 '이수'를 위해 가입한 학습자, '권유'를 받아 가입한 학습자 등 학습자의 학습 동기 또한 다양하다는 것을 알 수 있다.

1) http://konlpy.org

워드클라우드를 통해 다양한 가입 동기 관련 단어를 살필 수 있었는데, 예를 들어 '관심'이 있어 가입을 하게 된 학습자가 많았다는 것을 나타내고 있고, '평생'교육을 위해 가입한 학습자, 학점 '이수'를 위해 가입한 학습자, '권유'를 받아 가입한 학습자 등 학습자의 학습 동기 또한 다양하다는 것을 알 수 있다.

[그림 15] 전체 학습자의 가입 동기 워드클라우드

(상위 좌측부터 0대, 10대, 20대, 30대, 40대, 50대, 60대, 70대, 80대의 가입 동기 워드클라우드, 80대의 가입 동기의 경우 학습자의 수가 적으므로, 키워드가 많이 검출되지 않았다)

[그림 16] 각 연령대별 가입 동기 워드클라우드

그 다음, 가입 동기를 구체적으로 분석하기 위해 다양한 학습자 그룹별로 가입 동기에 차이가 있는지 살펴보았다. 먼저 연령대별 가입 동기 워드클라우드를 살펴보았는데, 연령대 또한 회원가입 시 선택적으로 목록에서 자율적으로 선택할 수 있는 형태이기 때문에 일부의 학습자만 선택을 하였다. 특히, 본인인증을 하는 부분이 없기에 0대(1세~9세)로 표기하는 등 데이터에 문제가 있기도 하였다. 각 연령대별 가입 동기 워드클라우드는 [그림 16]와 같다.

[그림 16]에서 볼 수 있듯이, 세대를 불문하고 '관심', '지식', '강의'와 같이 키워드는 모두 관찰되었다. 하지만 연령대별로 가입 동기에서도 차이가 나타났는데, 특히 10대의 경우 '진로'라는 키워드가 많이 나타났다. 이는 다수의 10대 학습자가 진로를 정할 때 K-MOOC을 활용하고자 하였음을 알 수 있다. 또한 '미리'라는 키워드를 통해서도 10대의 학습자가 진학 후 배울 공부를 미리 배우고 체험할 때 쓴다는 것을 알 수 있다. 20대의 학습자부터는 '관심'이라는 키워드가 대두하는데, 이는 관심 있는 분야에 대해 더 학습하기 위해 K-MOOC을 활용한다는 것을 나타낸다. 30대 학습자부터는 '평생'이라는 키워드가 등장하는데, 해당 키워드는 다음 연령 그룹으로 갈수록 더 큰 비율을 차지하는 추세를 보였다. 이는 연령대가 있는 학습자일수록 가입 동기가 '평생교육'을 위함인 것을 알 수 있다. 또한, '학교', '대학'의 키워드가 70대에 다시 나타나는 것 또한 주목할 만 하다. 이러한 연령대별 K-MOOC 활용 목적이 다름을 이용하여 연령대별 전략적 맞춤형 광고에 활용할 수 있다는 시사점을 제시해주고 있다.

이러한 학습자 그룹별 가입 동기의 차이가 학력별로 학습자를 구분 지을 때도 나타나는지 살펴보았다. 각 학력별 가입 동기 워드클라우드는 [그림 17]와 같다.

(상위 좌측부터 박사 학위 소지자, 석사/전문대학원 학위 소지자, 학사 학위 소지자, 전문학사 학위 소지자, 고등학교 졸업자(대학교 재학자), 중학교 졸업자(고등학교 재학자), 초등학교 졸업자(중학교 재학자), 정식 교육 경험 없는 학습자, 기타 형태의 교육을 이수한 학습자의 가입 동기 워드클라우드)

[그림 17] 각 학력별 가입 동기 워드클라우드

[그림 17]에서 볼 수 있듯이 학력의 차이는 가입 동기 워드클라우드는 단어의 선택에서도 나타난다. 예를 들어, 같은 의미를 내포하더라도 강의와 수업은 확연한 다른 어감을 나타낸다. 비슷한 예로는 대학과 학교가 있을 수 있다. 가입 동기 작성란이 자율형식이었기에 이러한 단어선택의 차이로 이어짐을 확인할 수 있다.

이외에도 이수한 강좌의 개수가 많고 적음이 가입 동기의 차이로 인해 발생한 것인지를 검증하고자 하였다. 따라서, 본 연구에서는 학습자별 이수한 강좌의 개수를 계산하여 이를 바탕으로 그룹 내 편차는 최소화하고, 그룹 간 편차는 최대화하는 알고리즘인 Jenks natural breaks classification 알고리즘을 이용하여 학습자를 3개의 그룹으로 나누었다. 해당 알고리즘을 통해 0개의 강좌를 이수한 학습자들(총 58,589명), 1~12개의 강좌를 이수한 학습자들(총 22,098명), 13개~74개의 강좌를 이수한 학습자들(총 37명)로 나눌 수 있었는데, 이렇게 생성된 3개의 그룹의 가입 동기 워드클라우드는 [그림 18]과 같다.

(좌) 이수 강좌 수가 0개인 학습자들의 가입 동기 워드클라우드 (우) 이수 강좌 수가 1~12개인 학습자들의 가입 동기 워드클라우드 (이수 강좌 수가 13~74개인 학습자들의 가입 동기에서는 키워드가 검출되지 않았다.)

[그림 18] 각 이수 강좌 수 클래스별 가입 동기 워드클라우드

[그림 18]에서 볼 수 있듯이, 이수 강좌 수별 가입 동기는 확연한 차이가 나타나지 않았다. 이를 통해, 이수 강의의 수의 차이는 단순히 가입 동기의 차이로부터 나타난 것이 아니라 추후의 외적 요인으로부터 발생한 것임을 알 수 있다.

