



## 초등수학강의 분류&추천 서비스

이름	곽태경	이수지	전동인
소속	빅데이터 9기	빅데이터 9기	빅데이터 9기

### 요약

수학 강의 영상을 음성 파일로 변환 후, 음성 파일에서 강의 내용을 텍스트로 추출. 추출한 텍스트에서 많이 나타난 글자의 빈도에 따라 키워드를 선정(openai 이용) 초등 수학 정규과정 단원에서 나오는 걸 기준으로 카테고리 정하고 그 카테고리에서 각 강의 내용과 맞는 키워드를 뽑아내서 강의에 태그를 부여. 그리고 그 정보들을 db에 저장함. 그리고 태그 검색하면 관련 영상 목록 출력하게 함. 그리고 라벨링한 데이터를 가지고 새로운 영상을 분류하는 딥러닝 모델을 구현하여 db에 데이터를 추가하는 기능도 구현.

## 도입

### 기획 의도

- 배경
  - 현재, 국내에는 수많은 강의 플랫폼이 있고 학생들은 플랫폼의 인지도, 선생님의 커리큘럼 등을 고려하여 강의를 선정합니다. 하지만 공부를 하면서 이미 배웠던 특정 영역에 대한 복습이 필요한 경우가 있는데 이럴 경우 해당 영역만 따로 복습하기에는 선택지가 너무 많은 상황입니다. 뿐만 아니라 교육과정의 구성은 조금 다르지만 외국 강의로도 복습에 효과적일 수 있고, 외국 강의를 복습이 필요한 영역에 맞게 선정하는 것은 어려운 일입니다. 해당 서비스에서는 학생들이 플랫폼과 한국 수학 교육 과정에 맞는 카테고리를 입력했을 때 관련 영상이 조회되는 수학 강의 허브를 만들어보고자 합니다.
  - 기존 서비스

한국 학년별 수학	수학	미국 학년별 수학
초등 1학년 1학기	기초 수학	칸아카데미 키즈
초등 1학년 2학기	연산	미국 유치원
초등 2학년 1학기	기초 대수학 (Pre-algebra)	미국 1학년
초등 2학년 2학기	대수학 입문 (Algebra basics)	미국 2학년
초등 3학년 1학기	대수학 1	미국 3학년
초등 3학년 2학기	대수학 2	미국 4학년
초등 4학년 1학기	삼각법	미국 5학년
초등 4학년 2학기	기초 미적분학	미국 6학년
초등 5학년 1학기	미분학	미국 7학년
초등 5학년 2학기	적분학	미국 8학년

해당 사진은 칸 아카데미에서 발췌한 사진입니다. '한국 학년별 수학'이라는 항목에서 한국 초등학교 커리큘럼에 맞게 칸 영상을 재분류 해놓은 모습을 볼 수 있습니다. 그러나 원하는 교육과정에 해당하는 영상을 찾으려면 각 학년에 들어가서 직접 확인해야 한다는 불편함이 있습니다. 그렇기에 영상이 다 모여있고 원하는 카테고리를 입력했을 때 필요한 영상이 바로 호출되는 서비스가 있으면 좋겠다고 생각했습니다.

- 플랫폼 선정 > EBS, 칸 아카데미

- 플랫폼 선정

- EBS

한국 대표 온라인 강의 플랫폼으로 공영 방송국입니다. 뿐만 아니라 한국의 수능 출제에 있어서도 연계를 하여 출제하는 입시에 큰 영향을 주는 플랫폼이고 많은 학생들이 이용하는 만큼 해당 서비스 구축에 있어서 이용하게 되었습니다.

- 칸 아카데미

비영리 온라인 교육 플랫폼이면서 미국 2만여개의 학교에서 교육 자료로 쓰고 있는 플랫폼입니다. EBS와 유사한 점이 많다고 생각했고, 수학의 경우 강의를 다양하고 미국 교육과정에 맞게 분류가 되어 있습니다. 추후 다양한 국가의 플랫폼을 추가할 계획이지만 교육과정에 맞게 다양한 영상이 존재한다는 점, 비영리 플랫폼이라는 점에서 EBS와 같이 활용하기 좋다고 판단했습니다.

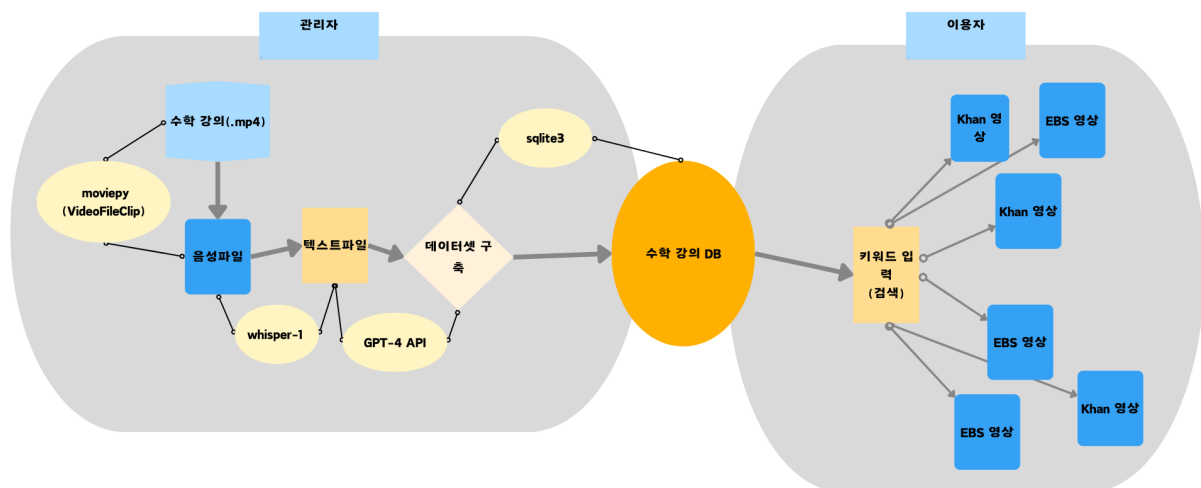
- 서비스

학생에게 맞는 교육을 위해서는 필요한 강의를 빠르게 매칭하는 것이 중요합니다. 사실 학생이 본인에게 필요한 강의를 선정할 때 그 선택지는 국내 강의에 국한되어 있습니다. 해외 강의는 편의상 유튜브 같은 플랫폼에서 쉽게 접근할 수 있는데 이 역시 교육 과정에 맞게 분류되어 있지 않아 빠르게 자료를 찾아 공부해야 하는 학생의 입장에선 적절하지 않을 수 있습니다.

그에 반해 강의 스타일에선 차이가 존재합니다. 학생마다 성취 수준이 다르기에 기존에 배웠던 내용을 다시 공부해야 할 경우도 많기 때문에 교육 과정에서의 5가지 키워드(수와 연산, 규칙성, 도형, 측정, 확률과

통계)를 입력하면 해당 과정과 관련 있는 영상을 호출할 수 있습니다. 추가적으로 교육 영상만 있다면 이용자가 검색해서 사용할 DB에 자동으로 키워드 맞게 분류되어 저장할 수 있습니다. 이용자와 관리자 입장에서 강의를 편하게 추가하고 원하는 강의를 편하게 이용할 수 있는 그런 서비스를 만들어보고자 하였습니다.

## 서비스 Flow Chart



## 데이터 정보(EDA)

### 1. 강의 영상

- 한국 : 총 88개

- EBS 유튜브 영상 135개(45강 분량이고 4개의 파트로 분리되어 있음, 유튜브에 영상들이 누락되어 있어 일부 주제는 내용 누락 있음.)

