제작 보고서

과정명, 팀명 : 빅데이터 7기 '퀘스트지니'팀

작성 일자	2024년 07월 25일	**제작 기간**	2024년 06월 14일 ~ 2024년 07월 29일
팀원	장하나, 민선영, 이민아, 최종은	**참가 주제**	문항 추천 시스템 제작
개요	1. 요약 2. R&R 3. WBS (Work Breakdown Structure) 4. 서론 5. 개발 환경 6. 기능 구현 7. 산출물 8. 개선점 9. 소스코드 10. 참고문헌		
요약	현재 공교육에서의 AI디지털교과서의 도입은 교육계의 큰 이슈로 떠오르고 있다. 고등학교의 경우 킬러 문항을 제거한다는 정부의 기조로 인해 출제 경향 파악 및 취약 문항 학습의 중요성이 높아지고 있으나 이를 대비하기 위해 원하는 문항만을 선별하기 위해서는 교강사가 수작업으로 교육 콘텐츠를 제작해야 하기에 인력 부족 및 시간이 많이 소요된다는 한계점이 있다. 이를 해결하기 위해 고등학교 교강사를 대상으로 주요 교과인 국어, 영어, 수학 교과에 대해 교강사가 학생들에게 제공하고자 하는 문항과 유사한 문항 추천, 나아가 유사 문항 생성에 이르는 서비스의 제공이 필요하다고 보았다. 따라서, 본		

	프로젝트는 천재교육 서비스를 사용하는 교강사를 대상으로 하는 문항 추천 시스템을 개발하고자 한다. 문항 추천 시스템 개발을 위해 고등학교 1학년과 2학년은 2015년도 3월부터 2024년도 6월까지 10개년의 문항을 수집했으며, 고등학교 3학년은 2019년도 3월부터 2024년도 6월까지 5개년의 문항을 수집하여 최종적으로 국어 5333문항, 수학 4821문항, 영어 4952문항을 데이터로 사용하였다. 정확한 문항 이미지를 추출하기 위해 이미지 crop시에 컨투어(contour)[1] 영역을 이용한 crop방식과 Easyocr을 활용한 OCR[2]방식을 사용하였으며, 텍스트 파싱을 위해 국어, 영어 교과에서는 Tesseract를 사용한 방식을, 수학에서는 OpenAI의 gpt4o모델을 사용한 프롬프트 엔지니어링을 사용하였다. 문항 추천 시스템의 주요 기능은 크게 2가지로 나뉜다. 새로운 데이터 발생 시 기존 데이터를 업데이트 해주는 데이터 버전 관리 기능과 사용자가 실제로 문제를 웹에 넣어 유사 문항을 추천해주는 문항 추천 기능으로 이루어져 있다. 데이터 버전 관리 기능의 경우 전처리 과정을 거친 데이터를 S3에 저장하고 텍스트 데이터는 S3에서 MariaDB로 적재되게 한다. 그 후 이미지와 텍스트 데이터를 병합한 데이터를 임베딩[3] 모델을 통해 임베딩 벡터값을 산출한 후 ElasticSearch에 저장한다. 그 후 ElasticSearch에서 코사인 유사도를 실시하고 그 값을 MariaDB에 적재한다. 사용자가 실제 이용하는 기능의 경우 웹에 들어온 이미지를 전처리하여 S3에 저장하고, 저장된 이미지와 텍스트를 대상으로 임베딩과 코사인 유사도를

[1] 컨투어(contour): 이미지에서 동일한 강도의 픽셀들을 연결한 곡선. 즉, 이미지 내에서 특정 객체의 외곽선.

[2] OCR(Optical Character Recognition, 광학 문자 인식): 이미지나 스캔한 문서에 포함된 텍스트를 인식하고 디지털 텍스트로 변환하는 기술

[3] 임베딩(embedding): 자연어 처리(NLP)와 기계 학습에서 단어, 문장, 문서 등의 텍스트 데이터를 고차원 벡터 공간으로 변환하는 기술

실시한다. 이 벡터값들을 Elasticsearch에 저장하여 기존 데이터셋과의 유사도를 비교하여 유사 문항을 추천해준다. 상위 5개의 유사 문항 추천 시 평균 유사돈 80~90% 사이로 높은 유사도와 일치도를 나타내고 있다. 본 프로젝트의 개선점은 다음과 같다. 첫째, 시간 및 인력 상의 어려움으로 인해 웹 상에서 사용자가 결과에 만족하지 못할 경우 다른 문제들도 보이게 하거나 자동 채점 기능을 구현하지 못한 부분이 개선점이라고 여겨진다. 둘째, 문항 분류 결과 기반으로 유사문제 추천/생성을 기존에 개발 목표로 잡았으나 생성의 경우 시간 및 인력 부족으로 인해 자동태깅 및 유사도를 정교화하는 방향으로 가게 되었다. 특히, AWS SageMaker를 통해 배포 및 관리할 수 있는 모델 중 교육 특화 및 연구 목적에 적합한 LLama 모델을 사용한다면 문항 생성도 가능할 것이라 보여진다.

* 위 프로젝트 진행과정에서 소요된 비용 산정은 별첨자료 참고.

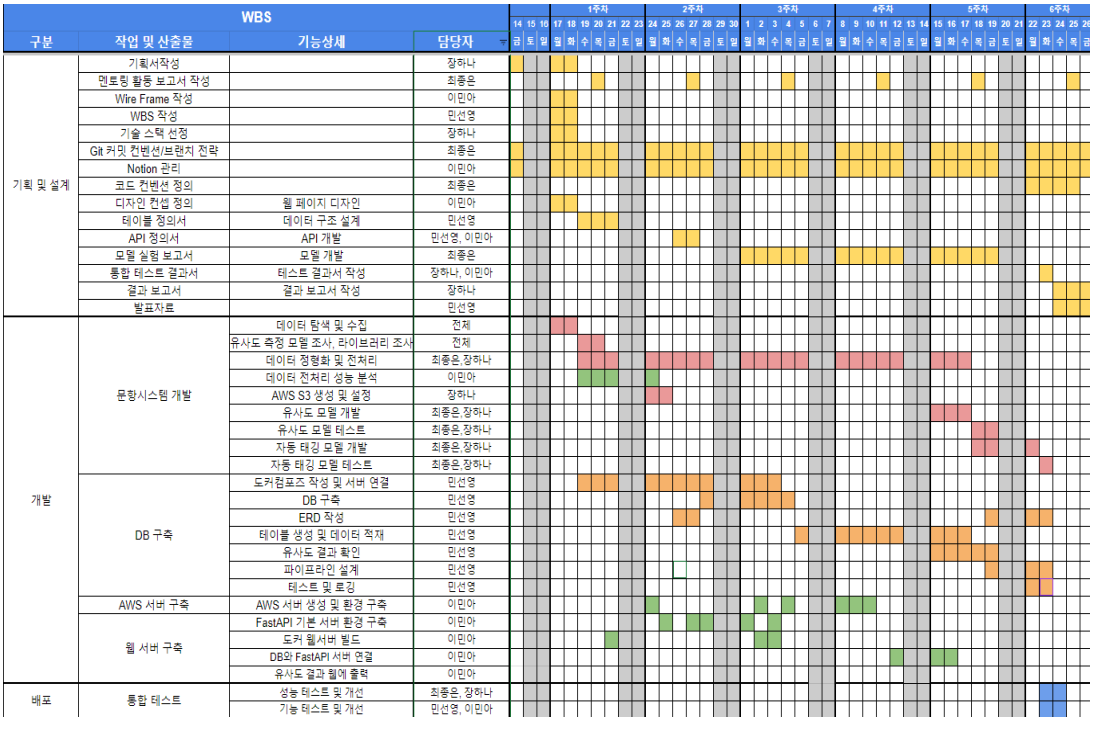
본문

2. R&R

담당자	업무
장하나	- 팀장 - 데이터 정량화 및 전처리 - 문항시스템 개발 및 테스트 - 자동 태깅 모델 개발 및 테스트 - AWS S3 생성 및 설정, 관리 - 기획서 및 최종 보고서 작성
민선영	- 데이터 수집 - DB 구축 및 관리 - 도커컴포즈 작성 및 서버 연결 - 파이프라인 설계 - PPT 제작 - 테이블 정의서 작성
이민아	- 데이터 수집

	- AWS 서버 생성 및 환경 구축 - FastAPI 기본 서버 환경 구축 - 도커 웹서버 빌드 - DB와 웹 서버 연결 및 관리 - Wire Frame 작성 - 노션 관리
최종은	- 데이터 수집 - 데이터 정량화 및 전처리 - 문항시스템 개발 및 테스트 - 자동 태깅 모델 개발 및 테스트 - 멘토링 활동 보고서 작성 - Git 커밋 컨벤션/브렌치 전략 - 코드 컨벤션 정의

3. WBS (Work Breakdown Structure)



WBS				1주차							2주차							3주차							4주차							5주차							6주차				
구분	작업 및 산출물	기능상세	담당자	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
				금	토	일	월	화	수	목	금	토	일	월	화	수	목	금	토	일	월	화	수	목	금	토	일	월	화	수	목	금	토	일	월	화	수	목	금	토	일	월	화	수	목	금
기획 및 설계	기획서작성		장하나																																									
	멘토링 활동 보고서 작성		최종은																																									
	Wire Frame 작성		이민아																																									
	WBS 작성		민선영																																									
	기술 스택 선정		장하나																																									
	Git 커밋 컨벤션/브랜치 전략		최종은																																									
	Notion 관리		이민아																																									
	코드 컨벤션 정의		최종은																																									
	디자인 컨셉 정의	웹 페이지 디자인	이민아																																									
	테이블 정의서	데이터 구조 설계	민선영																																									
	API 정의서	API 개발	민선영, 이민아																																									
	모델 실험 보고서	모델 개발	최종은																																									
	통합 테스트 결과서	테스트 결과서 작성	장하나, 이민아																																									
	결과 보고서	결과 보고서 작성	장하나																																									
	발표자료		민선영																																									
개발	문항시스템 개발	데이터 탐색 및 수집	전체																																									
		유사도 측정 모델 조사, 라이브러리 조사	전체																																									
		데이터 정형화 및 전처리	최종은,장하나																																									
		데이터 전처리 성능 분석	이민아																																									
		AWS S3 생성 및 설정	장하나																																									
		유사도 모델 개발	최종은,장하나																																									
		유사도 모델 테스트	최종은,장하나																																									
		자동 태깅 모델 개발	최종은,장하나																																									
		자동 태깅 모델 테스트	최종은,장하나																																									
	DB 구축	도커컴포즈 작성 및 서버 연결	민선영																																									
		DB 구축	민선영																																									
		ERD 작성	민선영																																									
		테이블 생성 및 데이터 적재	민선영																																									
		유사도 결과 확인	민선영																																									
		파이프라인 설계	민선영																																									
		테스트 및 로깅	민선영																																									
	AWS 서버 구축	AWS 서버 생성 및 환경 구축	이민아																																									
	웹 서버 구축	FastAPI 기본 서버 환경 구축	이민아																																									
		도커 웹서버 빌드	이민아																																									
		DB와 FastAPI 서버 연결	이민아																																									
		유사도 결과 웹에 출력	이민아																																									
배포	통합 테스트	성능 테스트 및 개선	최종은, 장하나																																									
		기능 테스트 및 개선	민선영, 이민아																																									

4. 서론

현재 공교육에서의 AI디지털교과서의 도입은 교육계의 큰 이슈로 떠오르고 있다. AI 교과서 도입을 앞두고 교육부는 2024년 관련 예산을 5333억원 편성했으며, AI 교과서는 연 구독료 기준 6만~10만원 선으로 예상되어 조 단위 교과서 시장이 창출될 것이라는 전망이 나오고 있다.[i] AI 교과서 확산에 따라 개별화 및 맞춤형 교육에 대한 기대도 커지고 있으며 이러한 기대를 충족하기 위한 AI 기술을 활용한 교육 콘텐츠가 요구되고 있다. 이는 입시에 높은 중요성을 갖고 있는 고등학교도 다르지 않다. 고등학교의 경우 입시와 직결되어 있는 시기인 만큼 각 학생마다 다양한 전략을 사용한 학습이 중요하다. 최근 킬러 문항을 제거한다는 정부의 기조로 인해 핵심적이고 기본적인 개념에 대한 이해가 중요성을 갖게 되며 출제 경향 파악 및 취약 문항 학습의 중요성이 높아지고 있으나[ii] 고등학교 교강사가 학생별 맞춤형 문항을 제공하기는 어려운 현실이다. 이를 대비하기 위해서는 교강사가 수작업으로 맞춤형 문항을 제공하기 위해 문항 판단 및 편집, 제작을 해야 하기에 인력 부족 및 시간이 많이 소요된다는 문제점이 있다고 보아 경쟁사 및 자사 분석을 진행하였다.