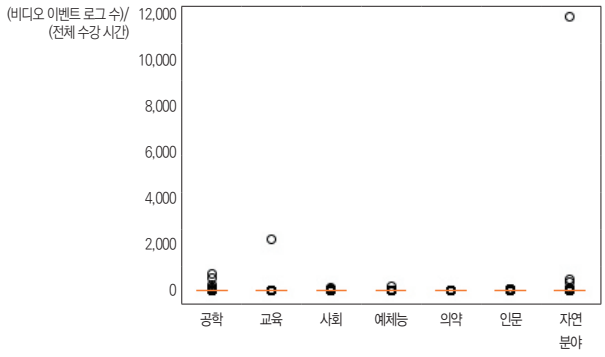
(3) 전체 수강 시간 대비 비디오 이벤트 로그 수

강의 수강 시 강의와의 상호작용 정도를 나타내는 지표는 다음과 같이 계산하였다. 단순히 비디오 이벤트 로그 수를 분석하는 것이 아닌 전체 수강 시간에 대비하여 정의하여 더 의미 있는 지표가 나타나도록 하였다. 위의 지표를 바탕으로 총 7개의 대분류 강좌별 (비디오 이벤트 로그 수)/(전체 수강 시간) 지표가 차이가

있는지를 분석하였다. 먼저 대분야별 지표의 평균값은 [표 9]와 같다 ([그림 19] 참고). 그룹 간 분산분석을 진행한 결과, 유의미한 차이가 발견되지 않았다 (F=0.9, p=0.51, α=0.05).

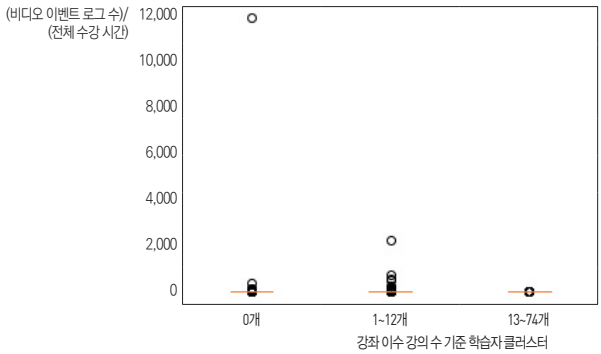
〈표 9〉 강좌 대분야별 (비디오 이벤트 로그 수)/(전체 수강 시간) 평균 값

강좌 대분류	공학	교육	사회	예체능	의약	인문	자연
지표 평균 값	0.15	0.38	0.05	0.07	0.04	0.06	0.9



[그림 19] 강좌 분야별 (비디오 이벤트 로그 수)/(전체 수강 시간) 분포도

이번에는 앞서 2-(2)에서 이수 강좌 수로 학습자를 나눈 것을 바탕으로 하여 각 집단별 (비디오 이벤트 로그 수)/(전체 수강 시간) 지표에 차이가 있는지를 분석하였다. 먼저, 각 집단 별 평균 (비디오 이벤트 로그 수)/(전체 수강 시간) 값에서는 이수 강좌 수가 0개인 학습자들이 0.26, 이수 강좌 수가 1~12개인 학습자들이 0.25, 이수 강좌 수가 13~74개인 학습자들이 0.05를 보였으며 ([그림 20] 참고), 그룹 간 분산분석을 진행한 결과 유의미한 차이가 발견되지 않았다(F=0.0, p=1.00, α=0.05).



[그림 20] 이수 강좌 수 클래스별 (비디오 이벤트 로그 수)/(전체 수강 시간) 분포도

3.3 단일 강좌에 대한 분석

(1) 분석 동기

약 1200개에 이르는 강좌에 대한 학습자의 학습 패턴을 분석하기에 앞서, 하나의 강좌에 대한 학습자의 로그 데이터 분석을 진행했다. 이를 통해 주어진 데이터에서 얻을 수 있는 정보가 무엇인지 구체적으로 알 수 있으며, 차후에 이어질 연구에 대한 아이디어를 얻을 수 있을 것이라 예상되어 단일 강좌에 대한 분석을 진행하게 되었다.

(2) 강좌 선택

특정 강좌 연구를 진행하기 위해 분석할 강좌로 "R을 활용한 통계학개론" 강좌를 선택하였다. 위 강좌는 부산대학교에서 제공하여 2018년 9월 3일에 개설한 강의로, 자연(수학 · 물리 · 천문 · 지리) 계열에 속하는 강의이다. [표 10]은 강좌별 로그 데이터 수를 내림차순으로 정렬한 것의 일부이고, [표 11]은 강좌별 학습자의 수를 내림차순으로 정렬한 것의 일부이다. 두 개의 표를 통해 확인할 수 있듯이, 본 강좌는 학습자의 이벤트 로그 데이터 수와 강좌를 수강한 학습자의 수가 모두 많은 강좌이기 때문에 케이스 분석을 하기에 적절한 강의로 판단할 수 있다.

〈표 10〉 강좌 id와 로그데이터 수

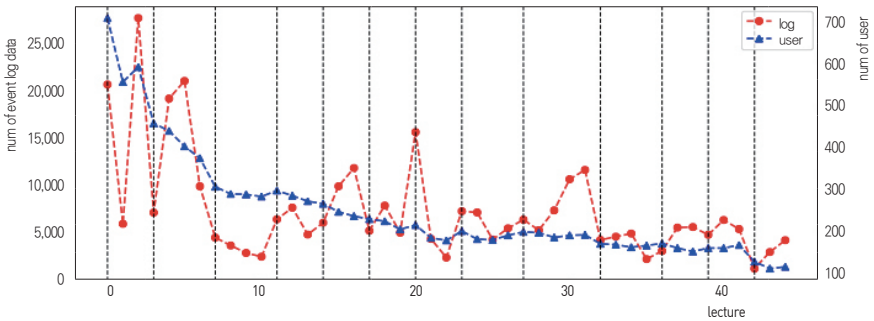
	강좌 id	로그 데이터 수
1	course-v1:EwhaK+EW11237K+2018_S06	462128
2	course-v1:HansungK+HSKMOOCk01+2018_1	420700
3	course-v1:EwhaK+EW11237K+2018_F06	410362
4	course-v1:SookmyungK+SM_soc_002k+2018_03SM_02	359357
5	course-v1:DGUl+DGUk_005k+DGU_005k_2018_3_5	358096
6	course-v1:PNUk+RS_C01+2018_KM007	339511
7	course-v1:EwhaK+EW36387K+2018_S03	316928
8	course-v1:DGUk+DGUk_005k+DGU_005k_2018_9_5	305832
9	course-v1:DGUk+DGUk_004k+DGU_004k_2018_3_4	303778
10	course-v1:EwhaK+EW36387K+2018_F03	295315