-EBS 사이트 영상 43강

강의	주제	누락된 파트
1강	네자리수와 큰 수	
2강	덧셈과 뺄셈의 여러가지 방법	
3강	곱셈의 계산 원리	
4강	나눗셈의 원리	
5강	혼합계산의 원리	
6강	분수 이해하기	6-2
7강	소수 이해하기	

8강	분수의 덧셈과 뺄셈 원리	
9강	소수의 덧셈과 뺄셈의 원리	
10강	평면도형 알아보기	10-3
11강	평면도형의 이동과 무늬 만들기	11-2
12강	원 알아보기	12-2
13강	삼각형 알아보기	13-1, 13-3
14강	수직과 평행	14-1, 14-3
15강	사각형과 다각형 알아보기	15-3
16강	길이와 시간 알고 합과 차 구하기	
17강	틀리와 무게 알고 합과 차 구하기	17-1, 17-2
18강	각도 알아보기	18-4
19강	평면도형의 둘레와 넓이 구하기	19-1, 19-3
20강	수의 범위와 어림하기1	20-2, 20-3
21강	수의 범위와 어림하기2	21-2, 21-3, 21-4
22강	자료정리의 방법	22-1, 22-3
23강	문제 해결의 여러가지 방법 찾기1	
24강	문제 해결의 여러가지 방법 찾기2	24-1
25강	약수와 배수	25-2
26강	약분과 통분	26-1
27강	분수의 덧셈과 뺄셈	27-3, 27-4
28강	분수의 곱셈	
29강	분수의 나눗셈	29-3, 29-4
30강	소수의 곱셈	
31강	소수의 나눗셈1	31-2
32강	소수의 나눗셈2	32-4
33강	분수와 소수의 혼합 계산	33-2
34강	평면도형의 넓이	34-4
35강	원주율과 원의 넓이	
36강	겉넓이와 부피	
37강	원기둥의 겉넓이와 부피	
38강	도형의 대칭	38-2
39강	자료의 표현과 해석	39-2, 39-3, 39-4
40강	비와 비율	40-2, 40-3
41강	비율 그래프	41-1, 41-3
42강	비례식	42-3
43강	연비와 비례배분	43-1, 43-3
44강	경우의 수와 확률	
45강	방정식	45-2, 45-4

...	...	
86강	자료의 정리(2)	
87강	자료의 정리(3)	
88강	자료의 정리(4)	

- 미국 : Khan Academy Kids 159개

강의	강의 제목
1강	Comparing numbers of objects
2강	Comparing numbers on the number line
3강	Counting by category
4강	Counting in pictures
5강	Counting objects 1
...	...
157강	Understanding place value when subtracting
158강	Understanding place value while subtracting
159강	Why a negative times a negative is positive

## 데이터 탐색

### EBS & Khan Academy Kids 강의 영상

#### 한국vs미국 강의 스타일 비교 분석

##### 목표 :

한국의 EBS와 미국의 칸 아카데미를 통해 각 플랫폼이 어떤 차이를 보이는지 확인해보고자 합니다. 한 강의 당 영상 시간은 얼마나 차이가 나는지, 초등 교육 과정을 타겟으로 하였을 때 각 카테고리 별로 어떤 분포를 보이는지, 영상에서 분당 글자 수는 어떻게 차이 나는지, 영상에서 나타나는 차이를 통해 두 플랫폼을 비교해보고자 합니다.

#### 1. 데이터(오디오파일) 변환

##### ▼ 1.1 mp3 형식의 오디오 파일을 wav 형식으로 변환

- 사용된 라이브러리: pydub
- 작업 설명: pydub의 AudioSegment를 사용하여 mp3파일을 wav형식이 음성 인식 라이브러리에 서 지원이 더 잘 되기 때문에 wav형식으로 변환하였습니다.

##### ▼ 1.2 오디오 파일에서 텍스트 추출

- 사용된 라이브러리: speech\_recognition, pydub
- 작업 설명: speech\_recognition를 사용하여 wav로 변환된 오디오 파일에서 음성을 텍스트로 추출 하였습니다. 이 과정에서 음성을 30초 간격으로 나누어 타임스탬프 정보와 함께 텍스트를 저장하였습니다.

니다.

- 텍스트 추출 시 타임스탬프 정보를 포함한 이유는, 시간 구간별로 세부 데이터를 수집함으로써 각 파일의 문자 밀도를 더 정확하게 분석할 수 있기 때문입니다. 구간 별로 노래가 나오거나 말을 하지 않는 부분에 대하여 제외시키고 말을 한 부분을 뽑아내기 위해 실시했습니다.

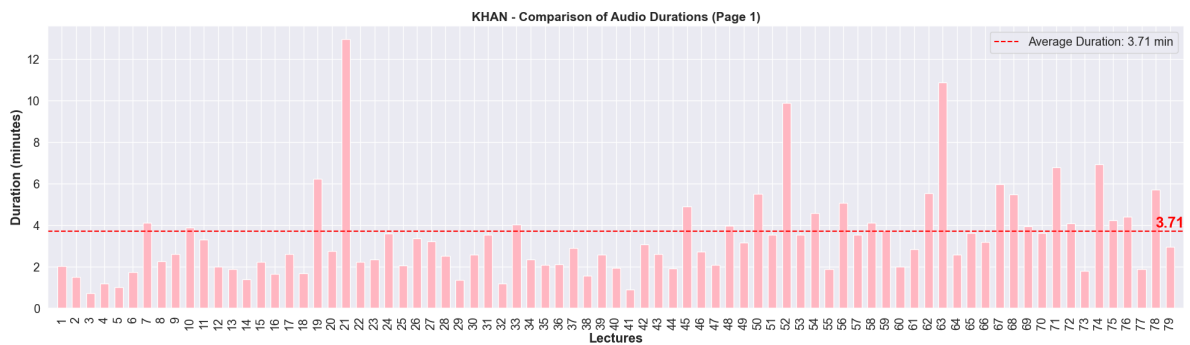
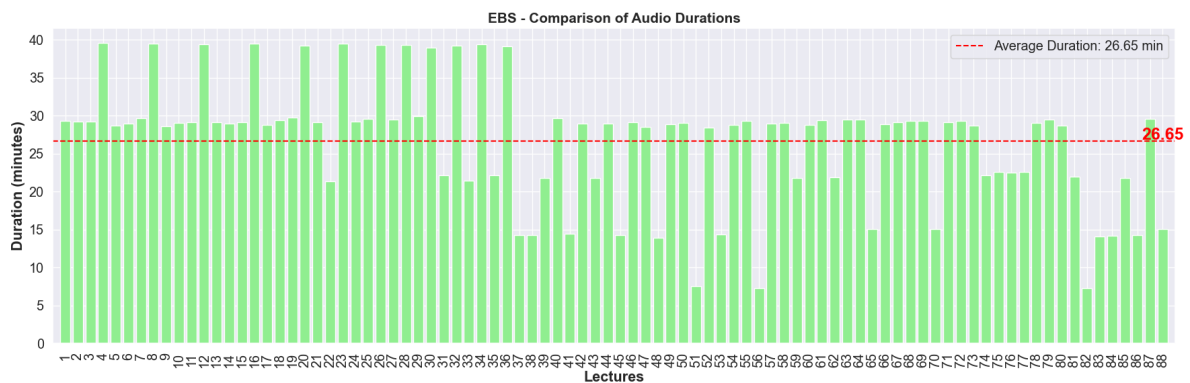
## 2. 데이터 분석 및 시각화(EDA)

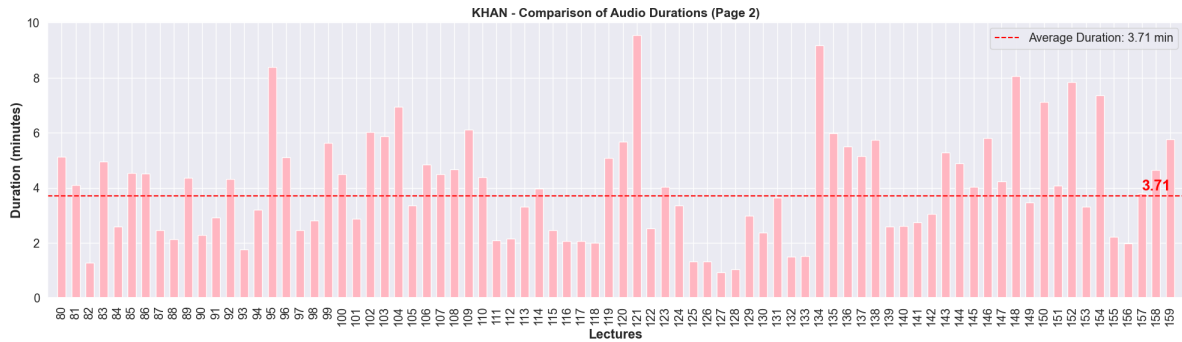
### ▼ 2.1.1 영상 길이 비교(오디오 파일 길이 추출)

- 사용된 라이브러리: os, pydub, re
- 작업 설명: 오디오 파일 mp3 형식의 각 오디오 파일의 길이를 분 단위로 계산하였습니다.