문항 추천 및 관리와 관련된 경쟁사의 서비스는 다음과 같다. 우선, 비상교육의 기출탭탭은 2015 개정 교육 과정이 모두 수록돼 있는 태블릿PC 기반의 수능 기출 학습 애플리케이션이다. 기출탭탭은 한국교육과정평가원이 출제한 최근 3개년 6, 9월 모의평가, 수능 전 영역 기출문제를 제공하며 취약문제 반복 풀기 및 유사 문항 제공, 유형별 기출문제 학습의 기능을 서비스하고 있다.[iii] 프리윌린은 교사를 위한 수학 문제은행 솔루션 '매쓰플랫'을 제공하고 있다. 매쓰플랫은 자체 개발한 70만개의 수학 문제를 교과서 및 시중 교재와 연동해 교사 주도의 교육이 가능하도록 도우며 수업 전에는 원하는 난이도와 범위의 학생 수준별 자료를 제공할 수 있도록 돕고 있다.[iv] 또한, B2G버전으로 수업 보조 도구로서 대시보드, 학습 관리 등에 최적화된 '스쿨 플랫'을 서비스하고 있다.[v]

문항 생성 분야의 기술 현황을 살펴보면 다음과 같다. '젠큐'의 경우 초등학교부터 고등학교 수준까지 원하는 난이도에 맞춰 국어, 영어 지문과 문제를 생성할 수 있으며 교재에서 시험 문제를 추출하고 지문과 문제 난이도를 변경하거나 형태가 유사한 문제도 쉽게 만들 수 있도록 서비스하고 있다.[vi] 비상교육은 교수 지원 플랫폼 '비바샘'에서 초등 교사를 위한 AI 기반 맞춤 수학 문항 자동 생성 서비스 '쌤핏수학'을 서비스하고 있다. 쌤핏수학은 AI 기반 수학 문항 자동 생성 엔진을 통해 학습자 맞춤형 수학 학습지를 생성할 수 있으며, 초등학교 교과서 일부 도형 및 통계 단원을 제외한 전 단원에 대해 문항 생성이 가능하다.[vii] 해외 서비스인 'QueaionWell'의 경우 지문을 프로그램에 제공하고 언어 및 읽기 수준을 설정하면 AI가 자동으로 그에 적합한 문항을 생성해주는 서비스를 제공하고 있다.[viii]

문항 추천 및 관리와 관련된 자사의 서비스는 다음과 같다. 천재교과서에서 서비스하고 있는 지니아튜터의 경우 초등학교, 중학교 학생들의 학습을 돕기 위한 자동채점, AI유사학습의 서비스가 제공되고 있다. 교육현장에서 학습 결손들이 누적되어 생기는 '학습 부진'을 방지하기 위해 형성평가를 차시 단위로 제공하고 그에 맞는 진단, 분석, 처방을 통해 완전학습을 지원[ix]하고 있다. 닥터매쓰의 경우 고등학생의 학습지원을 위해 AI '통합문항플랫폼'을 이용해 유형별, 난이도별 맞춤 문항과 쎈, 체크체크 등 시중교재의 유사문제를 제공한다. 뿐만 아니라 학습 과정에서 모르는 문제가 발생하는 경우 AI스마트렌즈 기능을 이용하여 한 번의 터치만으로 그와 유사한 문제를 제공받을 수 있어 스스로의 학습 상태를 점검할 수 있도록 하는 기능[x]을 갖추고 있다. 수능 및 모의고사, 내신 대비를 위한 다양한 문제를 수록하고 있으나 수학 과목에 한정되어 서비스가 이루어지고 있다. 이와 같은 분석에 따라 천재교육 서비스를 이용하는 교강사를 대상으로 하여, 고등학생의 개별화 및 맞춤형 교육을 지원하기 위해서 주요 교과인 국어, 영어, 수학 교과를 중심으로 교강사가 학생들에게 제공하고자

하는 문항과 유사한 문항 추천, 나아가 유사 문항 생성에 이르는 서비스의 제공이 필요하다고 보았다.

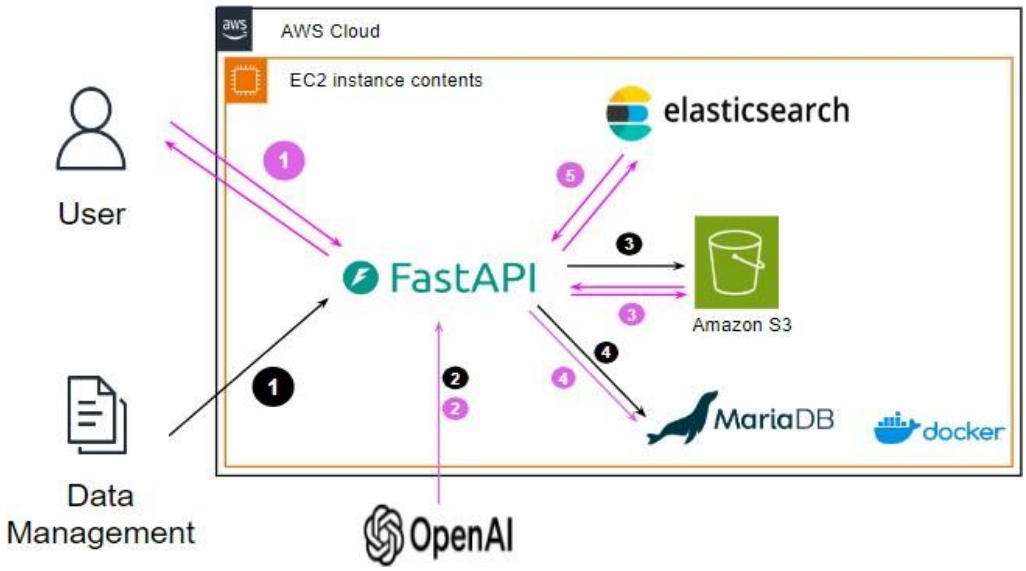
5. 개발 환경

분류	환경
운영체제	Windows 10, 11 / Ubuntu 22.04 LTS(AWS 상에서 사용)
버전 관리 시스템	Git, Github
개발 언어 및 프레임워크	Python 3.10.12 - pdf2image / opencv-python / pandas / numpy / pillow / FastAPI / Easyocr / pytesseract / pymupdf /pdf2image / glob2 / pymysql / uvlcorn / boto3 / keras / Tensorflow / torch / torchvision / transformers / sentence-transformers / nltk / python-bidi / OpenAI API
데이터베이스	AWS Maria DB, ElasticSearch
웹 서버 환경	AWS EC2

6. 기능 구현

문항 추천 시스템 개발은 다음과 같은 기능을 가지고 있다. 사용자가 외부 문제를 PNG와 같은 이미지 형태로 웹에 넣으면 웹에서는 이를 받아 필요시 이미지 crop과정을 진행한 후 텍스트 파싱 및 임베딩, 유사도 값 추출의 과정을 거친다. 그 후 기존 ElasticSearch에 저장된 임베딩, 유사도 값과 비교하여 사용자가 넣은 문항과 유사한 문항을 상위 20개 추천해주는 시스템으로 이루어져 있으며 추후 llm을 사용한 문항 생성의 과정까지 실형하기 위해 문항 자동 태깅 기능을 가지고 있다. 또한, 수능과 모의고사 문제지를 데이터셋으로 가지고 있기 때문에 수능 및 모의고사가 시행될 때마다 최신 문항을 업로드하기

위한 버전관리 기능도 구현되어 있다. 구체적인 기능 구현 과정은 다음과 같다.



(a) 기능 구현 아키텍처

구체적인 기능 구현 과정은 다음과 같다.

(0) 데이터 수집

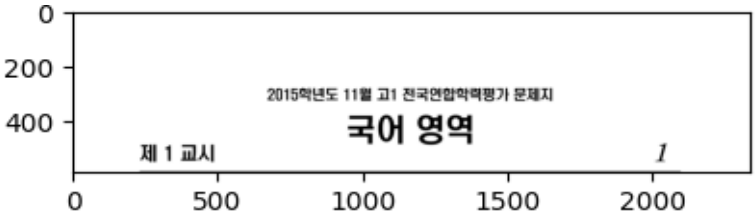
EBSi 홈페이지를 통해 국어, 영어, 수학 과목의 수능 및 모의고사 문항을 수집했다. 고등학교 1학년과 2학년은 2015년도 3월부터 2024년도 6월까지 10개년의 문항을 수집했으며, 고등학교 3학년은 2019년도 3월부터 2024년도 6월까지 5개년의 문항을 수집했다. 고등학교 1학년과 2학년의 경우 수능을 보지 않고 3월, 6월, 9월, 11월로 총 4번 모의고사를 치루는 반면, 고등학교 3학년은 수능을 포함하여 3월, 4월, 6월, 7월, 9월, 10월로 총 7번 모의고사 및 수능을 치룬다. 따라서, 학교급 별 데이터 수를 맞추기 위해 고등학교 1학년과 2학년은 2015년도부터 2023년까지 4번의 모의고사와 2024년도에 실시된 3월과 6월의 모의고사 2번을 포함하여 총 38번 실시된 모의고사의 데이터를 수집하였고, 고등학교 3학년은 2019년도부터 2023년도까지 7번의 모의고사 및 수능과 2024년도에 실시된 3월, 4월, 6월의 모의고사 3번을 포함하여 총 38번 실시된 모의고사 및 수능 데이터를 수집하였다. 이 과정에서 홀수형, 짝수형 문항의 경우 동일한 문항의 순서를 바꾼 것이므로 홀수형 데이터만 수집하였다. 반면, 가형과

나형의 경우 서로 다른 문항이 데이터로 들어 있기에 모두 데이터로 활용하였다. 따라서, 중복 문항을 제외하고 국어 5394문항, 수학 4824문항, 영어 5130문항를 이미지 데이터로 수집하였으며, pdf 및 이미지 파싱 과정에서 문항이 발생하여 사용하지 못하게 된 데이터를 제외하고 최종적으로 국어 5333문항, 수학 4821문항, 영어 4952문항을 데이터로 사용하였다. 데이터의 형태는 문항 및 해설의 경우 pdf로 저장하여 활용하였고, 답안은 png 형태로 저장하여 활용하였다.

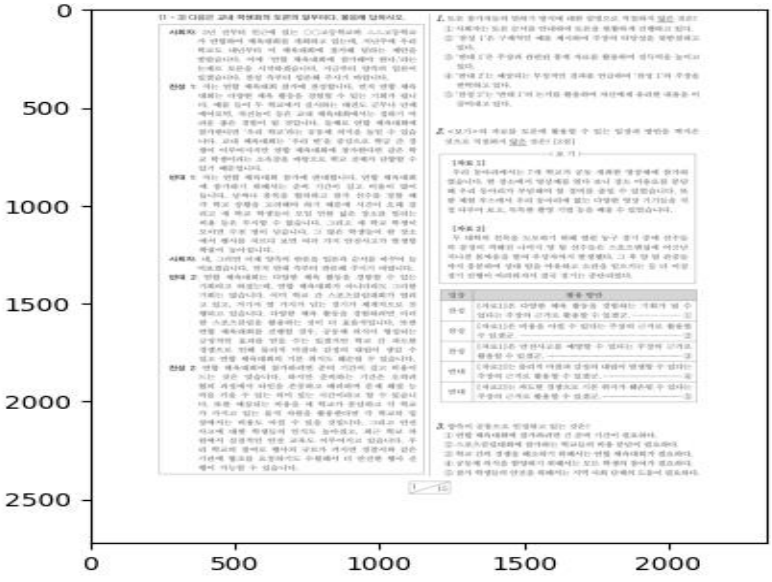
(1) 모델 – 이미지

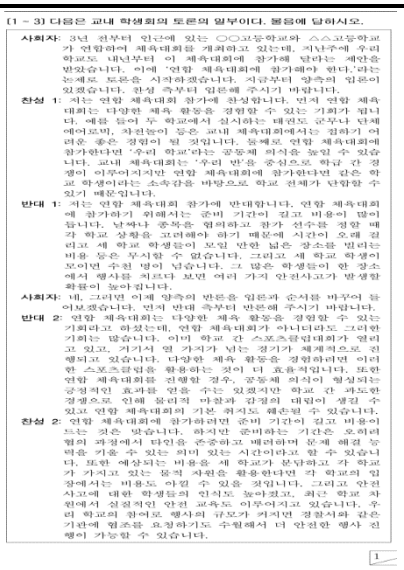
pdf에서 이미지를 추출하기 위해서는 각 pdf 페이지를 이미지 형태로 바꿔주어야 했다. 이를 위해 pdf2image 라이브러리를 사용하여 pdf를 이미지화 하였다. 이미지로 변환된 각 페이지에서 문항만 정확히 추출하기 위해서 상단에 불필요한 내용을 제거하고 좌/우를 나누는 작업을 하였다. 상단에서 불필요한 내용을 제거하기 위해 관련 블로그를 참고[xi]하여 이미지 전체를 header와 body로 분리하는 작업을 진행하였다. 이진화를 통해 윤곽선을 찾고 그 중에서 상단과 하단이 분리되는 기준 선을 찾기 위해 컨투어가 읽히는 전체 영역을 넘파이 배열로 확인하여 가로선의 특징을 찾았다. 너비가 700이상이고 높이가 50이하이며, y값이 가장 작은 가로선을 찾아 header와 body로 분리하였다. 그리고 body를 좌우로 나누기 위해 이미지 전체를 반으로 나누고 좌, 우에 5씩 여백을 주는 방식으로 좌우를 구분하였다. 이 과정에서 출제한 기관에 따라 형식이 약간씩 차이가 나타난다는 것을 발견하였다. 기존에 구성한 코드는 교육청에서 출제한 형태에 적합하였고, 평가원에서 출제한 형태에는 적합하지 않음을 깨달았다. 이를 해결하기 위해 평가원에 해당하는 고등학교 3학년 6월, 9월, 11월 데이터를 위한 별개의 전처리 코드를 구성하였다. 우선, 평가원임을 구분하기 위해 pdf 상에서 구분할 수 있는 특징을 찾았다. 평가원 문항의 경우 각 페이지 하단에 '이 문제지에 관한 저작권은 한국교육과정평가원에

있습니다.'라는 문구가 붙어있음을 발견하고 Easyocr을 사용하여 이미지를 header와 body로 구분하기 전 '한국교육과정평가원'이라는 글씨를 찾게 하여 이 단어가 발견되면 별도의 전처리 코드를 사용하도록 로직을 구성하였다. 이때, 문자 인식에 Easyocr을 사용한 이유는 다음과 같다. Easyocr은 내장된 이미지 전처리 기능을 통해 다양한 이미지 조건에서도 높은 정확도를 유지할 수 있다는 장점이 있으며, 한글과 같은 문자 인식에 높은 정확도를 보이는 것으로 알려져 있다. crop작업에서 사용하는 ocr의 경우 정확히 그 단어를 찾아야 하기 때문에 Easyocr을 사용하게 되었다. 평가원 문제지의 경우 교육청 문제지와 달리 컨투어가 읽히는 영역의 넘파이 배열을 확인한 결과 h값이 가장 큰 가로선이 header와 body를 구분해주는 선임을 찾고 이를 기준으로 상단과 하단을 분리하였으며 좌우 분리는 교육청 문제지와 동일하게 진행하였다. 이렇게 분리된 이미지는 다음과 같은 형태로 나타나게 된다.