〈표 11〉 강좌 id와 학습자 수

	강좌 id	유저 수
1	course-v1:KAISTk+KCS470+2017_K0203	2540
2	course-v1:EwhaK+EW11237K+2018_S06	1446
3	course-v1:PNUk+RS_C01+2017_KM_009	1429
4	course-v1:SookmyungK+SM_sta_004k+2018_02SM_02	1300
5	course-v1:SKKUk+SKKU_2017_01+2018_SKKU03	1243
6	course-v1:SKKUk+SKKU_EXGB506_01K+2018_SKKU02	1101
7	course-v1:SNUk+SNU044_008k+2018_T1	1065
8	course-v1:EwhaK+EW11237K+2018_F06	1061
9	course-v1:SKP_POSTECHk+SKP_CHEB411k+2018_1	1032
10	course-v1:CAUk+ACE_CAU02_01+2018_T!	1024
11	course-v1:SookmyungK+SM_soc_002k+2018_03SM_02	1008
12	course-v1:PNUk+RS_C01+2018_KM_007	926

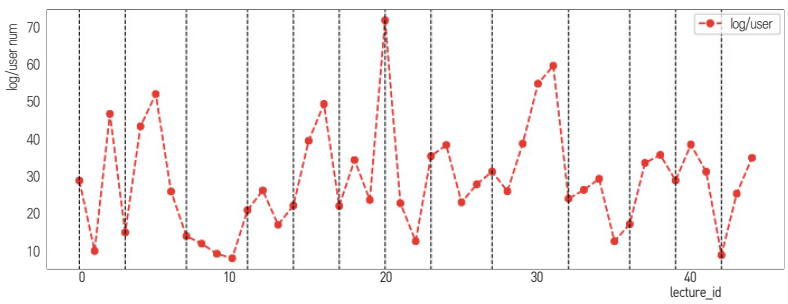
(3) 강좌 단위 분석

본 강좌는 총 13주차, 45개(lecture 0~ lecture 44)의 강의로 구성되어 있다. 강의의 진행에 따른 학습자 수 변화와 이벤트 로그 수 변화를 분석하기 위해서, [그림 21]과 같은 그래프를 그려보았다. x 축은 강의의 순서를 나타내고, 좌측의 y축은 강의에서 발생한 이벤트 로그의 개수, 우측의 y축은 강의의 학습자 수를 나타낸다. 또한 빨간색 그래프는 강의에 따른 이벤트 로그 개수의 변화를, 파란색 그래프는 학습자 수의 변화를 나타낸다. 또한 검정색 세로선은 새로운 주차의 강의가 시작되는 부분을 나타낸다.



[그림 21] 강의에 따른 학습자의 수와 발생한 이벤트 로그의 수 그래프

파란색 그래프가 지속해서 감소하는 것을 한눈에 확인할 수 있듯이, 강의의 후반부로 갈수록 학습자의 수가 적어지는 패턴을 보인다. 이는 MOOC 플랫폼에 오랫동안 지적되는 문제인 유저들의 참여도 저하를 단적으로 보여준다. 주목해볼 만한 부분은 강좌의 후반부로 갈수록 강의를 듣는 유저의 수가 줄어드는 경향을 보이는 가운데, 각 주차의 첫 강의는 수강한 유저의 수가 약간 증가하는 경향을 보이는 점이다. 발생하는 로그 데이터의 수는 유저의 수가 많은 강좌 초반부에 많이 발생하고, 유저의 수가 강좌 초반에 비해 매우 적은 강좌 후반부에서 적게 발생하는 것이 매우 자명하다. 따라서 유저의 수와 관계없이 강의 당 평균적으로 어느 정도 양의 로그 데이터가 발생하는지를 분석하기 위해 (발생한 로그 데이터 수) / (유저의 수)를 계산하여 새로운 척도로 사용하였다. 이 값을 이용하여 계산해본 결과, 학습자들은 강의 당 평균 29.5, 표준편차 13.8의 이벤트 로그를 발생시킨다는 통계를 얻었다.



[그림 22] 강의에 따른 (이벤트 로그 발생 횟수)/ (학습자) 값 그래프

앞선 그래프 [그림 21]에서 학습자들이 각 주차의 첫 강의를 보다 많이 수강하는 경향성을 보였다. 하지만, 각 주차의 첫 강의의 log/user 값은 다른 강의들보다 크지 않았고 오히려 각 주차의 중간에 위치하는 강의들이 큰 값을 갖는 것을 관찰할 수 있었다. 이를 종합하여 볼 때, '각 주차의 첫 강의의 경우 보다 많은 학습자들이 강의 동영상에 들어가보지만, 그 중 많은 수는 학습을 제대로 진행하지 않는다'는 가설을 세울 수 있었다. 이는 후속연구를 통해 검증이 가능할 것이라고 예상된다.