### ▼ 2.1.2 시각화

- 사용된 라이브러리: matplotlib, seaborn
- 작업 설명: 각 강의의 오디오 파일 길이를 막대 그래프로 시각화 하였습니다. 그래프에는 평균 길이를 나타내는 점선과 그 값을 텍스트로 표시하였습니다. x축은 간소화된 강의 번호로 설정하고, y축은 길이를 분 단위로 나타냅니다.





### ▼ 2.1.3 해석

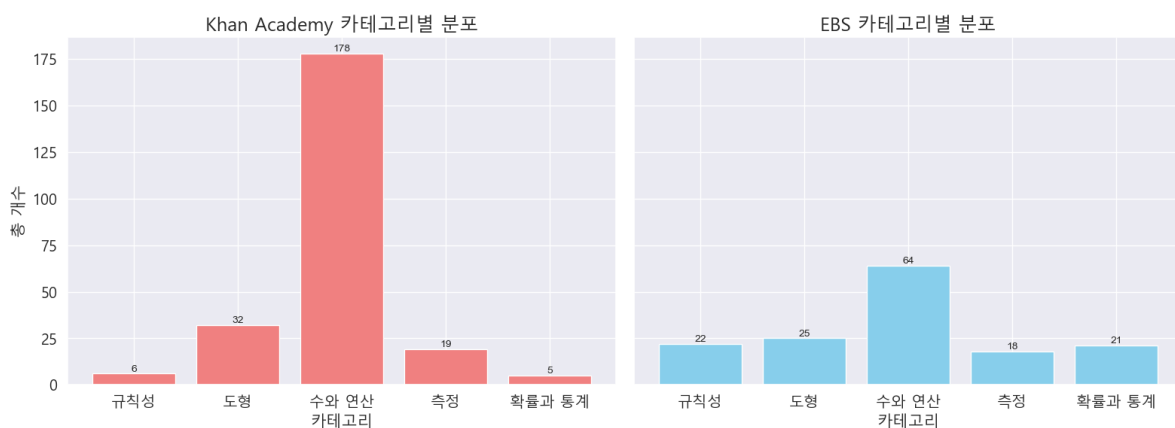
- EBS 강의 길이 평균은 26.65분으로, 대부분의 강의가 20~30분 사이에 분포되어 있습니다.
- KHAN 강의 길이 평균은 3.71분으로, 주로 2~4분 사이에 분포되어 있습니다.
- EBS 강의 길이는 변동폭이 크지 않게 일관된 길이의 강의를 하고, EBS와는 다르게 KHAN 강의 길이는 다양한 길이의 강의를 제공함을 알 수 있습니다.

### ▼ 2.2.1 카테고리 분포

- **사용된 라이브러리:** pandas, collection
- **작업 설명:** 모든 각 강의마다 연결해 놓은 카테고리 csv 파일에서 데이터를 불러와 Counter를 사용하여 각 카테고리의 빈도를 계산했습니다.

### ▼ 2.2.2 시각화

- **사용된 라이브러리:** matplotlib
- **작업 설명:** 각 카테고리의 빈도를 막대 그래프로 시각화하였습니다. 그래프에는 각 막대 위에 해당 개수를 표시하였고, x축은 카테고리, y축은 총 개수를 나타냅니다.



### ▼ 2.3.3 해석

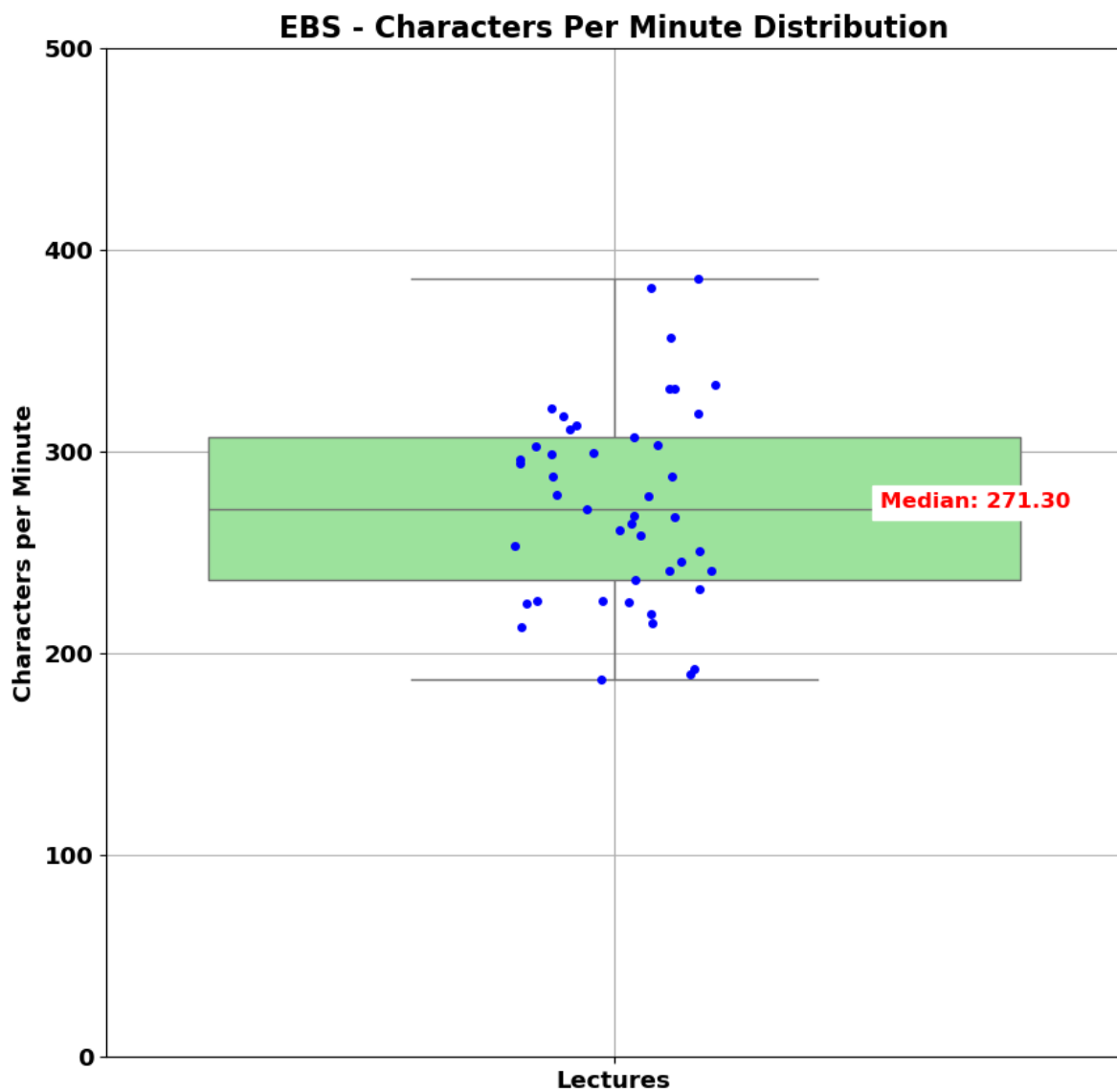
- EBS와 KHAN 강의 영상을 5개의 카테고리('규칙성', '도형', '수와 연산', '측정', '확률과 통계')로 나누었을 때 '수와 연산' 카테고리가 가장 많은 분포를 보였습니다.
- 카테고리별 항목 개수는 최소 5개에서 최대 178개까지 분포되어 있습니다.