(a) 문제 이미지 header영역





[1 ~ 3] 다음은 교내 학생회의 토론의 일부이다. 물음에 답하시오.

사회자: 3년 전부터 인근에 있는 ○○고등학교와 △△고등학교가 연합하여 체육대회를 개최하고 있는데, 지난주에 우리 학교도 대년부터 이 체육대회에 참가해 달라는 제안을 받았습니다. 이에 '연합 체육대회에 참가해야 한다.'라는 논제로 토론을 시작하겠습니다. 지금부터 양측의 입론이 있겠습니다. 찬성 측부터 입론해 주시기 바랍니다.

찬성 1: 저는 연합 체육대회 참가에 찬성합니다. 먼저 연합 체육대회는 다양한 체육 활동을 경험할 수 있는 기회가 됩니다. 예를 들어 두 학교에서 실시하는 태권도 군무나 단체 에어로빅, 치어리딩 등은 교내 체육대회에서는 경험기 어려운 좋은 경험이 될 것입니다. 등해로 연합 체육대회에 참가한다면 '우리 학교'라는 공동체 의식을 높일 수 있습니다. 교내 체육대회는 '우리 반'을 중심으로 학급 간 경쟁이 이루어지지만 연합 체육대회에 참가한다면 같은 학교 학생이라는 소속감을 바탕으로 학교 전체가 단합할 수 있기 때문입니다.

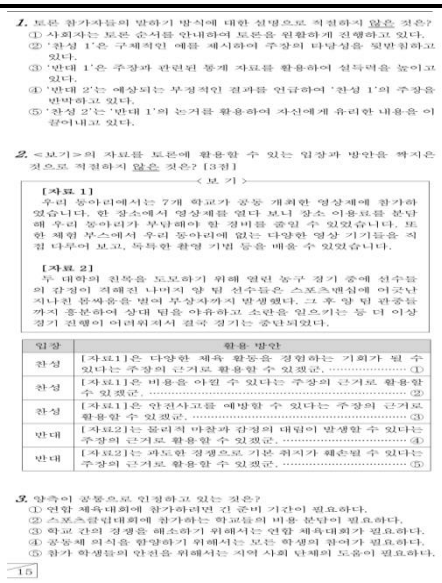
반대 1: 저는 연합 체육대회 참가에 반대합니다. 연합 체육대회에 참가하기 위해서는 준비 기간이 길고 비용이 많이 듭니다. 날씨나 중복을 염려하고 참가 선수를 정할 때 각 학교 상황을 고려해야 하기 때문에 시간이 오래 걸리고 세 학교 학생들이 모일 만한 넓은 장소를 빌리는 비용 등은 무시할 수 없습니다. 그리고 세 학교 학생이 모이면 수천 명이 넘습니다. 그 많은 학생들이 한 장소에서 행사를 치르다 보면 여러 가지 안전사고가 발생할 확률이 높아집니다.

사회자: 네, 그러면 이제 양측의 반론을 입론과 순서를 바꾸어 듣고 보겠습니다. 먼저 반대 측부터 반론해 주시기 바랍니다.

반대 2: 연합 체육대회는 다양한 체육 활동을 경험할 수 있는 기회라고 하셨는데, 연합 체육대회가 아니더라도 그러한 기회는 많습니다. 이미 학교 간 스포츠클럽대회가 열리고 있고, 가기서 열 가지가 넘는 경기가 체계적으로 진행되고 있습니다. 다양한 체육 활동을 경험하려면 이러한 스포츠클럽을 활용하는 것이 더 효율적입니다. 또한 연합 체육대회를 진행할 경우, 공동체 의식이 형성되는 긍정적인 효과를 얻을 수는 있겠지만 학교 간 과도한 경쟁으로 인해 물리적 마찰과 감정의 대립이 생길 수 있고 연합 체육대회의 기본 취지도 훼손될 수 있습니다.

찬성 2: 연합 체육대회에 참가하려면 준비 기간이 길고 비용이 드는 것은 맞습니다. 하지만 준비하는 기간은 오히려 협의 과정에서 타인을 존중하고 배려하며 문제 해결 능력을 키울 수 있는 의미 있는 시간이라고 할 수 있습니다. 또한 예상되는 비용을 세 학교가 분담하고 각 학교가 가지고 있는 물적 자원을 활용한다면 각 학교의 입장에서는 비용도 아낄 수 있을 것입니다. 그리고 안전사고에 대한 학생들의 인식도 높아졌고, 최근 학교 차원에서 실질적인 안전 교육도 이루어지고 있습니다. 우리 학교의 참여로 행사의 규모가 커지면 경찰서와 같은 기관에 협조를 요청하기도 수월해서 더 안전한 행사 진행이 가능할 수 있습니다.

1



1. 토론 참가자들의 말하기 방식에 대한 설명으로 적절하지 않은 것은?
① 사회자는 토론 순서를 안내하여 토론을 원활하게 진행하고 있다.
② '찬성 1'은 구체적인 예를 제시하여 주장의 타당성을 뒷받침하고 있다.
③ '반대 1'은 주장과 관련된 통계 자료를 활용하여 설득력을 높이고 있다.
④ '반대 2'는 예상되는 부정적인 결과를 언급하여 '찬성 1'의 주장을 반박하고 있다.
⑤ '찬성 2'는 '반대 1'의 논거를 활용하여 자신에게 유리한 내용을 이끌어내고 있다.

2. <보기>의 자료를 토론에 활용할 수 있는 입장과 방안을 짝지은 것으로 적절하지 않은 것은? [3점]

< 보 기 >

[자료 1]
우리 동아리에서는 7개 학교가 공동 개최한 영상제에 참가하였습니다. 한 장소에서 영상제를 열다 보니 장소 이용료를 분담해 우리 동아리가 부담해야 할 경비를 줄일 수 있었습니다. 또한 체험 부스에서 우리 동아리에 없는 다양한 영상 기기들을 직접 다루어 보고, 독특한 촬영 기법 등을 배울 수 있었습니다.

[자료 2]
두 대학의 친목을 도모하기 위해 열린 농구 경기 중에 선수들의 감정이 격해진 나머지 양 팀 선수들은 스포츠맨십에 어긋난 지나친 몸싸움을 벌여 부상자까지 발생했다. 그 후 양 팀 관중들까지 흥분하여 상대 팀을 야유하고 소란을 일으키는 등 더 이상 경기 진행이 어려워져서 결국 경기는 중단되었다.

입장	활용 방안
찬성	[자료1]은 다양한 체육 활동을 경험하는 기회가 될 수 있다는 주장의 근거로 활용할 수 있겠군. ····· ①
찬성	[자료1]은 비용을 아낄 수 있다는 주장의 근거로 활용할 수 있겠군. ····· ②
찬성	[자료1]은 안전사고를 예방할 수 있다는 주장의 근거로 활용할 수 있겠군. ····· ③
반대	[자료2]는 물리적 마찰과 감정의 대립이 발생할 수 있다는 주장의 근거로 활용할 수 있겠군. ····· ④
반대	[자료2]는 과도한 경쟁으로 기본 취지가 훼손될 수 있다는 주장의 근거로 활용할 수 있겠군. ····· ⑤

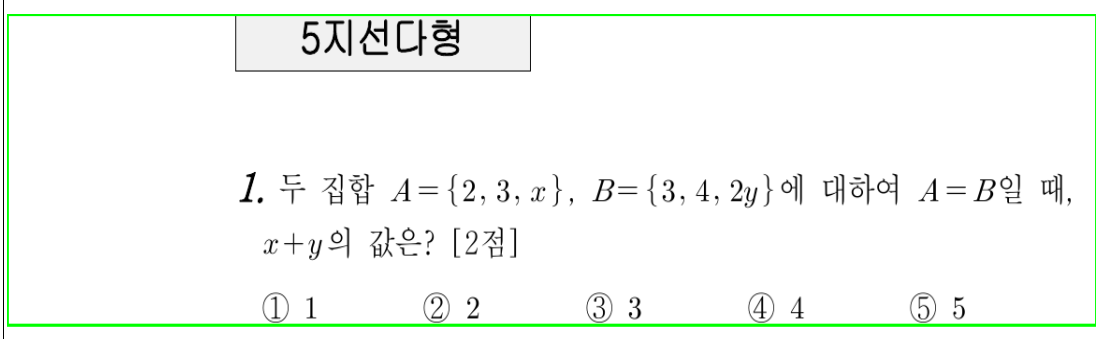
3. 양측이 공통으로 인정하고 있는 것은?
① 연합 체육대회에 참가하려면 긴 준비 기간이 필요하다.
② 스포츠클럽대회에 참가하는 학교들의 비용 분담이 필요하다.
③ 학교 간의 경쟁을 해소하기 위해서는 연합 체육대회가 필요하다.
④ 공동체 의식을 함양하기 위해서는 모든 학생의 참여가 필요하다.
⑤ 참가 학생들의 안전을 위해서는 지역 사회 단체의 도움이 필요하다.

15

(b) 문제 이미지 body영역

(c) 문제 이미지 body영역 좌/우로 분리

이러한 형태로 pdf를 이미지화 하여 분리한 뒤 각 문항을 자르는 작업을 과목에 따라 다르게 실행하였다. 우선, 수학의 경우 각 문항들이 넓은 빈 공간을 두고 문항간 분리가 명확하여 관련 블로그[xii]를 참고하여 컨투어 영역 확인을 통해 문항을 분리하였고 흰색과 아닌 색의 경계를 찾아 경계 좌표를 기준으로 문항 밖의 여백을 자르는 작업을 추가로 진행하였다. 또한, 수학의 경우 '5지선다형', '단답형'이라는 부분이 문항과 같이 붙어서 crop되는 문제가 발생하여, 이 글자를 Easyocr로 읽어 해당 키워드의 아래 부분에서 20의 offset만큼 아래 영역부터 상단까지 이미지를 자르도록 하여 최종 문항 이미지를 산출할 수 있었다. 다음은 최종 문항 이미지 산출 과정이다.



5지선다형

1. 두 집합 $A=\{2, 3, x\}$, $B=\{3, 4, 2y\}$에 대하여 $A=B$일 때, $x+y$의 값은? [2점]

① 1 ② 2 ③ 3 ④ 4 ⑤ 5

(a) 컨투어 영역 찾아서 컨투어 영억에 따라 crop

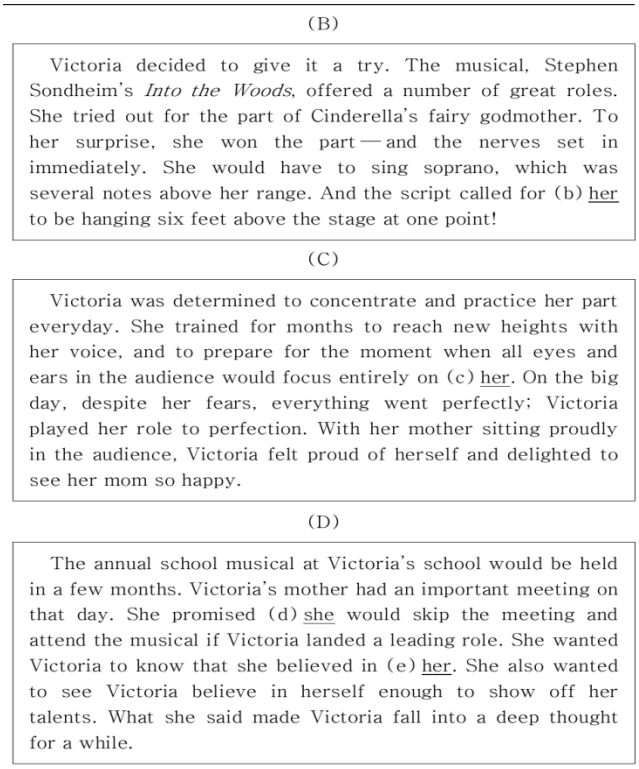
1. 두 집합 $A=\{2, 3, x\}$, $B=\{3, 4, 2y\}$에 대하여 $A=B$일 때, $x+y$의 값은? [2점]

① 1 ② 2 ③ 3 ④ 4 ⑤ 5

(b) 최종 산출된 crop 이미지

그러나 국어, 영어의 경우 문항간 불규칙하게 여백이 형성되어 있고 문항의 길이 또한 다양하게 이루어져 있기 때문에 컨투어 영역으로는 정확히 문항만 자르지 못하는 문제가 발생하였다. 문제를 해결하기 위해 템플릿을 사용하였다. 문항숫자와 동일한 형태의 템플릿을 따와 동일한 형태를 매칭하여 이미지를 crop 하도록 하였으나, 템플릿과 이미지 간의 해상도 차이로 인해 이미지가 원하는 대로 crop되지 못하는 문제가 발생하였다. 때문에 최종적으로는 정규식을 사용하여 문항번호를 찾아 Easyocr로 읽은 후 crop하는 방식을 사용하였다. OCR을 위한 이미지 픽셀은 1000픽셀의 이미지에도 기존 픽셀을 고려하지 않은 이미지와 유사한 성능을 보였기에 이미지의 해상도를 위한 전처리는 하지 않았다. 이 때, 영어 문제를 고려하여 Easyocr의 학습 데이터로 한국어 데이터와 영어 데이터를 함께 사용하였다. 정규식을 패턴으로 사용하여 '[숫자 ~숫자]', '숫자.' , '숫자. 윗글', '[숫자' 패턴를 찾도록 하여 지문영역과 문제영역 모두를 crop할 수 있도록 하였다. crop되는 영역은 패턴을 발견한 영역부터 다음 패턴이 발견된 영역까지로 지정하였고, 마지막 문제의 경우는 다음 패턴이 발견되지 않을 경우 이미지 끝까지 crop되도록 하였다. 또한, 이미지 상단 부분부터 문항번호가 나타나지 않는 경우가 발생할 수 있으므로 이미지의 50 픽셀까지 문제번호 패턴을 찾지 못하는 경우 이미지 최상단부터 패턴을 찾을 때까지 crop하도록 하였다. 마지막으로, 국어, 영어의 경우 지문이나 보기로 인해 다음 단으로까지 문제가 이어지는 경우가 발생하여 이 경우는 직접 단이 나눠진 파일을 찾아 파일명으로 '-1', '-2'로 태깅을 해준 후 이미지를 합치는 작업을

진행하였다. 이미지를 합치는 작업은 두 이미지의 크기를 가져온 후 가로 길이를 비교하여 작은 길이를 큰 길이에 맞추어 리사이즈를 해준 후 '-1'을 상단, '-2'를 하단에 위치하여 합쳐 최종 문항 이미지를 산출하였다. 이 과정에서 pdf에서 추출하는 이미지의 경우 다음은 국어, 영어 과목의 최종 문항 이미지 산출 과정이다.