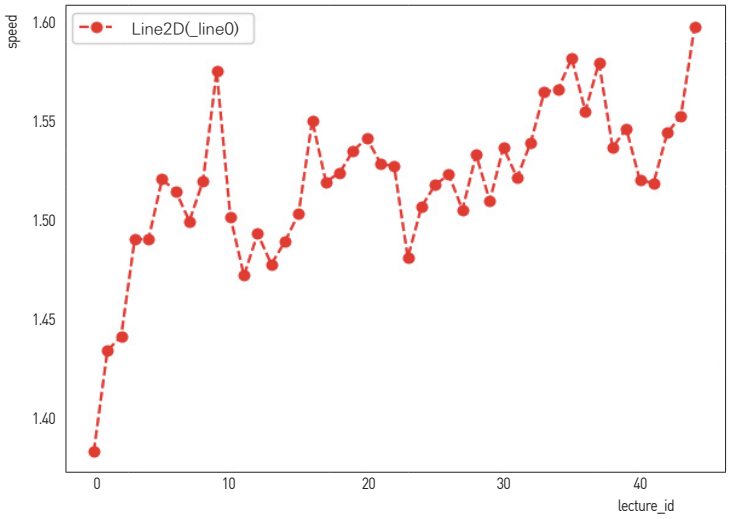
앞서 이벤트 로그 데이터는 8가지 이벤트 타입(Play_video, Pause_video, Seek_video, Speed_change, Video_show_cc_menu, Video_hide_cc_menu, Show_transcript, Hide_transcript)중 하나를 가짐을 언급했다. 이중 seek_video 이벤트를 강의의 뒷부분으로 이동한 seek_backward 이벤트와 앞으로 이동한 seek_forward 이벤트로 나누고, speed_change 이벤트 또한 속도가 전보다 빨라진 speed_up 이벤트와 속도가 느려진 speed_down 이벤트로 나누어 총 이벤트 타입을 10가지로 세분화했다. 또한 이벤트 타입별 각 강의에서 발생한 횟수를 분석해 어떤 이벤트 타입들이 서로 관련이 있는지 밝혀보고자 하였다. 이를 위해 이벤트 타입별 발생 횟수에 대한 Pearson상관관계 분석을 진행했으며 결과는 [표 12]와 같다.

〈표 12〉 event_type들의 correlation coefficient

	pause	seek forward	seek backward	speed up	speed down	show cc	hide cc	show transcript	hide transcript
pause	1	0.92	0.97	0.82	0.95	0.81	0.81	0.53	0.55
seek forward	0.92	1	0.95	0.8	0.91	0.71	0.71	0.51	0.49
seek backward	0.97	0.95	1	0.79	0.92	0.76	0.76	0.47	0.48
speed up	0.82	0.8	0.79	1	0.92	0.94	0.94	0.87	0.89
speed down	0.95	0.91	0.92	0.92	1	0.87	0.87	0.67	0.67
show cc	0.81	0.71	0.76	0.94	0.87	1	1	0.83	0.88
hide cc	0.81	0.71	0.76	0.94	0.87	1	1	0.83	0.88
show transcript	0.53	0.51	0.47	0.87	0.67	0.83	0.83	1	0.96
hide transcript	0.55	0.49	0.48	0.89	0.67	0.88	0.88	0.96	1

반대 관계에 있는 이벤트들(ex. hide_transcript, show_transcript)은 예상대로 강한 양의 상관관계가 있음을 확인할 수 있다. 주목해볼 점은 pause, seek_backward, speed_down 3개의 이벤트 발생에도 강한 양의 상관관계가 있다는 점이다. 현재까지는 왜 위 이벤트들의 발생에도 상관관계가 있는지는 알 수 없지만, 후속 연구를 통해 위 3개의 이벤트 발생과 강의의 복잡도를 연관 지어 본다면 흥미로운 결과를 얻을 수 있으리라 추측된다.

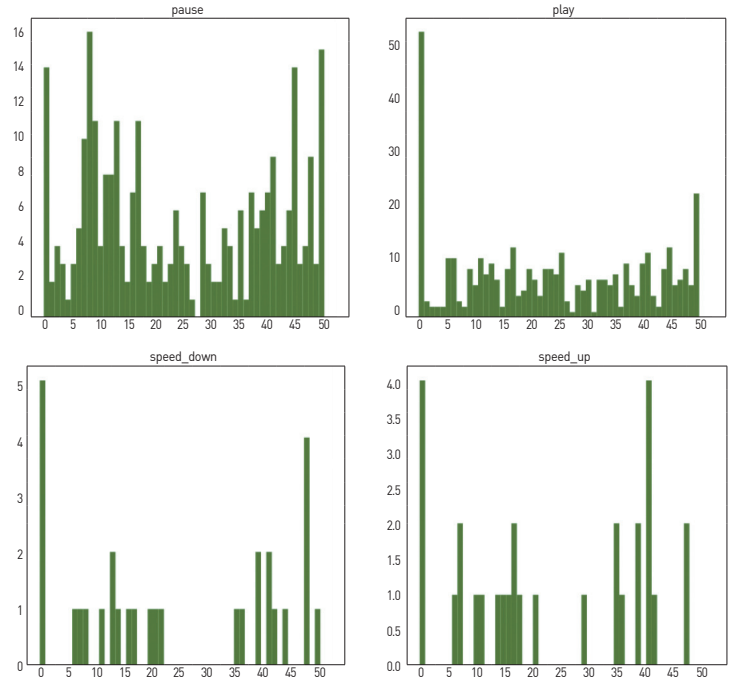
강의별 학습자들의 비디오 재생 평균 속도의 변화는 [그림 23]와 같았다. 강의에 후반부로 갈수록 비디오 재생 평균 속도가 빨라지는 경향성을 보였는데, 강좌 후반부를 듣는 학습자들이 초반부를 듣는 학습자들에 비해 매우 소수이고, 강좌를 끝까지 듣고자 하는 사람들이 다수일 것이라는 점을 고려하면 향후 관련 분석이 이루어질 수 있을 것으로 예상된다.