### ▼ 2.3.1 분당 문자 수 계산 및 비교

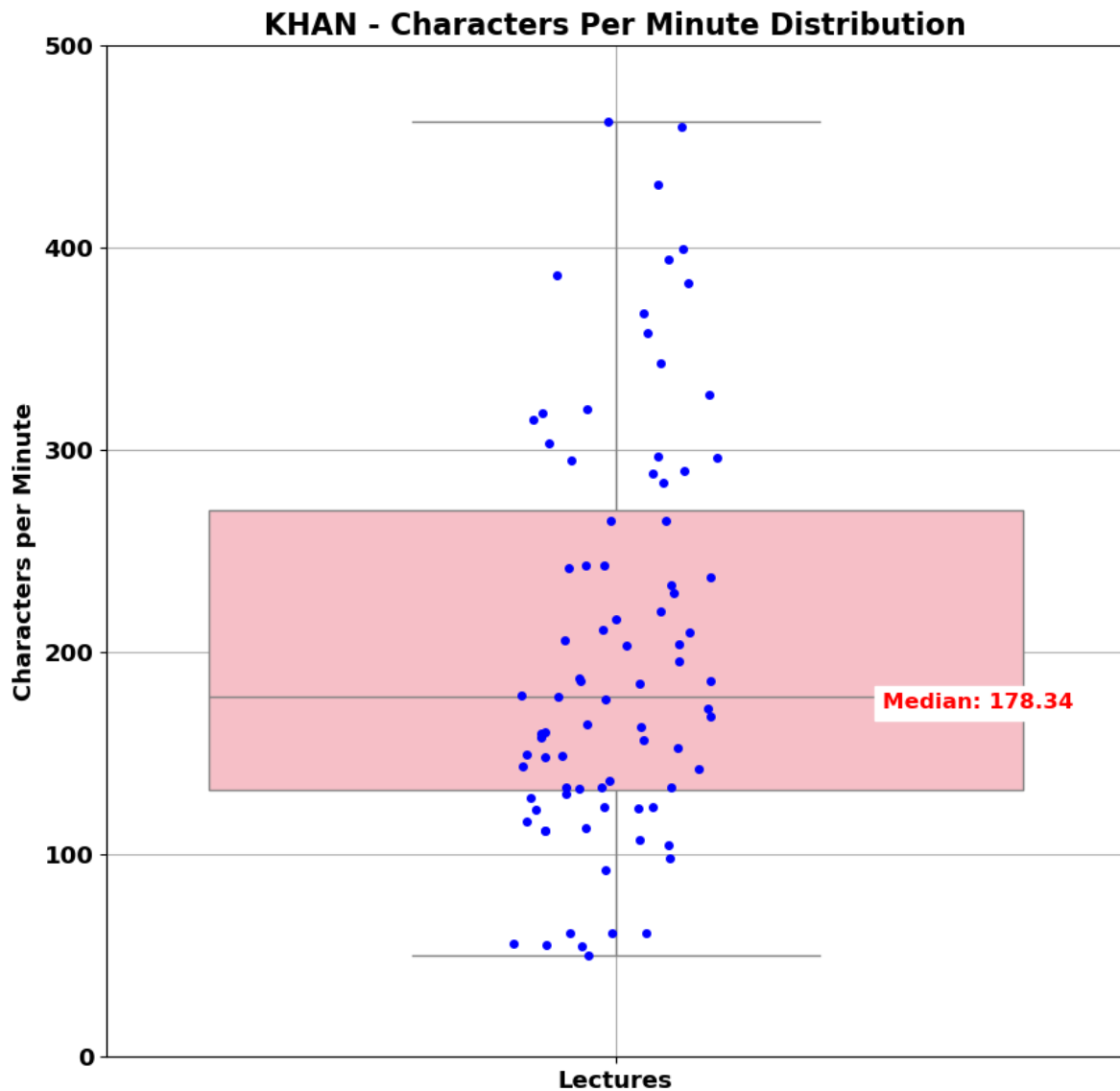
- 사용된 라이브러리: pydub, librosa
- 작업 설명: 각 강의의 오디오 길이와 해당 텍스트 파일의 총 글자 수를 이용하여 분당 문자 수를 계산 하였습니다.

### ▼ 2.3.2 시각화(박스플롯 및 스트립플롯 생성)

- 사용된 라이브러리: matplotlib, seaborn
- 작업 설명: seaborn을 사용하여 각 강의의 분당 문자 수를 박스플롯과 스트립플롯으로 시각화하였습니다.







#### ▼ 2.3.3 해석

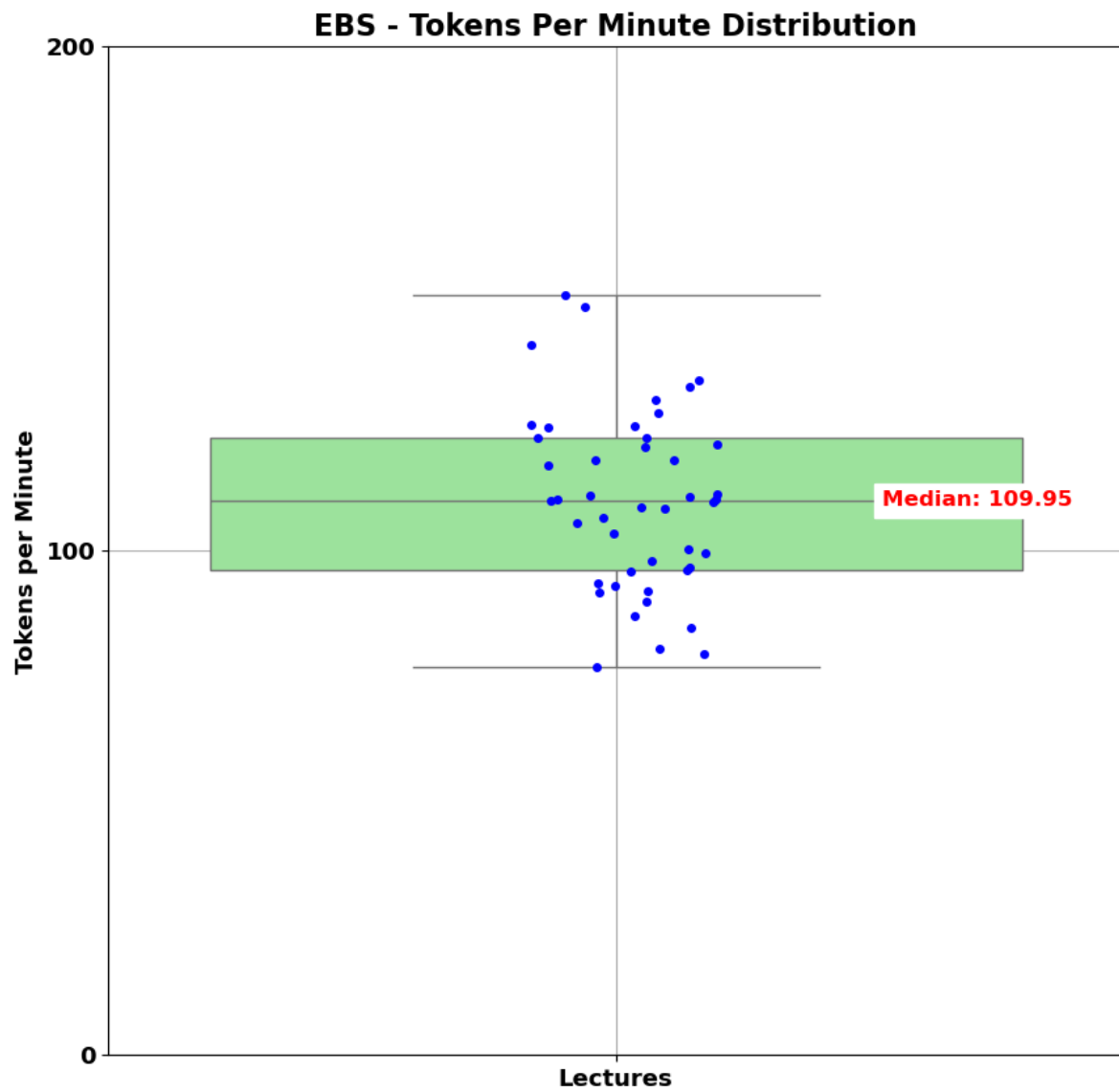
- EBS의 분당 문자 수는 200~400 사이에 분포되어 있고, 중앙값은 271.30입니다.
- KHAN의 분당 문자 수는 100~400 사이에 분포되어 있고, 중앙값은 178.34입니다.
- 한국과 미국 강의의 분당 문자 수 분포도에 따르면 EBS가 KHAN보다 더 좁게 분포된 것을 볼 수 있고, 중앙값 기준 분당 문자 수는 92.96만큼 차이가 납니다.