(B)

Victoria decided to give it a try. The musical, Stephen Sondheim's *Into the Woods*, offered a number of great roles. She tried out for the part of Cinderella's fairy godmother. To her surprise, she won the part — and the nerves set in immediately. She would have to sing soprano, which was several notes above her range. And the script called for (b) her to be hanging six feet above the stage at one point!

(C)

Victoria was determined to concentrate and practice her part everyday. She trained for months to reach new heights with her voice, and to prepare for the moment when all eyes and ears in the audience would focus entirely on (c) her. On the big day, despite her fears, everything went perfectly; Victoria played her role to perfection. With her mother sitting proudly in the audience, Victoria felt proud of herself and delighted to see her mom so happy.

(D)

The annual school musical at Victoria's school would be held in a few months. Victoria's mother had an important meeting on that day. She promised (d) she would skip the meeting and attend the musical if Victoria landed a leading role. She wanted Victoria to know that she believed in (e) her. She also wanted to see Victoria believe in herself enough to show off her talents. What she said made Victoria fall into a deep thought for a while.

(a) 이미지의 50픽셀까지 문제번호 패턴을 찾지 못한 경우

[43 ~ 45] 다음 글을 읽고, 물음에 답하시오.

(A)

William Miller stayed up after the family had gone to bed, then read until the morning. Candles were expensive, but there were plenty of pine knots, and all (a) he had to do was gather them from the woods. So William formed the habit of burning pine knots in the fireplace for his nightly reading light.

* pine knot: 관솔(송진이 엉긴 소나무의 옹이)

(B)

William's "secret life" continued for some time, though. Night after night he read as long as he could, then made (b) his way back upstairs, and slept until it was time to do the morning chores. But one night something happened that he hadn't expected. His father awoke and saw a glow downstairs. Thinking the house was on fire, (c) he came rushing down the stairs to save his home and family from going up in flames.

(C)

Instead of a house fire, however, he saw his son William lying peacefully before the fireplace reading a book he'd borrowed from a neighbor. His father grabbed a broomstick and chased his son around the room, yelling, "Young man, if you don't get to bed right now, I'll kick you out of the house!" William went up to bed, at least for this night. (d) He was only trying to get an education that he couldn't get from the teachers in the community.

(D)

But his father didn't like the habit and tried to stop it. His father felt that his son's late-night reading would cut into (e) his energy for the next day's work. And the farm required every ounce of work he could get from his son. He insisted that William retire for the night when the rest of the family did. And his father thought the growing boy should sleep soundly through the night.

44. 밑줄 친 (a)~(e) 중에서 가리키는 대상이 나머지 넷과 다른 것은?

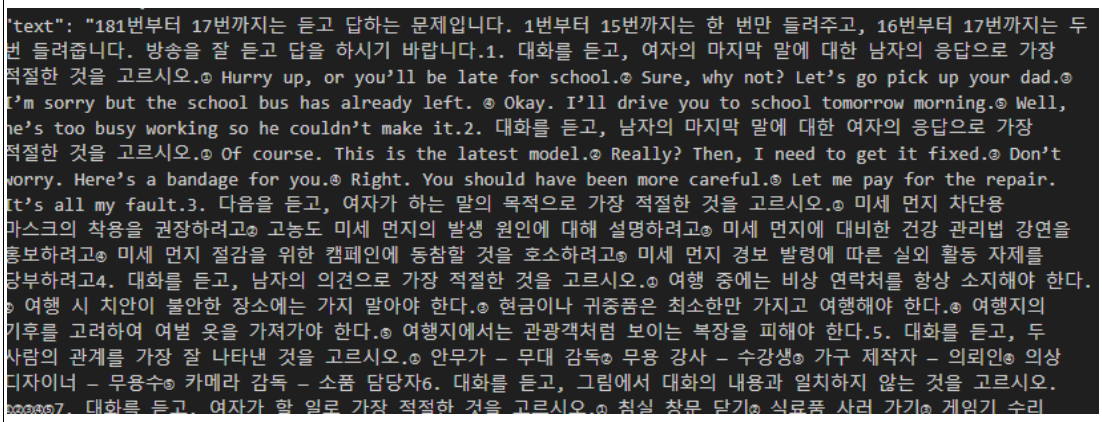
① (a) ② (b) ③ (c) ④ (d) ⑤ (e)

(b) 단이 나뉜 문제를 하나로 합친 이미지

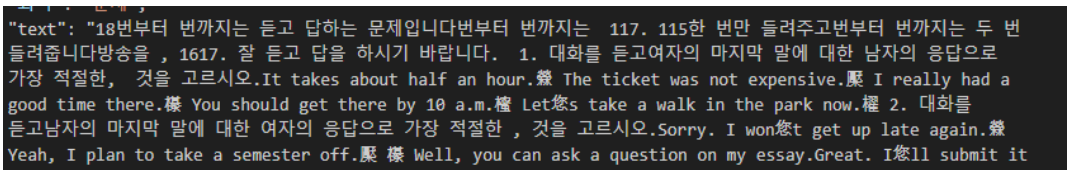
이와 같은 crop 방식을 사용하여 최종 이미지 crop 결과를 산출하였다.

(2) 모델 – 텍스트 및 텍스트 임베딩

국어, 영어, 수학 과목에 대해 수능 및 모의고사의 문항과 정답지, 해설지를 Tesseract OCR을 사용하여 파싱할 수 있는 코드를 정리하고자 하였다. OCR로는 Easyocr과 비교 결과 파싱의 정확도에는 큰 차이가 없는 것으로 판단되어 다량의 데이터를 빠른 속도로 파싱할 수 있는 Tesseract OCR를 선택하게 되었다. 국어, 영어 파싱 과정에서 동일한 모의고사임에도 평가원, 교육청에 따라 파싱 오류가 발생하는 경우를 확인하였다. 이를 해결하기 위해 파싱이 되지 않는 것들을 파악하여 새로운 코드를 적용하였다. 그러나, 영어가 한자로 파싱되는 등 전혀 다른 내용으로 파싱되는 경우는 코드수정, 로직수정의 방법을 사용하였으나 OCR로는 해결방법을 찾지 못하여 파싱된 파일의 정교성을 높이는 방향으로 진행하였다.



(a) 정상적으로 파싱된 문항



(b) 영어가 한자로 파싱되는 오류

반면, 수학의 경우 기존 Tesseract OCR로 수학기호를 파싱하고자 하였으나 OCR로는 수학기호를 아예 읽지 못하는 문제가 발생하였다. 이를 해결하기 위해 다양한 OCR모델 사용, math fix 프로그램의 API 사용의 방법을 사용하였으나

	다른 OCR모델 역시 비슷한 수준의 잘못된 파싱이 진행되었고, math fix 프로그램의 API 사용은 한 계정당 무료 1000개의 문제만이 제공되어 프로그램을 사용하여 파싱하는 방법은 적합하지 않다고 판단하였다. 따라서, 최종적으로 OpenAI의 API key를 활용하여 프롬프트 엔지니어링을 통해 수학 공식을 파싱하는 것이 가장 적합하다고 판단하여 프롬프트 엔지니어링을 통해 수학과목 파싱을 진행하였다. 프롬프트 엔지니어링에 사용한 모델은 gpt4o을 사용하였다. OpenAI사의 모델들 중 gpt4와 비교를 해보았을 때 gpt4는 데이터 분석 및 글쓰기에 적합한 모델인 반면 gpt4o는 실시간 처리 응답 속도가 가장 빠르고 적은 비용으로 사용할 수 있는 모델로 판단되어 gpt4o모델을 사용하였다. 파싱된 문항 데이터들 간 유사도를 확인하기 전 텍스트 임베딩을 실시하였다. 임베딩 모델은 국어, 영어, 수학 과목에 동일하게 'jhgan/KO-BERT-STS' 모델을 사용하였다. 위 모델은 BERT 기반 모델로서 다양한 자연어 처리에 적합한 모델이다. 또한, 한국어 데이터셋으로 추가 학습되어 한국어 문장 이해에 높은 성능을 발휘하는 한국어 최적화 모델이기에 한글 임베딩이 요구되는 현 프로젝트에 있어 유사도 평가에서 높은 정확도를 제공할 것으로 보았다. 이는 실제로 임베딩 모델 선정 과정에서 'jhgan/ko-sbert-sts', 'all-MiniLM-L6-v2', 'distilbert-base-nli-stsb-mean-tokens'의 모델과 비교, 사용해본 결과 'all-MiniLML6-v2', 'distilbert-base-nli-stsb-mean-tokens' 모델의 경우 70&를 웃도는 수준의 유사도가 나타났으나 'jhgan/KO-BERT-STS' 모델은 80~90% 수준의 유사도가 나타나 데이터셋에 가장 높은 성능을 보였기에 'jhgan/KO-BERT-STS' 모델을 채택하게 되었다.

```
Similarity Scores:
Index: 40, Similarity Score: 0.7379857897758484
Index: 4715, Similarity Score: 0.5531903505325317
Index: 3460, Similarity Score: 0.5439576506614685
Index: 769, Similarity Score: 0.5245111584663391
Index: 2245, Similarity Score: 0.5173482894897461
Index: 2739, Similarity Score: 0.5032526850700378
Index: 3787, Similarity Score: 0.5025597214698792
Index: 4400, Similarity Score: 0.5003668069839478
Index: 3821, Similarity Score: 0.49896669387817383
Index: 1572, Similarity Score: 0.4984367787837982
Index: 4892, Similarity Score: 0.4957675337791443
Index: 4368, Similarity Score: 0.4955410957336426
Index: 4054, Similarity Score: 0.49430668354034424
Index: 1944, Similarity Score: 0.4924786686897278
Index: 845, Similarity Score: 0.4916988015174866
Index: 4815, Similarity Score: 0.4913559556007385
Index: 1427, Similarity Score: 0.48853427171707153
Index: 3070, Similarity Score: 0.4818412065505981 4
Index: 2572, Similarity Score: 0.4798809587955475
Index: 1163, Similarity Score: 0.47917553782463074
```

(a) all-MiniLM-L6-v2 모델로 임베딩 후 유사도 결과

```
Similarity Scores:
Index: 4947, Similarity Score: 0.6923117637634277
Index: 2475, Similarity Score: 0.6860746145248413
Index: 1668, Similarity Score: 0.6655818819999695
Index: 1578, Similarity Score: 0.6581923365592957
Index: 4366, Similarity Score: 0.6394253373146057
Index: 2792, Similarity Score: 0.6347850561141968
Index: 986, Similarity Score: 0.6177663803100586
Index: 336, Similarity Score: 0.6086564064025879
Index: 1078, Similarity Score: 0.6066728830337524
Index: 2144, Similarity Score: 0.5975664258003235
Index: 4453, Similarity Score: 0.596991240978241
Index: 2637, Similarity Score: 0.592575192451477
Index: 2382, Similarity Score: 0.5917442440986633
Index: 468, Similarity Score: 0.5842276215553284
Index: 2881, Similarity Score: 0.5837739109992981
Index: 2703, Similarity Score: 0.5830985903739929
Index: 45, Similarity Score: 0.5810667276382446
Index: 988, Similarity Score: 0.5792607069015503
Index: 3285, Similarity Score: 0.5781310796737671
Index: 3380, Similarity Score: 0.5766295194625854
```

(b) distilbert-base-nli-stsb-mean-tokens 모델로 임베딩 후 유사도 결과

여기서 주목할만한 점은 영어문장이 대부분인 영어 교과에 있어서도 한국어 특화 모델이 가장 높은 성능을 보였다는 점이다. 이는 미루어 예상컨데 지문이 아닌 문제는 대부분 한글로 구성되어 있어 이것이 영향을 미쳤을 것이라고 보고 있다. 그러나 영어의 경우 한국어 특화모델만을 온전히 믿고 사용하기에는 정확한 임베딩이 나타나지 않을 것이라고 생각되어 임베딩 모델을 사용하기 전 추가로 더 조정을 주기 위해 nltk 패키지를 사용하여 불용어사전을 추가한 후 모델을 사용하였다.

```
Similarity Scores:
Index: 484, Similarity Score: 0.7310471534729004
Index: 4944, Similarity Score: 0.730719804763794
Index: 1162, Similarity Score: 0.7111301422119141
Index: 3292, Similarity Score: 0.6984366178512573
Index: 1394, Similarity Score: 0.6940234899520874
Index: 1068, Similarity Score: 0.6926692128181458
Index: 658, Similarity Score: 0.6900254487991333
Index: 3559, Similarity Score: 0.6844321489334106
Index: 4954, Similarity Score: 0.6820858120918274
Index: 3251, Similarity Score: 0.670322835445404
Index: 434, Similarity Score: 0.6699002385139465
Index: 3509, Similarity Score: 0.6681149005889893
Index: 2339, Similarity Score: 0.6678752899169922
Index: 1487, Similarity Score: 0.6671552062034607
Index: 3203, Similarity Score: 0.6656055450439453
Index: 211, Similarity Score: 0.6601696610450745
Index: 4317, Similarity Score: 0.6545317769050598
Index: 4593, Similarity Score: 0.6539310812950134
Index: 2473, Similarity Score: 0.6503515243530273
Index: 3143, Similarity Score: 0.6470848917961121
```