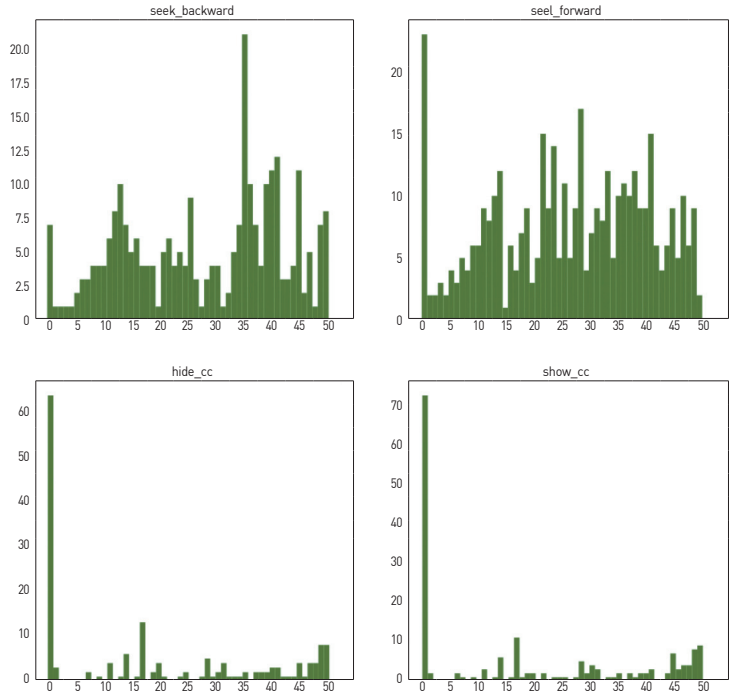
[그림 23] 강의별 학습자들의 평균 속도

(4) 강의 단위 분석

위 그래프 [그림 22]에서 확인할 수 있다시피, 강의 20 (21번째 강의)에서 가장 많은 양의 log/user가 발생하였다. 강의 20에서의 log/user 값은 log/user의 평균 + 2 * log/user의 표준편차 값보다도 큰 것으로 나타났다. K-MOOC 사이트에서 확인해본 결과, 이 강의는 week6[5-1장] 통계적 추론, 5-2. "구간추정 I 대표본" 강의였다. 위 강의에서 이벤트 로그가 특별히 많이 발생한 이유가 무엇인지, 어떤 학습자들의 행동이 있었는지, 그리고 강의 20에서 각 이벤트가 어느 위치에서 얼마나 발생했는지 분석하기 위해 강의의 진행에 따른 각 이벤트의 발생 횟수에 대한 히스토그램([그림 24], [그림 25])을 그려보았다. 하지만 히스토그램을 통해 아직까지 큰 통찰을 얻지는 못하였다.



[그림 24] 강의20에서 각 이벤트별 발생 위치 및 빈도



[그림 25] 강의20에서 각 이벤트별 발생 위치 및 빈도

4
결론

본 연구는 K-MOOC의 데이터를 이용해 플랫폼의 학습자들을 보다 실증적으로 분석하고, 이를 바탕으로 플랫폼 개선 계획 수립을 돕는 것을 목표로 진행되었다. K-MOOC 데이터 분석 과정은 데이터 분석 파이프라인 구축, 데이터 전처리, 전처리된 데이터 분석의 과정으로 이루어졌으며 이는 차후의 K-MOOC 데이터 분석 연구에서도 활용할 수 있도록 설계된 것이다. 또한 데이터 분석에서는 데이터의 전반적인 이해를 위한 분석, 학습자의 행동 및 정보에 대한 자세한 분석, 단일 강좌에 대한 사례 분석을 진행하였다.

본 연구는 특정 주제에 대한 깊이 있는 분석보다는 간단한 통계적, 알고리즘적 방법들을 사용하여 여러 가지 측면을 관찰하는 방향으로 진행되었다. 이는, 본 연구가 K-MOOC 데이터를 활용한 첫 연구인만큼 K-MOOC 데이터에 대한 통찰을 얻고, 후속 연구에서 발전 가능한 측면을 찾는 일에 집중하기 위함이었다. 해당 플랫폼에 적합한 시각화, 분석 기준 등을 정립하여 분석하는 등 해당 연구는 K-MOOC 플랫폼의 현황뿐만 아니라, 앞으로의 K-MOOC 연구의 방향성 또한 제시하고 있다.

참고문헌

· Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. JEDM| Journal of Educational Data Mining, 1(1), 3-17.

Bienkowski, M., Feng, M., & Means, B. (2012). Enhancing teaching and learning through educational data mining and learning analytics: An issue brief. US Department of Education, Office of Educational Technology, 1, 1-57.'

Chuang, I., & Ho, A. (2016). HarvardX and MITx: Four Years of Open Online Courses--Fall 2012-Summer 2016.

Educational Data Mining. (n.d.). Retrieved February 25, 2019, from http://educationaldatamining.org/

Guo, P. J., Kim, J., & Rubin, R. (2014, March). How video production affects student engagement: An empirical study of MOOC videos. In Proceedings of the first ACM conference on Learning@ scale conference (pp. 41-50). ACM.

Ho, A., Reich, J., Nesterko, S., Seaton, D., Mullaney, T., Waldo, J., & Chuang, I. (2014). HarvardX and MITx: The first year of open online courses, fall 2012-summer 2013.

Ho, AD, Reich, J., Nesterko, S., Seaton, DT, Mullaney, T., Waldo, J., & Chuang, I.(2014). HarvardX and MITx: The first year of open online courses (HarvardX and MITx Working Paper No. 1).

KDD Cup 2015. (2015). Retrieved February 25, 2019, from https://biendata.com/competition/kddcup2015/

Kim, J., Guo, P. J., Seaton, D. T., Mitros, P., Gajos, K. Z., & Miller, R. C. (2014, March). Understanding in-video dropouts and interaction peaks in online lecture videos. In Proceedings of the first ACM conference on Learning@ scale conference (pp. 31-40). ACM.

Kizilcec, R. F., Piech, C., & Schneider, E. (2013, April). Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In Proceedings of the third international conference on learning analytics and knowledge (pp. 170-179). ACM.

Papamitsiou, Z., & Economides, A. A. (2014). Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. Journal of Educational Technology & Society, 17(4), 49-64.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. Expert systems with applications, 33(1), 135-146.

Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40(6), 601-618.

Romero, C., & Ventura, S. (2017). Educational data science in massive open online courses. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 7(1), e1187.