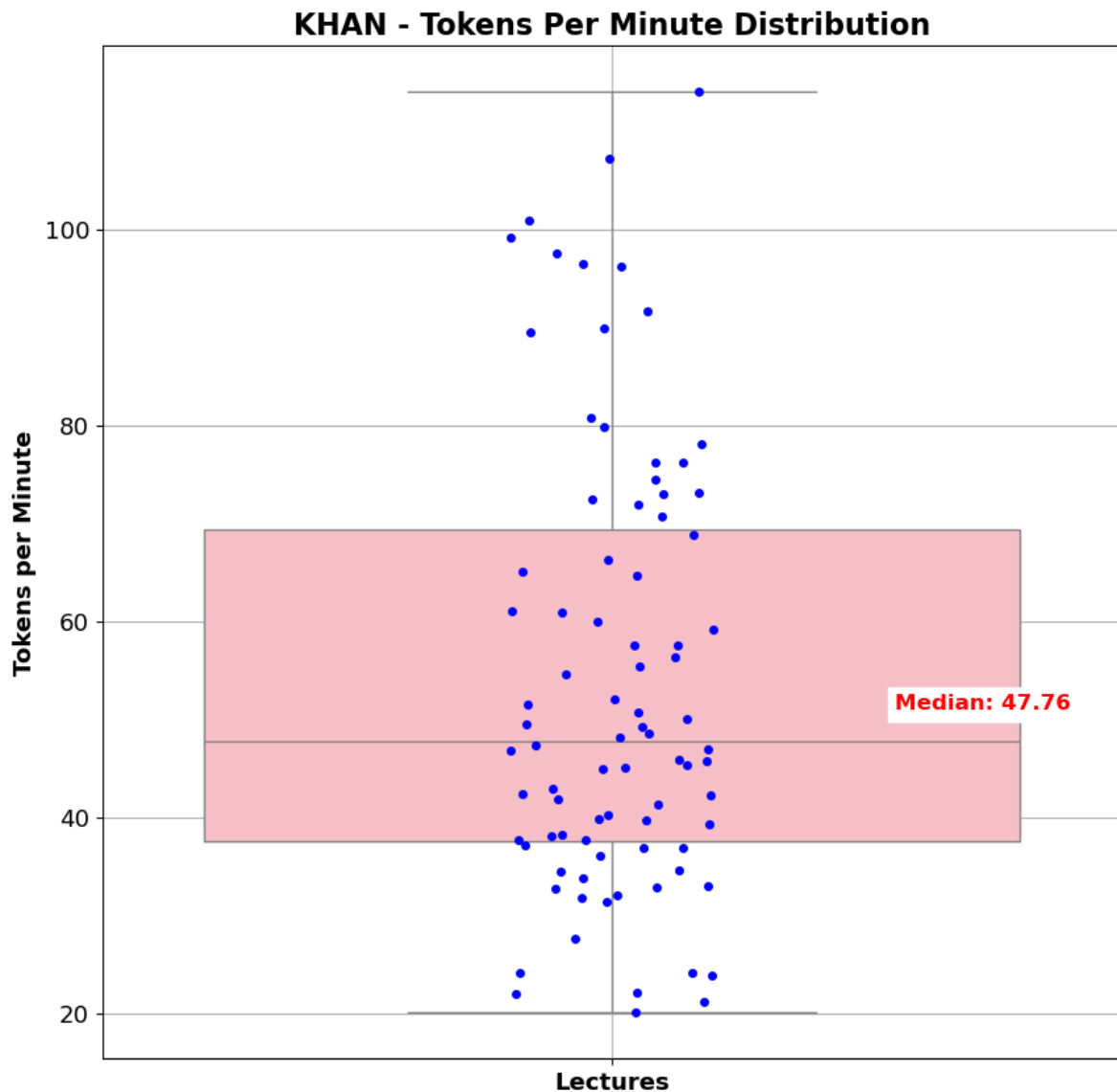
#### ▼ 2.3.4 분당 토큰 수 계산 및 비교

- 사용된 라이브러리: pydub, KoNLPy(EBS), NLTK(KHAN)
- 작업 설명: 각 강의의 오디오 길이와 해당 텍스트 파일의 총 단어 수를 이용하여 분당 토큰 수를 계산 하였습니다.

#### ▼ 2.3.5 시각화(박스플롯 및 스트립플롯 생성)

- 사용된 라이브러리: matplotlib, seaborn
- 작업 설명: seaborn을 사용하여 각 강의의 분당 토큰 수를 박스플롯과 스트립플롯으로 시각화하였습니다.





#### ▼ 2.3.6 해석

- EBS의 분당 토큰 수는 80~150 사이에 분포되어 있고, 중앙값은 109.95입니다.
- KHAN의 분당 토큰 수는 20~120 사이에 분포되어 있고, 중앙값은 47.76입니다.
- 한국과 미국 강의의 분당 토큰 수 분포도에 따르면 EBS가 KHAN보다 더 좁게 분포된 것을 볼 수 있고, 중앙값 기준 분당 토큰 수는 62.19만큼 차이가 납니다.

### 3. 비교

#### ▼ 강의 스타일

	EBS	KHAN
분당 문자수	271.30	178.34
분당 토큰(정보의 밀집)	109.95	49.94

정형성	정형(자유도가 낮음)	비정형(자유도가 높음)
시간(총 강의 길이)	30분 내외	10분 내외

⇒ 분당 토큰수가 많을 수록 정보가 밀집되어 있다고 이야기할 수 있습니다. EBS가 상대적으로 정보가 밀집되어 있으며 정형성을 갖춘 강의로 보입니다. 그에 반해 KHAN은 자유롭고 비교적 정보의 밀집도가 낮은 강의라고 할 수 있습니다.

⇒ KHAN이 EBS에 비해 학생들이 강의를 수강했을 때 이를 보상을 통해 성취감을 올려주는 시스템이 더 잘 구축되어 있습니다.

- 보상 : 수학 문제를 풀거나 강의를 볼 때마다 에너지 포인트(Energy Point)를 얻으며, 활동 정도에 따라 배지를 얻을 수 있습니다. 주로 달, 지구 등의 표현으로 우주를 지칭하는 표현으로 구성되어 있으며 얻은 배지는 본인의 등급을 나타내는 용도로 사용 가능합니다.(별도의 금전적 가치는 없음)

#### ▼ 학생의 성향

두 강의를 각각 정형성을 기준으로 EBS를 자유도가 낮은 강의, 칸을 자유도가 높은 강의라고 했을 때, 학생들의 성향에 따라 적절한 강의를 추천할 수 있습니다.

- 자유도가 낮은 강의
  - 해당 학년에서 시험 대비나 개념을 명확히 구축할 필요가 있는 학생
  - 학습 과정에서 앞선 개념에 대한 복습이 필요한 경우
- 자유도가 높은 강의
  - 심화 학습이 필요한 학생 중 이후에 배울 학년에 등장하는 개념을 새롭게 학습하는 학생
  - 한번도 배운 적 없는 이후 내용을 예습하는 학생

#### ▼ 결론

- 개념에 대해 복습을 하거나 시험에 대비하는 등 체계적인 학습이 필요한 경우 상대적으로 자유도가 낮은 강의를, 새로운 개념을 배우거나 심화학습에서 필요한 개념을 예습하는 경우, 자유도가 높은 강의를 보는 것을 추천합니다. 해당 서비스를 통해 원하는 강의 카테고리가 있다면 본인에게 필요한 스타일에 따라 강의에 자유롭게 접근할 수 있다는 점이 해당 서비스의 큰 특징 중 하나입니다.

## 데이터 전처리

### 텍스트 추출

시각화를 위한 텍스트 추출과 별개로, 데이터 라벨링을 위해서 타임스탬프가 없는 텍스트를 추출했습니다.