(c) jhganko-sbert-sts 모델 유사도 결과-불용어 처리 전

```
Similarity Scores:
Index: 426, Similarity Score: 0.8666638135910034
Index: 245, Similarity Score: 0.8259157538414001
Index: 246, Similarity Score: 0.8259157538414001
Index: 4213, Similarity Score: 0.8033315539360046
Index: 1467, Similarity Score: 0.7964079976081848
Index: 337, Similarity Score: 0.78397136926651
Index: 1336, Similarity Score: 0.7783372402191162
Index: 1154, Similarity Score: 0.7767747044563293
Index: 1419, Similarity Score: 0.7753781080245972
Index: 105, Similarity Score: 0.7714402079582214
Index: 152, Similarity Score: 0.7688501477241516
Index: 1016, Similarity Score: 0.7566654682159424
Index: 2147, Similarity Score: 0.7524469494819641
Index: 2326, Similarity Score: 0.7460530400276184
Index: 4680, Similarity Score: 0.7388414144515991
Index: 1513, Similarity Score: 0.7376527190208435
Index: 471, Similarity Score: 0.7350379228591919
Index: 4392, Similarity Score: 0.7339051961898804
Index: 1787, Similarity Score: 0.730992317199707
Index: 4258, Similarity Score: 0.7302708029747009
```

(d) jhganko-sbert-sts 모델 유사도 결과-불용어 처리 후

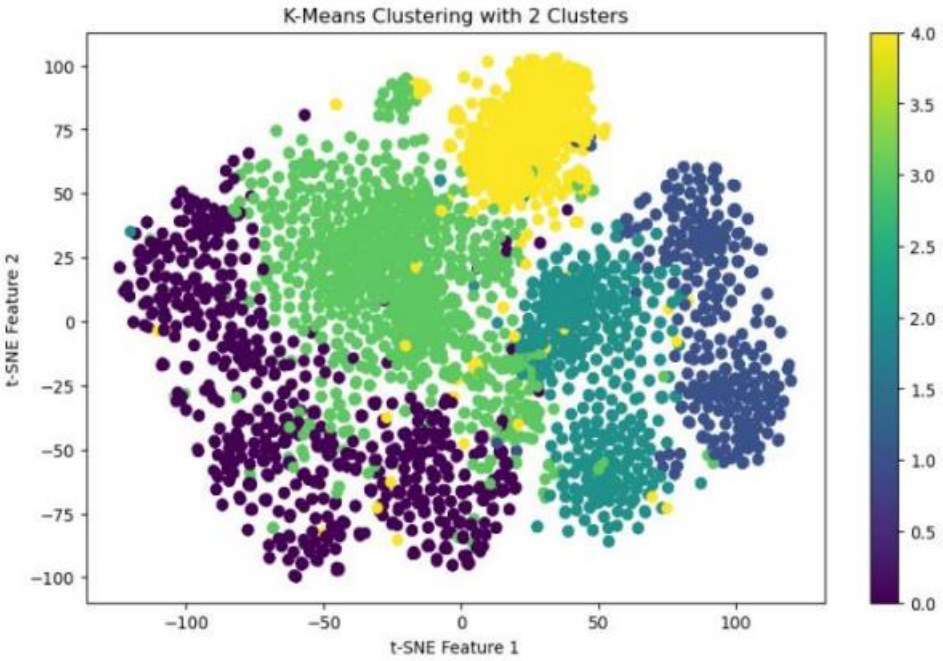
내부문제로 산출된 최종 임베딩 벡터값과 유사도 벡터값을 활용하여 외부문제와 비교하여 검증을 실시하고 실제 사용자가 외부문제를 넣었을 때 유사문항을 추천하고자 외부문제도 내부문제와 동일하게 임베딩, 유사도 값을 산출하기 위해 텍스트 파싱 과정을 진행하였다. 수학의 경우 기존 문제지 문항 파싱에 사용하였던 프롬프트 엔지니어링을 그대로 사용하였으나, 국어와 영어 과목의 경우 이미지 화질에 따라 OCR이 읽히는 정도가 너무 다르게 나타난다는 문제가 발생하여 다양한 화질의 이미지에 대해서도 유사문항 추천을 안정적으로 진행하기 위해 기존 OCR을 활용한 파싱 방법과 달리 프롬프트 엔지니어링을

사용하게 되었다. 이 과정에서 gpt4o-mini 모델의 개발로 OpenAI사의 gpt4o 모델과 gpt4o-mini를 비교하였다. 비교 결과 두 모델 간 정확도 및 속도 측면에서 차이는 크게 다르지 않았으나 비용 측면에서 큰 차이를 보였다. 비용 측정 결과 프롬프트 엔지니어링을 통해 파싱한 문항은 한 문항당 gpt4o 모델은 평균적으로 수학 0.14원, 국어 2원, 영어 0.14원의 비용이 요구되는 것으로 확인되었으나. gpt4o-mini 모델은 평균적으로 수학 1.8원, 국어 7원, 영어 1.9원의 비용이 요구되는 것으로 확인되었다. 이 결과는 모델 별 토큰 수 차이로 인해 발생되는 것으로 확인되었다. gpt4o-mini모델의 경우 토큰 가격이 저렴하다는 장점이 있었으나 실제 한 문항 당 토큰 수를 살펴보니 gpt4o에서는 토큰이 약 400여개 소요되나, gpt4o-mini의 경우 동일한 문제에서 사용되는 토큰 수가 약8500개로 나타났다. 때문에, gpt4o-mini는 토큰 가격이 저렴하나 토큰 수의 큰 차이로 인해 오히려 한 문항당 더 비싼 비용이 나타나게 되었다. 이에 따라 사용자가 넣는 외부문항 프롬프트 엔지니어링에는 gpt4o를 사용하였다.

(3) 모델 – 자동태깅

자동태깅을 실시하기 위해 우선 이미지 데이터를 사용하였다. 이미지 데이터를 가지고 DBSCAN, K-means 등 비지도 학습 모델을 중심으로 문항 분류를 실시하였다. 그러나, DBSCAN의 경우 과목 외에 아무런 조건 없이 이미지 데이터를 사용하자 클러스터 수가 658개가 나오는 등 전혀 분류를 하지 못하는 모습을 보였다. K-means의 경우 클러스터별 실제 분류된 데이터를 확인해보았을 때 부정확하게 분류하는 모습을 보여 이미지 데이터를 활용한 방식은 적합하지 않다고 여겨 최종적으로는 텍스트 데이터를 중심으로 문항 분류를 하였다.
국어와 영어 과목의 경우 파싱 과정에서 높은 성능을 보였던 'jhgan/KO-BERT-STS' 모델을 sentence-transformers를 활용하여 불러와 임베딩을 실시하여 그 결과를 바탕으로 K-means 클러스터링을 진행하였다. 클러스터의 수는 실루엣 계수와 엘보우 그래프를 활용하여 선정하였다. 국어 교과의 경우 5개의

클러스터로 분리되었고 각 분류는 현대문학, 고전문학, 비문학, 화법과 작문, 언어와 매체(문법)으로 나뉘어졌다.



(a) 국어교과 클러스터 분류 시각화

```
25  연재가 파탄에 직면한 것은 우묵배미의 맨 꼭대기 부잣집 김 씨네에서 어쩔 수 없이 ...
26  연재가 파탄에 직면한 것은 우묵배미의 맨 꼭대기 부잣집 김 씨네에서 어쩔 수 없이 ...
27  연재가 파탄에 직면한 것은 우묵배미의 맨 꼭대기 부잣집 김 씨네에서 어쩔 수 없이 ...
28  연재가 파탄에 직면한 것은 우묵배미의 맨 꼭대기 부잣집 김 씨네에서 어쩔 수 없이 ...
71  그의 대학 재학 시기 역시 학생 시위가 빈발하던 한일회담 진행기를 전후하고 있었다....
72  그의 대학 재학 시기 역시 학생 시위가 빈발하던 한일회담 진행기를 전후하고 있었다....
73  그의 대학 재학 시기 역시 학생 시위가 빈발하던 한일회담 진행기를 전후하고 있었다....
87  (가) 차례를 지내고 돌아온 구두 밑바닥에 고향의 저문 강물소리가 묻어 있다 겨울보...
88  (가) 차례를 지내고 돌아온 구두 밑바닥에 고향의 저문 강물소리가 묻어 있다 겨울보...
89  (가) 차례를 지내고 돌아온 구두 밑바닥에 고향의 저문 강물소리가 묻어 있다 겨울보...
```

(b) 국어교과 클러스터 분류 – 현대문학

```
15  (가) ㉠아득한 녯날에 나는 떠났다 부여를 숙신을 발해를 여진을 요를 금을 흥안령을...
16  (가) ㉠아득한 녯날에 나는 떠났다 부여를 숙신을 발해를 여진을 요를 금을 흥안령을...
17  (가) ㉠아득한 녯날에 나는 떠났다 부여를 숙신을 발해를 여진을 요를 금을 흥안령을...
18  <앞부분 줄거리> 양태백은 첩 송녀에게 미혹되어 부인과 세 남매 를 내쫓는다. 부인...
19  <앞부분 줄거리> 양태백은 첩 송녀에게 미혹되어 부인과 세 남매 를 내쫓는다. 부인...
20  <앞부분 줄거리> 양태백은 첩 송녀에게 미혹되어 부인과 세 남매 를 내쫓는다. 부인...
35  (가) 팔월이라 중추되니 백로 추분 절기로다 북두칠성 자로 돌아 서천(西天)을 가리...
36  (가) 팔월이라 중추되니 백로 추분 절기로다 북두칠성 자로 돌아 서천(西天)을 가리...
37  (가) 팔월이라 중추되니 백로 추분 절기로다 북두칠성 자로 돌아 서천(西天)을 가리...
38  (가) 팔월이라 중추되니 백로 추분 절기로다 북두칠성 자로 돌아 서천(西天)을 가리...
```

(c) 국어교과 클러스터 분류 – 고전문학

```
21  시야란 시선을 한곳에 고정하고 한 번에 볼 수 있는 범위를 의미한다. 한쪽 눈의 시...
22  시야란 시선을 한곳에 고정하고 한 번에 볼 수 있는 범위를 의미한다. 한쪽 눈의 시...
23  시야란 시선을 한곳에 고정하고 한 번에 볼 수 있는 범위를 의미한다. 한쪽 눈의 시...
24  시야란 시선을 한곳에 고정하고 한 번에 볼 수 있는 범위를 의미한다. 한쪽 눈의 시...
29  현대 사회는 정보 통신 기술의 발달로 매일 엄청난 양의 자 료가 생성 · 축적되고 있다...
30  현대 사회는 정보 통신 기술의 발달로 매일 엄청난 양의 자 료가 생성 · 축적되고 있다...
31  현대 사회는 정보 통신 기술의 발달로 매일 엄청난 양의 자 료가 생성 · 축적되고 있다...
32  현대 사회는 정보 통신 기술의 발달로 매일 엄청난 양의 자 료가 생성 · 축적되고 있다...
33  현대 사회는 정보 통신 기술의 발달로 매일 엄청난 양의 자 료가 생성 · 축적되고 있다...
34  현대 사회는 정보 통신 기술의 발달로 매일 엄청난 양의 자 료가 생성 · 축적되고 있다...
```

(d) 국어교과 클러스터 분류 – 비문학

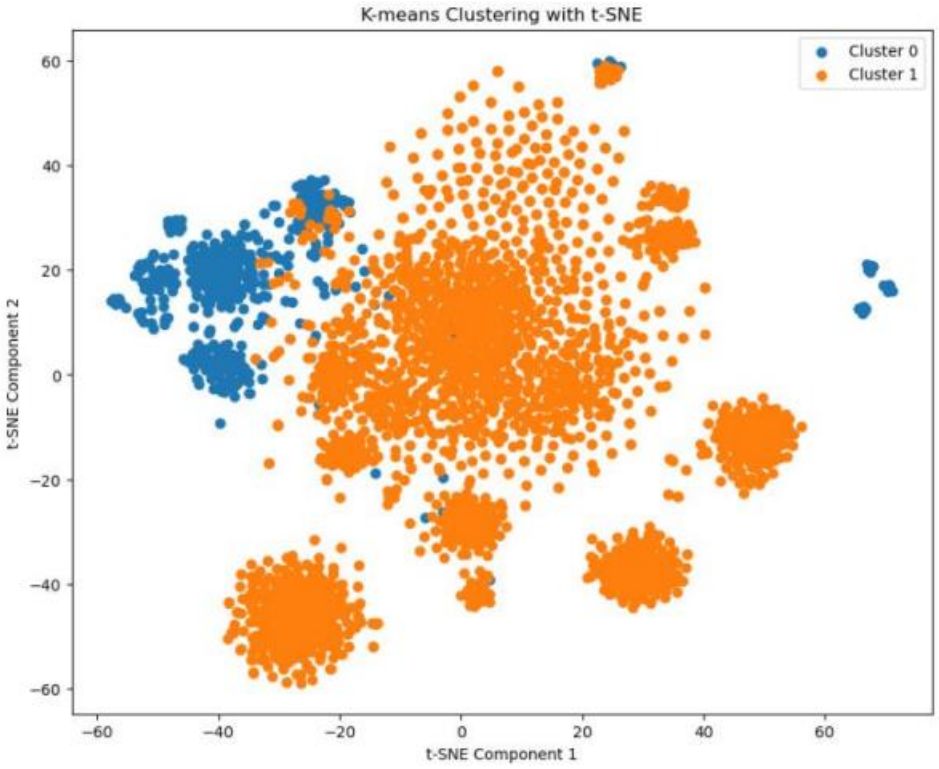
```
0  지난 체험 학습 때 저희 천문대에 오셔서 별을 관측했던 것 을 기억하시죠? (대답을...
1  지난 체험 학습 때 저희 천문대에 오셔서 별을 관측했던 것 을 기억하시죠? (대답을...
2  지난 체험 학습 때 저희 천문대에 오셔서 별을 관측했던 것 을 기억하시죠? (대답을...
3  (가) 학생1 : ㉠도서부에서 매년 진행하고 있는 '∞ 독서 대화'의 참여 인원이...
4  (가) 학생1 : ㉠도서부에서 매년 진행하고 있는 '∞ 독서 대화'의 참여 인원이...
5  (가) 학생1 : ㉠도서부에서 매년 진행하고 있는 '∞ 독서 대화'의 참여 인원이...
6  (가) 학생1 : ㉠도서부에서 매년 진행하고 있는 '∞ 독서 대화'의 참여 인원이...
7  (가) 학생의 일기 ∞로부터 최근에 겪은 일을 들었다. ㉠친구 관계를 고민 하면서...
8  (가) 학생의 일기 ∞로부터 최근에 겪은 일을 들었다. ㉠친구 관계를 고민 하면서...
9  (가) 학생의 일기 ∞로부터 최근에 겪은 일을 들었다. ㉠친구 관계를 고민 하면서...
```