Ruipérez-Valiente, J. A., Muñoz-Merino, P. J., Pijeira Díaz, H. J., Santofimia Ruiz, J., & Delgado Kloos, C. (2017). Evaluation of a learning analytics application for open EdX platform. Computer Science & Information Systems, 14(1), 51-73.

Zhang, T., & Yuan, B. (2016, October). Visualizing MOOC User Behaviors: A Case Study on XuetangX. In International Conference on Intelligent Data Engineering and Automated Learning (pp. 89-98). Springer, Cham.

부록

부록 1 로그 데이터 항목

본 연구는 2018년 1월 1일 ~ 2018년 12월 31일 기간 안에 발생한 K-MOOC 학습자의 이벤트 로그 데이터를 중심으로, 해당 학습자와 강좌에 대한 기본 정보 데이터를 모두 활용하여 진행되었다. 동영상 강의의 transcript 데이터를 제외한 모든 데이터는 국가평생교육진흥원에 의해 수집 및 제공 되었으며, transcript 데이터의 경우 위임 받은 베타 테스터 권한을 활용해 직접 K-MOOC 웹 사이트를 크롤링하여 수집하였다.

MOOC는 edX 플랫폼 표준 API[2]에 따른 JSON(JavaScript Object Notation) 형식으로 된 이벤트 로그 파일을 남긴다. 이벤트 로그는 강의, 학습자, 학습자의 학습 환경, 발생 이벤트 등에 대한 다양한 정보를 포함한다. 이때, 사전에 이벤트 로그 파일에서 학습자 ID가 해쉬처리가 되는 등의 개인 비식별화 과정이 진행된다. 아래의 표와 같이 이벤트 로그에는 다양한 정보가 저장된다.

2018년 한 해 동안에 수집된 이벤트 로그 데이터의 총 크기는 약 240GB 였다. 해당 데이터 내에는 올바르지 않은 형식으로 저장된 데이터, 특정 속성 값이 들어있지 않은 데이터 등 오류가 있는 데이터도 포함되어 있었으며, 오류를 포함한 데이터는 데이터 전처리 과정을 통해 분석에서 제외되었다.

2) https://edx.readthedocs.io/projects/devdata/en/latest/internal_data_formats/event_list.html#event-list

〈표 13〉 K-MOOC의 이벤트 로그 예시


```
{
        "accept_language": "ko-kr",
        "agent": "Mozilla/5.0 (Linux; Android 7.1.2; LGM-V300L Build/N2G47H)
AppleWebKit/537.36 (KHTML, like Gecko) Chrome/50.0.2661.102
Crosswalk/20.50.533.55 Mobile Safari/537.36 NAVER(inapp; search; 590; 8.8.2)",
        "context": {
                "course_id": "course-v1:SookmyungK+SM_sta_004k+2018_02SM_02",
                "org_id": "SookmyungK",
                "path": "/event",
                "user_id":
"2a8511ec909f8a7748fd55aea2bf71686c6d9efe07e239e217c9624fd47d239f"
        },
        "event": {
                "code": "html5",
                "id": "a5256fd3fe95437a81cb8d626df14ac4",
                "new_time": 113,
                "old_time": 3.783938,
                "type": "onSlideSeek"
        },
        "event_source": "browser",
        "event_type": "seek_video",
        "host": "www.kmooc.kr",
        "ip": "***.***.***.***",
        "name": "seek_video",
        "page":
"http://www.kmooc.kr/courses/course-v1:SookmyungK+SM_sta_004k+2018_02SM_02
/courseware/108b063e366744d0ba97c50e9ce4f246/91b88af65e8a4006907bbf3c530ba
841/",
        "referer":
"http://www.kmooc.kr/courses/course-v1:SookmyungK+SM_sta_004k+2018_02SM_02
/courseware/108b063e366744d0ba97c50e9ce4f246/91b88af65e8a4006907bbf3c530ba
841/",
        "session": "5ce7c43a92d6a3df1185a30244da295a",
        "time": "2018-07-24T10:18:05.757844+00:00",
        "username": "******"
}
```


여기서, 각 로그별로 다음과 같은 항목들을 포함한 정보가 저장된다.

〈표 14〉 edX(K-MOOC의 기반)의 로그 목록

accept_language	학습자가 이해할 수 있는 언어 (HTTP상)	event_type	이벤트 종류
agent	학습자의 브라우저 등	page	강좌 URL
course_id	강좌 ID	session	학습자의 session 코드
org_id	강의를 제공하는 기관 ID	time	이벤트가 발생한 시각
user_id	학습자를 식별할 수 있는 암호화된 ID	event	event_type에 따른 세부사항
표준편차	1160.9	3.14	47.2

〈표 15〉 speed_change_video 이벤트의 event 필드

```
"event": {
        "code": "html5",
        "current_time": 45.7893082,
        "id": "f5c4c6c9d5d04129b6a5ddae33100a6e",
        "new_speed": "2.0",
        "old_speed": "1.50"
}
```

〈표 16〉 speed_change_video 이벤트의 event 필드 설명

항목	설명
current_time	이벤트가 발생했을 때의 비디오 내 위치(s)
new_speed	바뀐 속도
old_speed	기존 속도

〈표 17〉 seek_video 이벤트의 event 필드

```
"event": {
        "code": "html5",
        "id": "a5256fd3fe95437a81cb8d626df14ac4",
        "new_time": 113,
        "old_time": 3.783938,
        "type": "onSlideSeek"
}
```

〈표 18〉 seek_video 이벤트의 event 필드 설명

항목	설명
new_time	바뀐 시간
old_time	기존 시간

본 연구에서는 동영상 강의를 시청하는 K-MOOC 학습자들의 클릭을 통해 발생하는 이벤트로그(Event_type항목이 Play_video,Pause_video, Seek_video, Speed_change_video, Video_show_cc_menu, Video_hide_cc_menu, Show_transcript, Hide_transcript 인 이벤트 로그)를 분석하는 것을 목표로 하였다.