#### EBS

- 강의 영상 → 음성파일로 변환 : moviepy의 VideoFileClip을 이용하여 추출했습니다.

- 음성 파일 → 텍스트 추출 : openai의 whisper-1 모델 api를 사용하여 음성 파일의 내용을 텍스트 파일에 저장했습니다.
- 강의가 분할되어 있으므로, 같은 주제의 영상들은 한 파일에 합쳐서 저장했습니다.
- ebs 스크립트 개수 : 45개

## Khan Academy Kids

- 강의 영상 → 음성파일로 변환 : moviepy의 VideoFileClip을 이용하여 추출했습니다.
- 음성 파일 → 텍스트 추출 : openai의 whisper-1 모델 api를 불러오고 사용하여 음성 파일의 내용을 텍스트 파일에 저장했습니다.
- 영어 강의이므로, 음성 파일 → 텍스트 추출 과정에서 한국어 텍스트로 번역하여 추출하도록 했습니다.  
(GPT-4 API 이용)
- Khan 스크립트 개수 : 84개

## 카테고리 분류

모든 강의 영상을 보고 카테고리를 구분하기 어렵기 때문에 다음과 같은 방법으로 카테고리를 분류하였습니다.

- **상위 카테고리 설정:** 초등학교 수학 교과과정(2009년 교육교육과정)을 바탕으로 상위 카테고리(예: 수와 연산, 도형, 측정, 규칙성, 확률과 통계)를 먼저 설정했습니다..
- **텍스트 요약 및 태그 추출:** 이후 GPT-4 API를 활용하여 강의 텍스트를 요약하고, 요약된 각 텍스트 파일에서 빈도수에 따라 상위 3개의 키워드를 추출했습니다. 이 과정에서 총 30개의 세부 카테고리(태그)를 생성했습니다.
- **태그 분류:** 생성된 30개의 태그는 초등학교 교과과정 기반의 상위 카테고리에 맞추어 분류하였습니다.
- **모델 학습 및 데이터베이스 저장:** 모델 학습 시 상위 카테고리(수와연산, 도형, 측정, 규칙성, 확률과 통계)를 클래스로 활용하였고 상위 카테고리를 입력하면 관련된 영상 정보가 호출되게 하였습니다.

Category	Values and Labels
수와 연산	0(수), 1(덧셈), 2(뺄셈), 3(곱셈), 4(나눗셈), 5(분수), 6(소수), 7(소수점), 25(약수), 26(배수)
도형	15(도형), 13(평면도형), 14(입체도형), 17(원), 16(대칭)
측정	8(길이), 9(시간), 10(화폐), 11(측정), 12(각도), 18(직각)
규칙성	19(비율), 20(비례식), 21(비), 29(규칙)
확률과 통계	22(그래프), 23(통계), 24(확률)

Category	value
수	0
덧셈	1
뺄셈	2
곱셈	3

나눗셈	4
분수	5
소수	6
소수점	7
길이	8
시간	9
화폐	10
측정	11
각도	12
평면도형	13
입체도형	14
도형	15
대칭	16
원	17
직각	18
비율	19
비례식	20
비	21
그래프	22
통계	23
확률	24
약수	25
배수	26
함수	27
미지수	28
규칙	29

reference : 2009년 초등수학 교과과정

## 초기 데이터셋 구성

영상을 카테고리에 따라 분류하였고 source라는 컬럼을 추가해 어떤 플랫폼의 영상인지를 저장한 데이터셋을 최종적으로 구성했습니다. 추후 영상이 계속 추가되는 것을 염두에 두고 초기 데이터셋이라 언급했습니다.

	text	category	source
	이 텍스트를 한국어로 번역: \n\nGavin은 이런 말들이 있습니다. 아래...	수와 연산	khan
	어떤 숫자들이 6보다 큰가요? 모두 선택해 보세요. 그래서 여기 수직선에서 6을 볼...	수와 연산	khan
	우리는 별, 숫자, 또는 글자를 가장 많이 세어 봅시다. 이것을 한 번 보겠습니다....	수와 연산	khan
	이 사진에 얼마나 많은 사람들이 보이나요? 보세요, 저는 하나, 둘, 셋, 넷, 다...	수와 연산	khan
	어떤 상자에 12마리의 고래가 있는지요? 그래, 이런 이 녹색 상자를 보세요, 한번...	수와 연산	khan

# 서비스

## 이용자

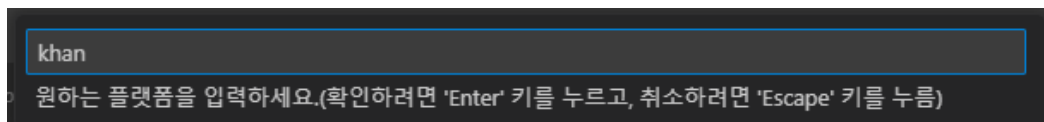
### DB 구축

sqlite3를 이용하여 video\_edu.db라는 DB를 생성했습니다. math\_elementary\_ 라는 테이블을 생성하여 초기 데이터셋을 저장하였습니다

### 활용

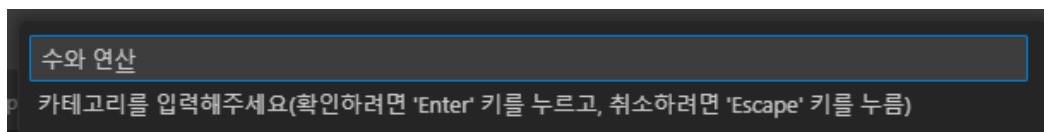
- 현재 UI없이 input을 통하여 원하는 플랫폼과 카테고리를 입력하여 호출하는 식으로 구현하였습니다.

- 원하는 플랫폼 입력



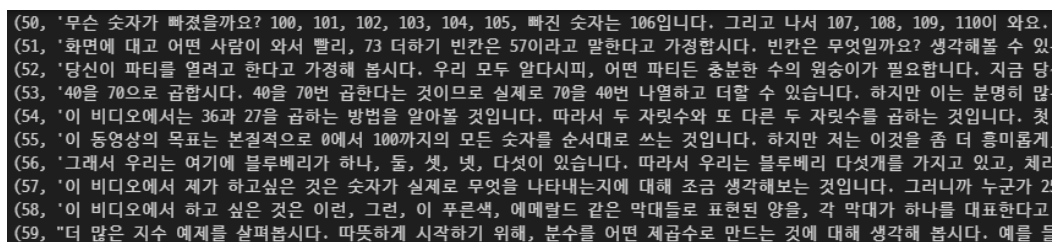
A terminal window with a dark background. A light blue rectangular input field contains the text 'khan'. Below the input field, a line of text reads: '원하는 플랫폼을 입력하세요.(확인하려면 'Enter' 키를 누르고, 취소하려면 'Escape' 키를 누름)'

- 필요한 카테고리 입력



A terminal window with a dark background. A light blue rectangular input field contains the text '수와 연산'. Below the input field, a line of text reads: '카테고리를 입력해주세요.(확인하려면 'Enter' 키를 누르고, 취소하려면 'Escape' 키를 누름)'

- 인덱스, 강의, 카테고리 출력(+추후 영상과 연결, 위 작업 실행 시 영상이 출력되도록 개선)