(e) 국어교과 클러스터 분류 – 화법과 작문

```
10    <보기>의 ⓐ~ ⓓ를 발음할 때 일어나는 음운 변동을 탐구한 내용으로 적절한 것...
11  관형사형 어미는 용언의 어간에 붙어 용언이 관형사와 같 은 기능을 수행하게 하는 어...
12  관형사형 어미는 용언의 어간에 붙어 용언이 관형사와 같 은 기능을 수행하게 하는 어...
13    <보기>의 밑줄 친 단어의 품사에 대한 이해로 적절하지 않은 것은? < 보 기 ...
14    <보기>에 제시된 '선생님'의 질문에 대한 답으로 적절하지 않은 것은? < 보 ...
55  현대 국어에서는 음절의 종성에서 실제로 발음되는 소리가 제한되어 있다. ㉠음절의 종..
56  현대 국어에서는 음절의 종성에서 실제로 발음되는 소리가 제한되어 있다. ㉠음절의 종..
57    <보기>의 선생님 물음에 대한 답으로 가장 적절한 것은? < 보 기 >선생님: ...
58    사전 자료의 일부인 <보기>를 바탕으로 어미의 쓰임을 탐구한 학습지 활동의 결과...
59    <보기>의 ㉮~㉰에 대한 설명으로 적절하지 않은 것은? < 보 기 >㉮ 그 사람...  }
```

(f) 국어교과 클러스터 분류 – 언어와 매체(문법)

영어 교과의 경우 영어과의 각론을 참고하여 문항을 분류하였다. 우선적으로 2개의 클러스터로 나누어 '그림, 사진, 도표, 대상, 주제' 분류인 것과 아닌 것을 분리하였고, 아닌 것으로 분리된 것 중에서 세부 클러스터링을 다시 진행하여 3개의 클러스터로 구분하였다. 세부 클러스터링으로 분류된 각 클러스터는 '일이나 사건의 순서/전후 관계/원인/결과, 필자의 의도/목적, 필자의 심경/태도'와 '빈칸에 들어갈 문장/단어 찾기' 그리고 '문맥 속 낱말/어구/문장의 의미, 글의 숨겨진 의미, 줄거리/주제/요지'의 특징을 띈 문항들로 구성되어 있음을 확인할 수 있었다.



(a) 영어교과 클러스터 분류 시각화

Cluster 0 - 연도별 상위 3문제:

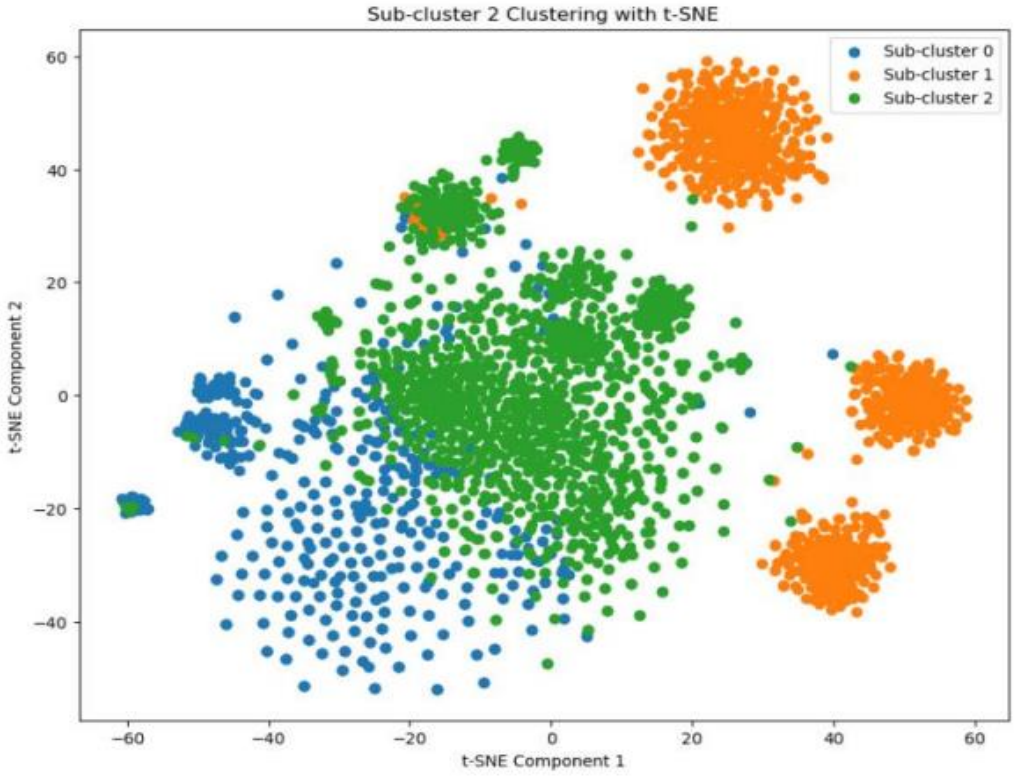
	grade	yyyy	mm	host	subject_cat	question_num	combined_text
1812	2	2016	11	0	1	24	다음 도표의 내용과 일치하지 않는 것은? Preferred Banking Method, 2013-2014 pie charts show preferred banking method based annual survey 1,000 U.S. adults 2013 2014 American Bankers Association. sum percentages respondents preferring Internet Banking Branches surpassed 50 percent years. 2013, 39 percent respondents named Internet Banking favorite way conducting banking, preference dropped 8 percentage points 2014. However, compared 2013, preference Branches ATMs increased 3 percentage points respectively 2014. 2013, preference Mobile less 10 percent, reached double digits 2014. 2013 2014, Telephone Mail remained preference 7 percent.'t Know10%Mail MI1796 Internet Banking909095 Internet Banking 31%the 39% SMobile8% . QC | 10% .. . ~Branches /18% WZBranches2013 201421%영어 영역48
3423	1	2015	9	1	1	25	다음 도표의 내용과 일치하지 않는 것은? Percentage Male Female Teachers UK, 2010 graph shows percentage male female teachers five educational settings UK 2010. Overall, percentage females larger males three five educational settings. Females took highest percentage nursery schools, lowest universities. situation nursery primary schools onesided, ninety percent teachers female. secondary schools, percentage gap male female teachers larger primary schools. However, difference percentage male female teachers colleges. Males showed highest percentage universities, percentage males seventy percent.
249	3	2019	9	1	1	25	다음 도표의 내용과 일치하지 않는 것은? Note: Respondents choose one three options listed location. Due rounding, percentages may sum 100%.The three pie charts show percentages American adults' responses survey conducted 2017. survey asked whether people allowed fly drones three locations: public parks, beaches, near people's homes. 44% respondents said people allowed fly drones public parks, 25% said people allowed asked people allowed fly drones beaches, 35% respondents said allowed 32% said half respondents said people allowed fly drones near people's homes Less 10% respondents said people allowed fly drones near people's homes three locations, proportion respondents chose "it depends" 30%

(b) 영어교과 클러스터1 - 그림, 사진, 도표, 대상, 주제

Cluster 1 - 연도별 상위 3문제:

	grade	yyyy	mm	host	subject_cat	question_num	combined_text
420	3	2020	6	1	1	19	다음 글에 드러난 Sharon의 심경 변화로 가장 적절한 것은? Sharon received ticket upcoming tangc concert friend. surfing Internet, came across review concert. reviewer harsh, calling "an awfu performance." raised Sharon's mind question whether worthwhile go, end, reluctantly decidec attend concert. hall located old town ancient run-down. Looking around, Sharon wondered kinc show could expect. soon tango started, everything changed. piano, guitar, flute, violin magically flew harmony. audience cheered. "Oh goodness! fantastic music!" Sharon shouted. rhythm tempc energetic sensational shook body soul. concert far beyond expectations. excited →bored doubtfu →amazed calm→upset ashamed →grateful envious →indifferent 2C
4807	1	2023	6	1	1	22	다음 글의 요지로 가장 적절한 것은? promise computerized society, told, would pass machines repetitive drudgery work, allowing us humans pursue higher purposes leisure time. didn't work way. Instead time, us less. Companies large small offloaded work onto backs consumers. Things used done us, part valueadded service working company, expected ourselves. air travel, we're expected complete reservations checkin, jobs used done airline employees travel agents. grocery store, we're expected bag groceries and, supermarkets, scan purchases.* drudgery: 고된 일 컴퓨터 기반 사회에서는 여가 시간이 더 늘어난다. 회사 업무의 전산화는 업무 능률을 향상시킨다. 컴퓨터화 된 사회에서 소비자는 더 많은 일을 하게 된다. 온라인 거래가 모든 소비자를를 만족시키기에는 한계가 있다. 산업의 발전으로 인해 기계가 인간의 일자리를 대신하고 있다
332	3	2020	10	2	1	20	다음 글에서 필자가 주장하는 바로 가장 적절한 것은? One funniest things becoming boss causes awful lot people forget everything know relate people. complaint somebody personal life, would never occur wait formally scheduled meeting tell them. Yet, management bureaucratized point throw away effective strategies everyday communication. Don't let formal processes like annua performance reviews take over. meant reinforce, substitute, every day. You'd never let fact go dentist cleaning couple times year prevent brushing teeth every day. 정확하고 구체적으로 직원들에게 피드백을 제공하라. 업무에 대한 동료의 건전한 비판을 겸허히 수용하라. 직원 결속을 위해 회사 내 비공식적 모임을 활성화하라. 직장에서 상사에게 이의를 제기할 때는 격식을 존중하라. 절차에만 의존하지 말고 부하 직원들과 일상적으로 소통하라. 정확하고 구체적으로 직원들에게 피드백을 제공하라

(c) 영어교과 클러스터2



(d) 영어교과 클러스터2의 세부 분류 시각화

	grade	yyyy	mm	host	subject_cat	question_num	combined_text
18	3	2019	10	2	1	19	다음 글에 드러난 'I'의 심경 변화로 가장 적절한 것은? Mary held hand made follow her. eyes blindfolded. wondering fantastic place taking me. stopped suddenly played alltime favorite song: Stars Go Blue. took deep, shaky breath. Mary pulled blindfold, jaw dropped gasped sight me. hill. city lights anywhere sight. things giving light moon stars. Mary took hand again. next thing knew dancing, staring other's eyes. wished night would last forever. anticipating →delighted anxious →frightened disappointed →satisfied ashamed →relaxed grateful →annoyed
19	3	2019	10	2	1	20	다음 글에서 필자가 주장하는 바로 가장 적절한 것은? human brain wired look threats — trait kept us alive living savannas prevent happiness modern lives. socalled "negativity bias" keep focused what's going wrong (which explains complaining popular pastime). break neural rut, train acknowledge things go right, keep calendar journal, make point write went well. you're verbal processor, start conversations friends sharing recent win (anything gives yesssss feeling). mind goes, reality follows. appreciate life, reasons celebrate it. * rut: 고정된 틀 삶의 긍정적인 면을 인식하도록 자신을 훈련하라 경쟁자의 장점을 칭찬하고 따라 배우려 노력하라 글쓰기를 통해 부정적인 감정을 배출하라 실패의 원인을 다양한 각도에서 분석하라 불만을 자기 혁신의 동력으로 삼으라
42	3	2019	10	2	1	43	(A) (B) incident, teacher invited church dinner Grace's mom attended, too. talking her, teacher happened remark, "I know Grace bright, I'm worried days. doesn't recite answer question class. can't understand it." Mom couldn't understand either. heard Grace reading book home, brother drilled sums (b) knew well. (C) Grace felt rather proud known Billy did. (c) pride didn't last long, however. brother, Justin, reported Mom happened. said, "Grace made Billy feel like fool today." Grace tossed head defiantly. "Well, know words, Billy didn't," said proudly. "Your brother right, Grace," said Mom. "You made Billy feel bad reading him. this, speak up, even (d) know answer." Grace nodded head. understood knew something, keep herself. (D) Mom approached subject suppertime, asking, "Grace, read lessons?" Grace said, "Sure, Mom. read whole book!" Mom puzzled. "Then why," asked, "does teacher say don't recite school?" Grace surprised. "Why, Mom," answered, "you told to!" Mom exclaimed, "Why, Grace, thing!" "Yes, (e) did," Grace said. "You told speak up, even knew answer." Mom remembered. matter soon straightened out, Grace recited class. [43 ~ 45] 다음 글을 읽고, 물음에 물음에 답하시오. 주어진 글 (A)에 이어질 내용을 순서에 맞게 배열한 것으로 가장 적절한 것은? (B) - (D) - (C) (C) - (B) - (D) (C) - (D) - (B) (D) - (B) - (C) (D) - (C) - (B)