부록 2 기타 데이터 크롤링

다량의 K-MOOC 로그 데이터를 분석하는 데에 있어서, 본 연구는 각 강좌의 자막(Caption)을 추출하여 각 현상을 설명하고자 하였다.

먼저, 자막 데이터를 분석하는 것은 K-MOOC 학습자들의 패턴을 보다 실증적으로 분석하는 데에 의의가 있다. 영상의 특정 부분에서 학습자들이 다른 범위와는 다른 패턴을 보인다면, 강좌의 특정 부분이 학습자들에게 어떠한 영향을 미치는지 직접 확인해볼 필요가 있다. 예를 들어 한 영상의 특정 부분에서 View가 급격하게 증가했다면, 이 부분에서의 내용은 다른 부분과 다른 양상을 보일 수 있기 때문에 반드시 확인해보아야 한다. 강의 내용을 분석하려면 그 부분에서의 문맥을 파악해야 하는데, 본 연구에서는 '자막'을 통해 해당 문맥을 파악하고 설명하고자 하였다.

또한, 자막은 전체적인 흐름의 기준이 되기 때문에 이를 이용하여 분석하는 것이 중요하다. 자막은 [시작점(시간) : 자막(문자열)]의 키 : 값 배열로 이루어져 있다. 강의자의 발화 속도는 수강생의 영상 시청 패턴에 많은 영향을 미치는데[citation needed], 강의자의 발화 속도를 이용하여 패턴을 분석하기 위한 지표로 자막 분석이 이용될 수 있다. 예를 들어, 교수자가 1:00부터 1:06까지 "I study hard"라는 발화를 한다면, 해당 구간에서의 발화속도는 3(words)/0.1(minute) = 30wpm이 된다[3]. 영상의 각 구간에 이러한 특성을 부여하며 영상 전체의 패턴을 분석함으로써, 학습자들의 학습 패턴을 보다 객관적으로 분석할 수 있다. 이에, 본 연구에서는 각 강좌의 자막을 분석함으로써 설명하고자 한다.

3) 본 연구에서는 보다 한국어에 적합한 말하기 속도 특성을 도출하기 위하여, wpm (분당 발화 단어 수)가 아닌 특성 (cpm, character per minute, 분당 발화 글자 수)을 사용하였다.

본 연구에서는, K-MOOC 플랫폼 내 강좌들의 자막을 추출하기 위하여 크롤링(Crawling) 기법을 사용하였다. 크롤링이란, 웹에 존재하는 수많은 문서를 특정 기준으로 색인하여 필요한 정보를 가져오는 기술을 말한다. 특정 웹사이트에 접속했을 때 보이는 정보를 이용 가능한 기준으로 정렬하기 위해 크롤링을 이용해야 하는데, 본 연구에서는 각 강좌 영상의 자막을 크롤링 기법을 통해 추출하여 분석하였다.

크롤링은 다음과 같은 순서로 진행된다.

〈표 19〉 크롤링 프로세스 목록

순서	내용
1	전체 강좌 목록을 이용해, 각 강좌 웹에 접근한다
2	해당 강좌에 있는 모든 세부강좌 웹에 접근한다
3	각 세부강좌의 HTML 요소를 확인하여, 자막에 해당하는 부분을 Javascript 언어로 추출한다
4	해당 부분을 분석할 수 있도록 json 형식으로 정렬한다

위의 [표 19]의 순서에서 알 수 있듯이, 크롤링 기법은 웹 요소(HTML)와 이를 작동시키는 요소(Javascript), 그리고 이를 총괄하여 실행하고 결과물을 추출해내는 요소로 이루어져있다. 결과물을 추출할 때 다양한 언어와 방법으로 추출할 수 있는데, 본 연구에서는 Python 언어 내 Selenium[4] 라이브러리를 이용하여 Chrome 웹 브라우저를 자동화하는 방식으로 크롤링을 진행하였다.

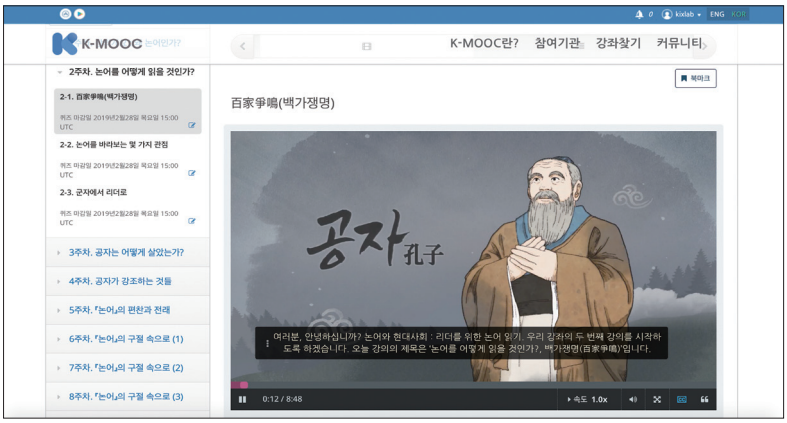
먼저, K-MOOC의 웹에서 자막은 다음과 같은 링크의 HTML 문서에 포함되어 있다.

〈표 20〉 HTML 링크

http://www.kmooc.kr/courses/강좌명/courseware/카테고리_식별자/강의_식별자/

예를 들어, 서울대학교 (기관명) 〉 논어와 현대사회 - 리더를 위한 논어 읽기 (강좌명) 〉 2주차. 논어를 어떻게 읽을 것인가? (카테고리명) 〉 2-1. 百家爭鳴(백가쟁명) (강의명) 영상을 위 링크와 같은 기준으로 접속하면, 다음과 같은 홈페이지가 로드된다.

4) https://www.seleniumhq.org/



[그림 26] 로드된 K-MOOC 홈페이지

해당 웹에서의 자막을 추출하기 위해서는 본 링크 HTML 상의 구조를 파악해야 한다. 모든 K-MOOC 강좌는 동일한 HTML 구조를 가지는데, 각 영상의 자막은 HTML 구조 내에 다음과 같이 존재한다.