A terminal window with a dark background displaying a list of math problems in Korean. The text is as follows:  
(50, '무슨 숫자가 빠졌을까요? 100, 101, 102, 103, 104, 105, 빠진 숫자는 106입니다. 그리고 나서 107, 108, 109, 110이 와요.  
(51, '화면에 대고 어떤 사람이 와서 빨리, 73 더하기 빈칸은 57이라고 말한다고 가정합니다. 빈칸은 무엇일까요? 생각해볼 수 있  
(52, '당신이 파티를 열려고 한다고 가정해 봅시다. 우리 모두 알다시피, 어떤 파티든 충분한 수의 원숭이가 필요합니다. 지금 당  
(53, '40을 70으로 곱합니다. 40을 70번 곱한다는 것이므로 실제로 70을 40번 나열하고 더할 수 있습니다. 하지만 이는 분명히 많  
(54, '이 비디오에서는 36과 27을 곱하는 방법을 알아볼 것입니다. 따라서 두 자릿수와 또 다른 두 자릿수를 곱하는 것입니다. 첫  
(55, '이 동영상의 목표는 본질적으로 0에서 100까지의 모든 숫자를 순서대로 쓰는 것입니다. 하지만 저는 이것을 좀 더 흥미롭게  
(56, '그래서 우리는 여기에 블루베리가 하나, 둘, 셋, 넷, 다섯이 있습니다. 따라서 우리는 블루베리 다섯개를 가지고 있고, 체리  
(57, '이 비디오에서 제가 하고 싶은 것은 숫자가 실제로 무엇을 나타내는지에 대해 조금 생각해 보는 것입니다. 그러니까 누군가 2  
(58, '이 비디오에서 하고 싶은 것은 이런, 그런, 이 푸른색, 에메랄드 같은 막대들로 표현된 양을, 각 막대가 하나를 대표한다고  
(59, '더 많은 지수 예제를 살펴봅시다. 따뜻하게 시작하기 위해, 분수를 어떤 제곱수로 만드는 것에 대해 생각해 봅시다. 예를 들

## 관리자

새로운 영상을 추가했을 때 음성을 텍스트로 추출하고 한글로 번역하는 작업을 거칩.

이후 딥러닝 모델을 이용하여 5개의 카테고리에 맞게 분류하여 DB에 추가하는 시스템 구축.

## 영상 데이터 처리

'수와 연산', '도형', '측정', '규칙성', '확률과 통계'

5개의 클래스로 분류한 KHAN 강의와 EBS 강의 데이터로 새로운 강의의 클래스를 분류하는 딥러닝 모델을 구현했습니다.

feature : 강의 스크립트

target : 카테고리

먼저 KHAN, EBS 강의의 퓨어 텍스트 데이터를 tf-idf 벡터라이징을 통해 텍스트 데이터를 수치화하여 모델이 이해할 수 있는 형태로 만들었습니다.

(TfidfVectorizer - Max\_features : 10000)

그리고, 클래스를 One-Hot Encoder로 인코딩하여 모델이 학습하기에 용이하도록 했습니다.

학습 데이터와 테스트 데이터를 8 : 2 비율로 분리했습니다.

클래스 비율을 원본 데이터와 동일하게 유지하기 위해서 stratify 파라미터를 이용했습니다.

(랜덤 변수 : 4)

	규칙성	도형	수와 연산	측정	확률과 통계
학습	22	46	193	30	21
테스트	6	11	49	7	5

## 모델 구현

### 모델

tensorflow 기본 Sequential

	비고
Dense	512개 노드 활성 함수 : relu Input_shape : 10000
Dropout	0.5 비율
Dense	256개 노드 활성 함수 : relu
Dropout	0.5 비율
Dense	Output_layer : 5(클래스 개수) 활성 함수 : softmax
Compile	optimizer : adam loss : categorical_crossentropy metrics : accuracy

### 기타 사항

Total params: 5253125 (20.04 MB)

Trainable params: 5253125 (20.04 MB)



Non-trainable params: 0 (0.00 Byte)

epoch : 9

배치 크기 : 32

학습률 : 0.001 (기본값)

검증 데이터셋 : 학습 데이터 비율의 20% 랜덤 지정

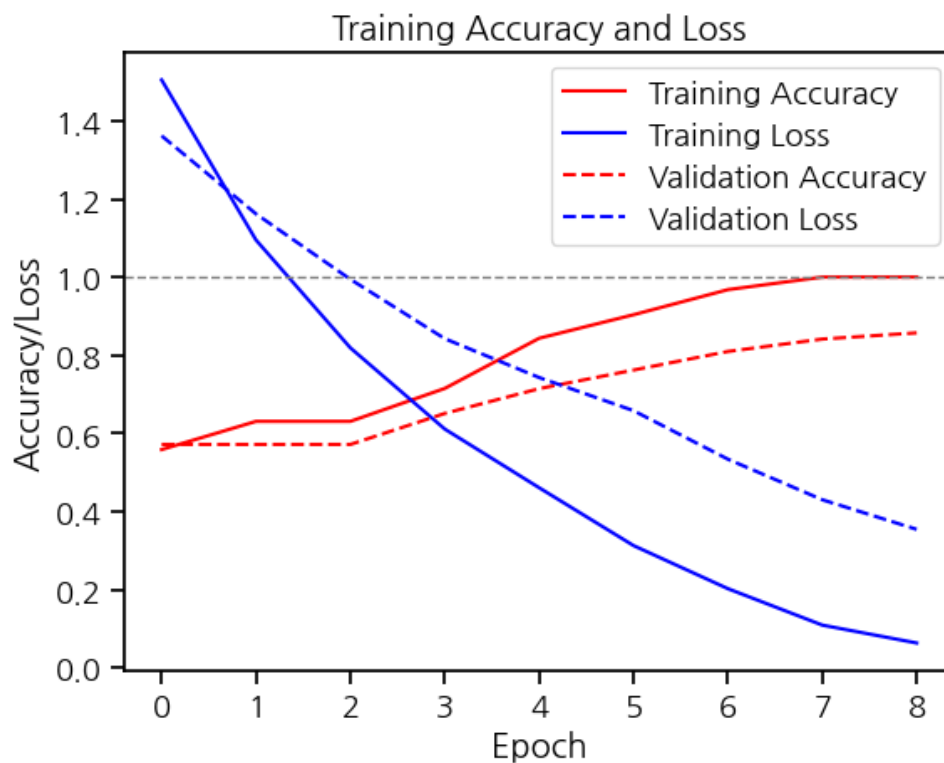
## 성능평가

학습 완료된 모델의 학습, 검증 데이터의 정확도와 loss를 계산했습니다.

학습/검증 데이터 정확도 : 1.0000 / 0.8571

학습/검증 데이터 loss : 0.0629 / 0.3539

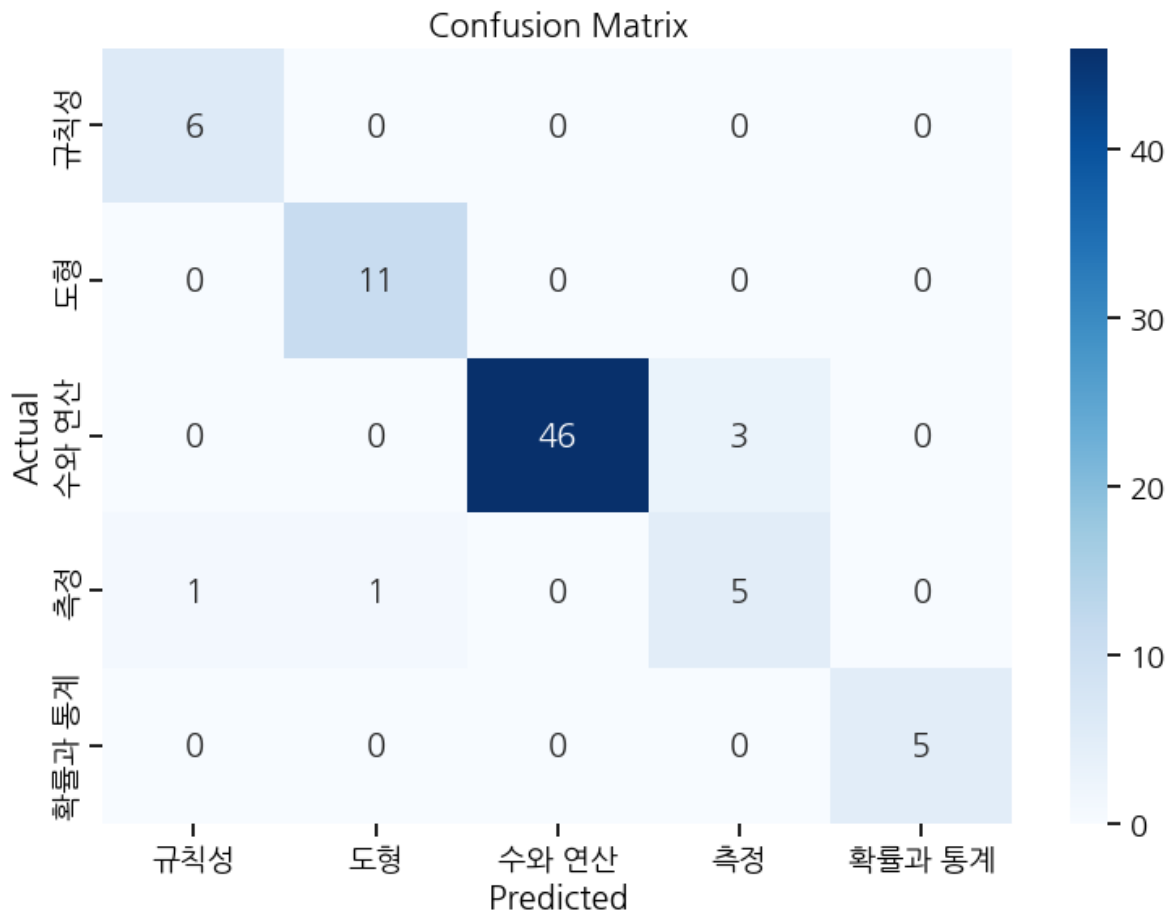
학습과 검증 데이터의 정확도와 loss 추이를 시각화해 보았습니다.