(e) 영어교과 클러스터2의 세부 분류1

- 일이나 사건의 순서/전후 관계/원인/결과, 필자의 의도/목적, 필자의 심경/태도

Cluster 2 - Sub-cluster 1 - 연도별 상위 3문제:

	grade	yyyy	mm	host	subject_cat	question_num	combined_text
30	3	2019	10	2	1	31	[31 ~ 34] 다음 빈칸에 들어갈 말로 가장 적절한 것을 물음에 답하시오. 31 developmental control children certain serious medical problems exert physical activity relevant . example, infant crib cognitively intact 14-yearold confined bed due illness injury may relatively inactive. adolescent can, however, expected awareness control movements rolling might dislodge otherwise impair functioning medical device breathing tube feeding tube. Likewise, 5yearold 25yearold cardiac pacemaker implanted may know need protect device. developmental differences understanding risk causation control impulses increase probability risky behavior child, example, jumping porch.* dislodge: 떼어 내다 ** cardiac pacemaker: 심박 조율기 device safety mental health pain reactions athletic training medical diagnoses
31	3	2019	10	2	1	32	[31 ~ 34] 다음 빈칸에 들어갈 말로 가장 적절한 것을 물음에 답하시오. 32 There's striving majority one's group merely acquiring power. work majority groups majority controls material psychological resources, also largely defined claim us own. Drawing distinctions who's who's out, who's right who's wrong, privileged disadvantaged — short, us — motivates us . seek belong majority group, even group minority. majority holds power, privilege attached majority position commonly viewed others deserved. coming, perception contributes sense worth, are, others' assessments value well. [3점] speak put silence empower powerless political processes counted among counting value inner self appearance take outsiders fashionable rule breakers
32	3	2019	10	2	1	33	[31 ~ 34] 다음 빈칸에 들어갈 말로 가장 적절한 것을 물음에 답하시오. Eating original science, original study environment. Kids, like primitive lifeforms, learn reality putting mouths. mouth knowledge knows abstracts. world either sweet bitter, smooth prickly, pleasant unpleasant. Mouth knowledge comes gutlevel certainty. eat literally know. know what? know self nonself. Mouth knowledge taught us boundaries bodies. When, babies, sucked object, pacifier, felt one side, side mouth. sucked thumbs, felt outside, mouth. Inside, feeling thumb sucked on. mouth knowledge —unlike later school knowledge — gave us glimpse paradoxical nature: somehow .. * pacifier: (유아용) 고무 젖꼭지 ignorant things remain confident gain pleasure serve people find unpleasant situations pleasant children attracted things go intuition subject object experience

(f) 영어교과 클러스터2의 세부 분류2 - 빈칸에 들어갈 문장/단어 찾기

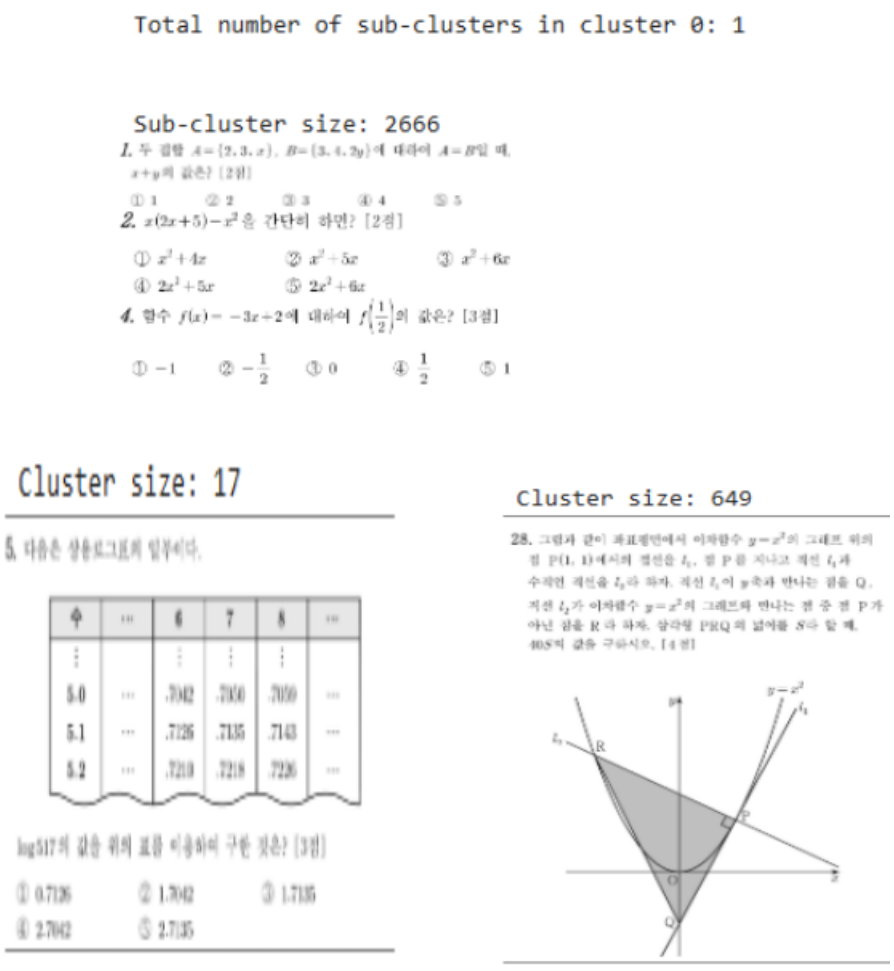
Cluster 2 - Sub-cluster 2 - 연도별 상위 3문제:

	grade	yyyy	mm	host	subject_cat	question_num	
17	3	2019	10	2	1	18	다음 글의 목적으로 가장 적절한 것은? would like thank suggestion switching new ABC software maintaining company's database s attend meeting technical team 2 p.m. October 8th Meeting Room A. assessing feasibility proposal, would like proceed implementatior
20	3	2019	10	2	1	21	밑줄 친 put proverbial cart horse가 다음 글에서 의미하는 바로 가장 적절한 것은? people try slow put proverbial cart horse. make drama life quickly discover slowerpaced life country drives crazy. habitual, hectic thinking won't allow adjust superficial changes make. Sec person city, you'll also hurried, rushed person country. mend problem, slow life inside out.* temperamentally: 기질적으로볼 reflect lookin
21	3	2019	10	2	1	22	다음 글의 요지로 가장 적절한 것은? tend think technology shiny tools gadgets. Even acknowledge technology exist disembodied form, (the computer code web page), thousand lines letters English (Hamlet) must qualify well. change behavior, alter course events, enable fu can't separate multiple overlapping technologies responsible Lord Rings movie. literary rendering original novel much invention digital r 형의 기술은 예술적 상상력을 구현할 수단을 제공한다 상상력을 발휘하여 물리적인 한계를 극복할 수 있다 그것은 현대 사회에서 새

(g) 영어교과 클러스터2의 세부 분류3

- 문맥 속 낱말/어구/문장의 의미, 글의 숨겨진 의미, 줄거리/주제/요지

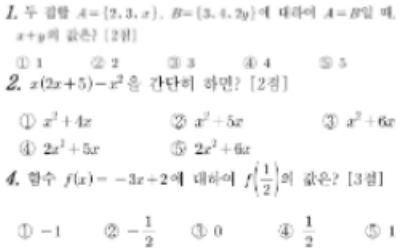
이미지 클러스터링의 경우 기존 수학공식 파싱과정에서 사용한 OpenAI의 프롬프트 엔지니어링에서 착안하여 OpenAI의 CLIP 모델의 변형인 'clip-ViT-B32'을 사용하였다. 이미지를 임베딩한 후 커뮤니티 감지(Community Detection) 기법을 통해 군집분석을 진행하였다. 첫번째 군집분석을 시도했을 때 총 7개의 군집이 나왔으나 실제 데이터를 확인하였을 때 도형이나 확률과 통계, 기하와 같이 그림이 들어간 문제들은 군집이 잘 나뉘었으나 문제에 그림이 없는 짧은 문제의 경우 나뉘지 않고 하나의 군집으로 묶이는 모습을 볼 수 있었다. 이를 해결하기 위해 두번째 시도로 해당 군집을 대상으로 한 번 더 군집분석을 실시하였으나 군집이 나뉘지 않고 하나의 군집으로 나오는 것을 확인할 수 있었다.



(a) 수학교과 클러스터 – 첫번째 시도

Total number of sub-clusters in cluster 0: 1

Sub-cluster size: 2666


1. 두 집합 $A=\{2,3,x\}$, $B=\{3,4,2y\}$에 대하여 $A=B$일 때, $x+y$의 값은? [2점]

① 1 ② 2 ③ 3 ④ 4 ⑤ 5

2. $x(2x+5)-x^2$을 간단히 하면? [2점]

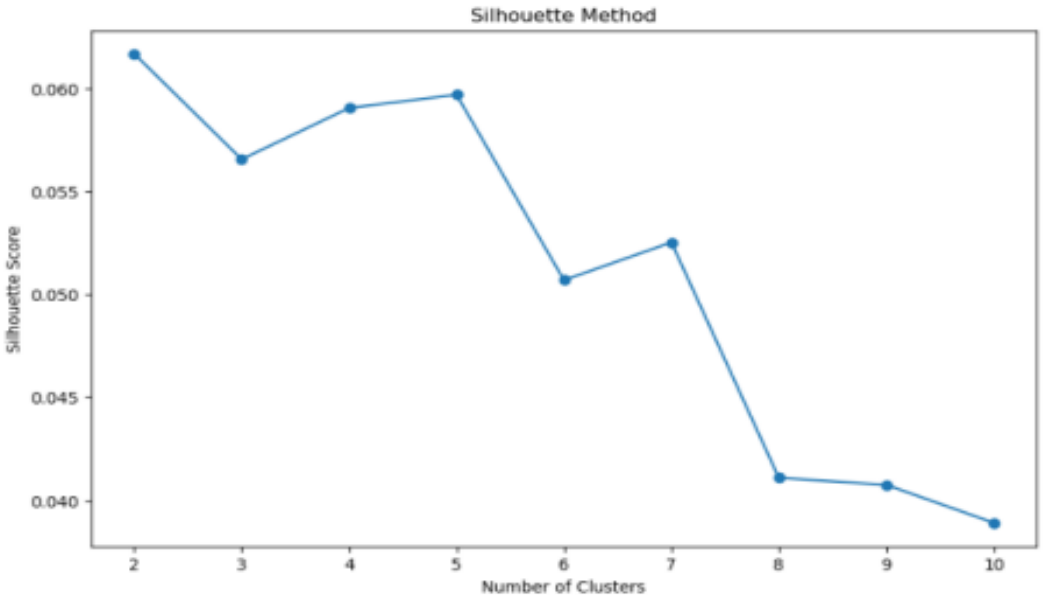
① x^2+4x ② x^2+5x ③ x^2+6x
④ $2x^2+5x$ ⑤ $2x^2+6x$

4. 함수 $f(x)=-3x+2$에 대하여 $f\left(\frac{1}{2}\right)$의 값은? [3점]

① -1 ② $-\frac{1}{2}$ ③ 0 ④ $\frac{1}{2}$ ⑤ 1

(b) 수학교과 클러스터 – 두번째 시도

이미지 클러스터링에서 결과가 잘 나오지 않아 텍스트 클러스터링을 실시하였다. 텍스트 클러스터링의 경우 BERT모델의 변형인 'jhgan/ko-sbert-sts'을 사용하여 텍스트를 임베딩 한 후 K-mean클러스터링을 사용하여 군집분석을 진행하였다. 최적군집은 실루엣계수를 사용하여 8인 것을 확인하였으나 군집들을 확인해본 결과 몇몇의 군집에서 통계, 기하 등 특수한 기호나 한글이 많은 문제들은 잘 분류하나 수식기호나 함수, 도형(sin, cos)등은 잘 분류하지 못하고 여러 군집 내에 섞여있는 모습을 볼 수 있었다. 이러한 문제를 해결하기 위해 텍스트 임베딩을 진행하는 컬럼을(question, choice1 ,choice2 ,choice3 ,choice4 ,choice5) ->(question)만 사용하여 군집분석을 진행하였으나 결과는 동일하였다. 추가적으로 개정교육과정이 2017년 전 후로 2009와 2015 개정교육과정으로 변화가 이루어졌고 고등학교 1학년과 고등학교 2, 3학년의 과목분류가 다르다는 점을 이용하여 타겟층을 나누어 군집분석을 진행하였으나 비슷한 수준으로 군집이 분석되어 최종적으로 현재 상황에서는 수학교과의 자동태깅을 실시하기 어렵다는 결론에 도달하였다. 다만, 일부 클러스터는 분리가 잘 이루어졌다는 점을 미루어 보아 추후 다른 모델을 더 다양하게 사용하여 임베딩을 시도하거나 비지도 뿐만 아니라 지도학습을 실시한다면 나은 성능을 나타낼 것으로 기대할 수 있다고 보여진다.



(c) 텍스트 데이터를 활용한 수학교과 클러스터 – 실루엣계수

Cluster 0 - Top Questions per Year:

	grade	yyyy	mm	host	subject_cat	question_num	combined_text
2737	2	2018	3	2	2	16	이 식당에서 두 메뉴 A, B를 합하여 하루 최대 35인분을 만들 수 있을 때, 하...
2747	2	2018	3	2	2	26	티켓 1500장을 모두 판매한 금액이 6000만 원이 되도록 하기 위해 판매해야 할...
2765	2	2018	3	2	3	14	2018 평창 동계 올림픽 대회 및 동계 패럴림픽 대회 자원봉사 활동 중에서 하나만...
9	3	2019	10	2	2	10	한 개의 주사위와 6개의 동전을 동시에 던질 때, 주사위를 던져서 나온 눈의 수와 ...
12	3	2019	10	2	2	13	어느 도시의 시민 한 명이 1년 동안 병원을 이용한 횟수는 평균이 14, 표준편차가...

Cluster 2 - Top Questions per Year:

	grade	yyyy	mm	host	subject_cat	question_num	combined_text
2667	2	2018	11	0	2	6	공비가 3인 등비수열 ({a_n })이 [lim_{{n to infty}} l...
2668	2	2018	11	0	2	7	(int_{ 1}^{1} left(4x^3 + x^2 frac{1}{2}x + a...
2670	2	2018	11	0	2	9	부등식 (2 log_{ frac{1}{2}} (x 2) < log_{2} (3x +...
0	3	2019	10	2	2	1	두 벡터 (vec{a} (1, 2)), (vec{b} (2, 5))에 대하...

(d) 텍스트 데이터를 활용한 수학교과 클러스터 결과

자동태깅의 경우 시간의 부족으로 인해 웹 상에 자동태깅 결과를 띄우지 못하여 DB에 자동태깅 결과가 저장되는 방식으로 진행하였다.