〈표 21〉 자막이 포함된 HTML 형식

〈div class="subtitles" ~~〉

또한, 위 영역의 하위 요소를 분석해보면 다음과 같은 자막이 모여 있는 집합을 추출할 수 있다.

〈표 22〉 자막 HTML 형식

〈li role="link" data-index="자막_목차" data-start="시작시간" tabindex="0" class="current"〉자막내용〈/li〉

HTML 문서 내의 이러한 자료를 추출하기 위해, 위 영역을 조작할 수 있도록 다음과 같은 Javascript 스크립트를 실행할 수 있다.

〈표 23〉 자막을 가져오는 Javascript 코드

```
var dicObject = {};
for (var i = 1; i < document.getElementById('transcript-captions').children.length - 1; i++) {
dicObject[parseInt(document.getElementById('transcript-captions').children[i].getAttribute('data-start
') / 1000)] = document.getElementById('transcript-captions').children[i].textContent;
}
console.log(dicObject)
```

수만 개에 해당하는 영상에서 위와 같은 작업을 통해 자막을 추출하기 위해, 본 연구에서 다음과 같은 파이썬 코드로 위 과정을 자동화하여 모든 자막을 json 형태로 추출하였다.

〈표 24〉 자막을 가져오고 Json으로 반환하는 Python 코드

```
from selenium import webdriver #import chrome webdriver
from time import sleep #explicit sleeping method
import json #convert to json format

def subtitle(courseURLList):
    driver = webdriver.Chrome('chromedriver')
    driver.get("https://www.kmooc.kr/login")
    sleep(1)
    id = driver.find_element_by_id("login-email")
    id.send_keys(아이디)
    pw = driver.find_element_by_id("login-password")
    pw.send_keys(비밀번호)
    driver.find_element_by_xpath("//*[@id=\"login\"]/button").click()
    sleep(1) #LOGIN

    for url in courseURLList:
        driver.get(url)
        sleep(1)
        try:
            driver.find_element_by_xpath('//*[@id="audit_mode"]').click()
            sleep(1)
            driver.find_element_by_xpath('/html/body/div[6]/div/div[2]/div[2]/button').click()
            sleep(1)
        except:
            driver.get(url)
            sleep(1)

        tempCourseList = (driver.execute_script(
            "var dict = []; for (var i=0; i <
document.getElementsByClassName('accordion-nav').length;
i++){dict.push(document.getElementsByClassName('accordion-nav')[i].href)} return dict"))
        for item in tempCourseList:
            driver.get(item)
            sleep(5)
            try:
                with open(path + item.split("/")[len(item.split("/")) - 2] + ".json",
                          'a') as writeFile:
                    print(item.split("/")[len(item.split("/")) - 2])
                    execute_file = driver.execute_script(
                        "var dicObject = {}; for(var i = 1; i <
document.getElementById('transcript-captions').children.length-1;
i++){dicObject[parseInt(document.getElementById('transcript-captions').children[i].getAttribute('data-st
art')/1000)]=document.getElementById('transcript-captions').children[i].textContent;} return
(dicObject)")
                    dict = {int(k): v for k, v in execute_file.items()}
                    tempFile = json.dumps(dict, sort_keys=True)
                    writeFile.write(tempFile)
            except:
                print('quiz_or_no_subtitle')
subtitle()
```

먼저, 일반적으로 발화 속도를 정할 때에는 일반적으로 WPM(분당 발화 단어 수)를 이용한다. 대부분의 영미 문화권에서 해당 기준을 사용하는데, 이는 단위 시간동안 의미의 최소 단위인 단어를 얼마나 많이 발화했는지가 특정 화자의 의미 전달 속도를 잘 나타낼 수 있다는 데에서 비롯된다.

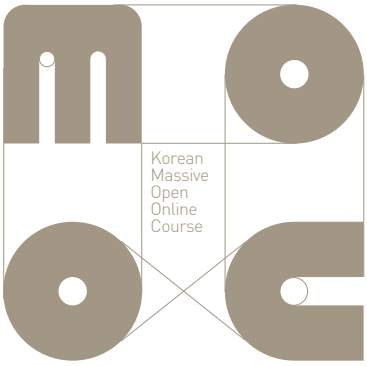
하지만, 의미가 한 글자로 축약된 한자가 많이 적용되는 한국어에서는, 실제로 청자가 화자의 발화를 이해하는 기준으로서 단어 수보다는 글자 수가 보다 적합하다. 이에, 본 연구에서는 각 범위에서 WPM 대신 CPM(분당 발화 글자 수)을 사용하였다. 이를 추출하기 위해, 이전 과정에서 추출했던 자막을 다음과 같은 코드를 이용하여 정렬하였다. (환경: Python 3.7)

〈표 25〉 자막에서 글자 수와 단어 수를 Json으로 반환하는 Python 코드

```
import json
import os

def charcount(string):
    return len(string)-string.count(' ')

def wordcount(string):
    return len(string.split(' '))

def write(course_id, lecture_id):

    with open(Path'+course_id+'/'+lecture_id, 'r') as readfile:
        temp = json.load(readfile)
        dictlist = [[k, v] for k, v in temp.items()]

        tot = []

        for idx, item in enumerate(dictlist[0:len(dictlist) - 1]):
            tot.append({"start_time": item[0], "end_time": dictlist[idx + 1][0], "char_count":
str(charcount(item[1])),
                        "word_count": str(wordcount(item[1]))})

    with open(Path+course_id+'/'+lecture_id, 'w') as writefile:
        writefile.write(json.dumps(tot))

curIdx = 0

for dir in os.listdir(Path):
    try:
        for jsondir in os.listdir(Path + dir):
            curIdx += 1
            try:
                write(dir, jsondir)
            except:
                print('file error')
    except:
        print('directory error')
```

K-MOOC 학습데이터 분석 및 활용방향 탐색

An Analysis of K-MOOC Learners' Data and an Investigation of Its Future Applications



국가평생교육진흥원
National Institute for Lifelong Education