그래프를 보아, 학습이 진행될수록 loss는 크게 떨어지고 정확도는 점점 올라 마지막 2회의 에포크에서는 100%를 달성했고, 검증 데이터 정확도에서도 점점 정확도가 오르고 loss가 떨어지는 모습을 보여 학습은 잘 되었다고 판단했습니다.

그래서 모델의 성능을 테스트해보기 위해서 테스트 샘플에 대한 정확도를 계산하고, 혼동행렬도 시각화해 보았습니다.

테스트 데이터 정확도 : 0.9359



전체적으로 테스트 데이터를 분류를 잘 하는 모습을 보여 모델의 성능이 좋은 것으로 보입니다. 하지만 '측정' 데이터를 상대적으로 잘 분류하지 못하고, '수와 연산' 데이터에 편향되어 있어서 모델이 대부분을 '수와 연산'으로 분류해서 정확도가 높게 나왔을 가능성이 있습니다.

## DB 추가

- 영상을 폴더에 넣으면 전처리 과정과 모델을 통한 분류 작업을 통해 DB에 저장되도록 순차적인 코드를 작성하였습니다.(동봉된 파일 중 model\_db라는 파일을 통해 구현하였습니다.)
- 모델 분류 작업 이후 텍스트(text)와 예측된 분류 카테고리(predicted\_labels[i])를 변수로 저장하여 DB에 insert문을 통해 저장됩니다.

	text	category	source
0	이 텍스트를 한국어로 번역: \r\n\r\nGavin은 이런 말들이 있습니다. 아래...	수와 연산	khan
1	어떤 숫자들이 6보다 큰가요? 모두 선택해 보세요, 그래서 여기 수직선에서 6을 볼...	수와 연산	khan
2	우리는 별, 숫자, 또는 글자를 가장 많이 세어 봅시다. 이것을 한 번 보겠습니다....	수와 연산	khan
3	이 사진에 얼마나 많은 사람들이 보이나요? 보세요, 저는 하나, 둘, 셋, 넷, 다...	수와 연산	khan
4	어떤 상자에 12마리의 고래가 있는지요? 그래, 이런 이 녹색 상자를 보세요, 한번...	수와 연산	khan
...	...	...	...
126	2를 곱했잖아요. 4 2 8 한거죠. 4 곱하기 2를 한 것입니다. 그래서 이것이 ...	수와 연산	ebs
127	여러분 안녕하세요 반갑습니다. 김연경 선생님이예요. 자 우리 친구들 오늘이요 굉장...	확률과 통계	ebs
128	7종류의 퀴즈 힌트를 찾아가보시죠 可以 하apter 요 sign 여 여 여 만남...	수와 연산	ebs
129	오르노르 오르노르 오르노르	오르노르 오르노르 오르노르	오르노르 오르노르 오르노르
130	새로운 텍스트	새로운 카테고리	새로운 플랫폼

- 새로운 영상 정보가 추가되는 쿼리 구현

## 결론

- 이용자
  - 한국의 교육과정에 맞게 칸 아카데미의 영상을 저장하여 한국 학생이 본인에게 맞는 플랫폼을 체험하기 용이하게 서비스를 구축해보았습니다. 또한 수학 교육과정에 맞게 영상들을 분류하여 넣었기에 한국 학생이 본인이 필요한 영상을 쉽게 찾아볼 수 있습니다.
  - 본인의 학습 상황이나 성취 정도에 따라 스타일이 상반되는 강의를 선택하여 시청할 수 있습니다. 한국 수학 커리큘럼에 맞게 영상을 분류했지만 그 전달 방식에 있어서 한국과 미국의 강의는 차이를 보인다는 것을 볼 수 있었습니다. 그런 차이를 학습 상황에 맞게 이용할 수 있습니다.
- 관리자
 

영상만 가지고 있다면 이를 분류하여 DB에 저장하는 것이 용이합니다. 대량의 영상을 처리할 때 영상을 다 시청하지 않고도 적절하게 분류하고 저장하여 이용자가 이용할 수 있습니다.

## 한계

현재는 강의 텍스트와 카테고리, 플랫폼으로만 DB를 구성했기 때문에 실질적인 영상이랑 연결되어 있는 상태는 아닙니다.

### ▼ 개선

차후 개선 작업을 통해 영상을 연결하여 키워드를 입력하면 영상이 바로 호출될 수 있게 개선해보고자 합니다.

가지고 있는 데이터가 대부분 초등 과정이기에 대부분 수와 연산과 관련된 데이터가 많았고 그래서 모델이 수와 연산을 제외한 다른 카테고리를 분류해내는 능력이 떨어집니다.

### ▼ 개선

영상에서 추출한 텍스트 데이터를 전처리하는 과정을 개선하거나 초등 고학년 과정이나 수와 연산을 제외한 카테고리의 영상을 추가하는 방식을 통해 개선하고자 합니다.

현재는 교육 과정을 크게 5가지 카테고리로 나누어 분류해두었습니다. 모델 훈련 과정에서 기존에 나누었던 30가지 카테고리로 분류하기에는 어려움이 있어 이렇게 설정을 했었습니다. 즉 더욱 상세한 카테고리를 통해 조회하기에는 어려움이 있습니다.

#### ▼ 개선

30가지의 카테고리로 분류할 수 있게 모델을 다시 구성하거나 데이터를 추가하여 훈련시키는 방법을 통해 개선하고자 합니다.

## 개발 환경

---

- Vscode 1.92.2
- Anaconda 1.12.3
  - Python 3.12.4
  - Numpy 1.26.4
  - Pandas 2.2.2
  - matplotlib 3.8.4
  - Scikit\_learn 1.4.2
  - Seaborn 0.13.2
  - jupyter notebook v2024.7.0
  - moviepy 1.0.3
  - openai 0.28.0
  - tensorflow 2.13.0
  - torch 2.4.1
  - transformers 4.44.2
  - speech\_recognition 3.10.0
  - pydub 0.25.1
  - librosa 0.10.2.post1
  - KoNLPy 0.6.0
  - NLTK 3.8.1

## 출처

---

강의 영상: EBS 유튜브 재생목록 '초등수학 개념잡기'

[https://youtube.com/playlist?  
list=PLQvhoNACU3iEcwBgT3WIWXHmIS8y2Bpl2&si=xT1LXa7lZqPboZrZ](https://youtube.com/playlist?list=PLQvhoNACU3iEcwBgT3WIWXHmIS8y2Bpl2&si=xT1LXa7lZqPboZrZ)

Khan academy

<https://ko.khanacademy.org/>

참고

국제학업성취도평가(PISA) 2022

## 소스코드

model\_db 파일과 같은 폴더 안에 영상을 넣고, model\_db 파일의 첫 번째 셀에 video\_path에 영상의 상대 경로를 설정해 주고, 실행시켜 주면 됩니다. api키는 천재교육에서 제공해주신 api키를 활용했습니다. 이를 넣어야 원활한 실행이 가능합니다.