(4) DB

프로젝트의 초기 단계에서는 과목별로 2개의 테이블(문항 테이블, 풀이이력 테이블)을 정의했다. 하지만, 처음 정의한 풀이이력 테이블에는 사용자가 풀었던 문 제의 정보가 포함되지 않았음을 파악하고 한 회차마다 사용자가 푼 문제에 대한 정보가 주어진 랜덤 문제 메타 정보 테이블을 추가했다. 이후 프로젝트가 진행되면서 전체 문제에 대한 유사도 쌍 테이블과 사용자의 외부 문제에 대한 유사도 쌍 테이블이 추가되었다. 프로젝트 진행 과정 중에 프로젝트의 데이터베이스와 검색 엔진 설정을 위해 Docker를 사용했다. Docker는 환경 설정 문제를 해결해주며, 애플리케이션을 신속하게 배포하고 관리할 수 있는 컨테이너화 기술이다. Docker에 MariaDB와 Elasticsearch & Kibana 이미지를 받아서 실행했다. 여기에서, Mariadb와 Elasticsearch를 선택한 이유는 다음과 같다. MariaDB은 MySQL 기반으로 만들어진 RDBMS로, 전반적인 사용법은 MySQL과 유사하며, 동일한 하드웨어 사양으로 MySQL보다 향상된 성능을 제공하고 다양한 기능을 제공하는 이점을 가지는 점에서 MariaDB를 선택하였다. Elasticsearch는 오픈소스 검색엔진 솔루션으로, 매우 빠른 속도와 확장성, 복원성뿐만 아니라 정형/비정형 데이터를 모두 수용할 수 있는 유연성을 가지고 있는 이점을 가지는 점에서 Elasticsearch를 선택하였다.

그러나 Docker에 Mariadb와 Elasticsearch&Kibana 이미지 받아서 실행하는 과정에서 다음과 같은 문제점이 발생했다. 첫째, Elasticsearch와 Kibana가 싱글 노드로 동작하지 않는 문제가 발생했다. 각각의 이미지를 받아 실행했으나, 두 서비스 간의 통신 문제가 발생했다. 이 문제를 해결하기 위해 Docker-compose를 사용하 여 Elasticsearch 클러스터를 구성하였다. 둘째, Docker-compose를 이용한 Elasticsearch 클러스터 구성 문제: Docker-compose 파일을 작성하여 Elasticsearch에 3개의 노드를 구성하고 실행했으나, 3개의 Elasticsearch 노드가 차례대로 멈추는 상황이 발생하여 클러스터가 정상적으로 작동하지 않은 문제가

	발견되었다. 이를 해결하기 위해, 다양한 시도를 해보았다. 우선, Elasticsearch와 Kibana 이미지 버전을 8.x.x에서 7.x.x로 낮추었다. 다음으로, max_map_count 설정을 확인하여 메모리를 늘려주었다. 'max_map_count'는 Elasticsearch가 사용하는 가상 메모리 영역의 최대 개수를 지정하는 시스템 설정이다. Elasticsearch를 실행할 때, 이 값이 충분히 높지 않으면 메모리 맵핑이 제한되어 성능에 문제가 생기 거나 정상적으로 실행되지 않을 수 있기 때문에 이 설정을 통해 Elasticsearch가 충분한 메모리 맵핑을 사용할 수 있도록 보장해야 함을 깨닫게 되었다. 위의 과정을 거치며 포트 충돌 문제도 함께 발생하여 Elasticsearch와 Kibana의 포트가 각각 2개씩 잡혀서 포트 충돌 문제가 발생했다. 충돌난 포트를 확인하여 제거해 주는 작업을 하였다. 이와 같은 과정을 거쳐 MariaDB와 Elasticsearch & Kibana를 성공적으로 실행하고, 데이터 적재 및 분석 환경을 구축했다. 다음으로, 프로젝트의 데이터베이스 관리를 위해 DBeaver를 사용하여 MariaDB와 연결하였다. DBeaver를 통해 테이블 정의서에 맞춰 과목별 테이블을 손쉽게 생성하고, 데이터의 스키마를 시각적으로 확인하며 데이터베이스 구조를 효율적으로 관리할 수 있었다. 실제 데이터 수집 및 전처리 과정에서 AWS S3에서 이미지 데이터(.png)와 텍스트 데이터(.json)를 받아와서 합치는 작업을 수행하였다. 이후, 텍스트 데이터로만 이루어진 데이터프레임은 MariaDB에 적재하였고, 이미지와 텍스트가 합쳐진 데 이터에서 임베딩 값을 추출하여 Elasticsearch에 적재하였다. 이때, DB에 적재하기 위해 다음과 같이 데이터 전처리 작업을 수행하였다. 1. 고유컬럼(pk) 생성: grade, yyyy, mm, subject_cat, question_num 컬럼을 합 쳐 G320240601Q1과 같은 형태로 생성하였다. 2. 타입 변경: 타입이 float형인 컬럼을 int형으로 바꿔주었다. 3. 값 변경: subject_cat(과목 세분류) 컬럼을 올바르게 바꿔주었다. 4. 중복값 제거 및 빈 값을 NULL값 또는 빈 문자열로 채워 각 DB에 잘 들어

	가도록 설정하였다. 이미지 임베딩 값과 텍스트 임베딩 값을 추출하여 Elasticsearch에 적재하였다. 이때, 이미지 임베딩 모델은 Open AI에서 제공하는 CLIP 모델을 사용하였다. CLIP 모델은 이미지와 텍스트를 동시에 임베딩할 수 있는 모델로, 두 가지 다른 형식의 데이터를 동일한 벡터 공간에서 표현할 수 있다. 이를 통해 이미지와 텍스트간의 유사도 분석이 가능하며, 정확한 유사도 계산을 할 수 있다. 위와 같은 모델을 사용하여 나온 임베딩 결과를 포함하여 Elasticsearch에 적재한 후, 다음과 같은 문제가 발생하였다. img_vec와 text_vec 컬럼의 타입을 densor_type으로 지정했으나, Elasticsearch에서 type을 확인해본 결과 float 타입으로 들어간 문제가 발생하였다. 첫번째로 해결한 방법은 float 타입의 컬럼을 densor_vector 타입으로 reindex한 방법이다. 두번째로 해결한 방법은 Elasticsearch 라이브러리를 최신 버전(8.x.x)으로 업그레이드한 방법이다. 두번째 방법은 기존에 Elasticsearch 라이브러리 버전을 7.x.x로 한 결과 타입이 제대로 반영하지 못한 이유이다. Elasticsearch에서 제공하는 코사인 유사도 모델을 사용하여 모든 문제에 대해 유사도를 계산했다. 각 문제에 대해 유사도가 높은 20개의 문제를 구하고, 이 20개의 문제에 대해서도 다시 유사도를 계산했다. 위 과정의 MariaDB 테이블 간 ERD는 다음과 같다.

(a) MariaDB 테이블 간 ERD

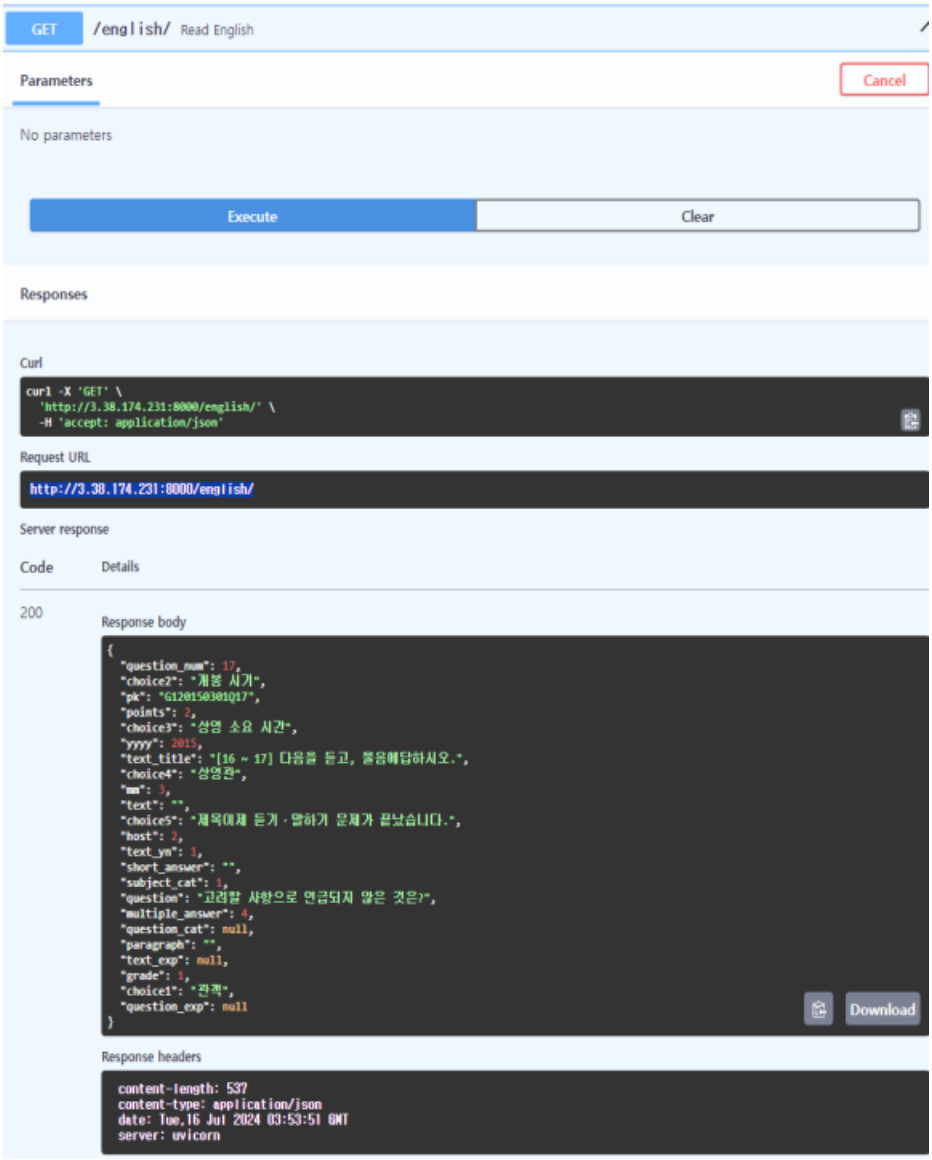
(5) 웹

AWS환경에서 FastAPI 웹 서버 사용을 비롯한 프로젝트 과정 중의 작업을 하기 위해 AWS환경을 구축하였다. AWS에서 VPC를 생성하여 클라우드 내 논리적으로 격리된 네트워크를 구축하였다.

(a) public, private 서브넷 생성, 각각 라우팅 테이블 생성해 연결

	생성한 VPC 내에 EC2 인스턴스를 연결하여 가상 서버를 구성하였으며, EC2 인스턴스를 통해 FastAPI 웹 서버를 호스팅하고 데이터 처리를 수행하고자 하였다. FastAPI는 비동기 처리를 지원하여 대량의 데이터 요청을 처리해야 하는 본 프로젝트에 적합하다고 여겼으며, 프로젝트 내에서 사용되는 Elasticsearch, MariaDB, S3의 여러 데이터베이스와 통합이 요구되는 상황에서 다양한 벡엔드 데이터베이스와의 통합이 용이하다는 특징을 가진 FastAPI가 가장 적합하다고 여겼다. Amazon S3버킷과의 원활한 통신을 위해 S3 엔드포인트를 생성한 후 기존에 생성된 VPC애 연결하였다. 이 과정에서 정책 편집 권한이 없어 담당자님께 권한 요청을 통해 정책 편집을 실행하였다. 이를 통해 VPC 내의 리소스가 인터넷을 경유하지 않고 직접 S3버킷과 통신할 수 있게 되어 데이터 전송의 효율성과 보안성을 높였다. 이는 대량의 이미지, pdf, json 데이터를 전송해야 하는 본 프로젝트의 특성에서 높은 중요성을 가지고 있다. 인스턴스 생성 과정에서 프로젝트 진행을 위한 인스턴스 유형이 어떤 것이 적합한 것인지 혼동이 발생하였다. 멘토님께 자문을 구한 결과 최종 m5.xlarge로 인스턴스를 설정하였다. 구축한 서버를 도커로 빌드하여 컨테이너화하였다. 이 때, 인스턴스 환경에서 아나콘다를 설치하여 가상환경을 만드는 것도 고려하였지만 설치 과정이 복잡하고 오랜 시간이 소요될 것으로 예상되어 파이썬 도커 이미지를 받아 빌드하였다. 도커를 사용하여 애플리케이션을 일관된 환경에서 실행할 수 있게 했으며, 이를 통해 배포와 관리가 용이하도록 하였다. 도커 이미지를 생성하고 이를 바탕으로 컨테이너를 실행하여 서버 환경을 구축함으로써 프로젝트 기간 중 이루어지는 개발과 실제 배포 후 운영 환경의 일관성을 유지하고자 하였다. pc로컬 vscode에 ssh로 인스턴스를 연결하여 개발을 진행하고자 하였으나 ssh연결 과정에서 일부 pc 환경에서 ssh 연결이 안되는 문제가 발생하였다. 여러 원인들을 찾아본 결과 인스턴스 연결 과정에서 필요한 pem키의 권한 문제인

	것을 확인하여 동인한 문제를 겪었던 블로그[xiii]를 참고하여 문제를 해결하였다. 빌드한 도커 컨테이너 내에 FastAPI 웹 서버를 구축하여 이후 데이터 전처리, 임시저장, 임베딩 및 유사도 추출 등의 기능을 담당하도록 하였다. 모델 개발 및 파이프라인 구축을 위해 sagemaker 도메인을 생성하였다. 빠른 설정으로 도메인을 생성하면 vpc지정이 불가능하여 직접 세팅을 해주었다. 세팅 과정은 다음과 같다. 1. 기본 ML활동 선택지에 manage pipelines, manage model monitoring 추가 선택. 2. 네트워크 설정 지정 시 vpc전용으로 했더니 sagemaker에서 인터넷 연결이 안되는 문제 발생. 3. 퍼블릭 인터넷 액세스로 설정 변경하여 해결. Sagemaker에서는 인스턴스와 동일 사양으로 jupyterlab space 생성해 진행하였고, 한 space에서 팀원들이 동시에 작업할 경우 충돌 문제가 발생하여 space를 여러 개 생성하여 진행하였으나 비용문제로 인해 사용하지 않는 경우에는 space를 종료하는 것으로 하였다. DB구축 후 FastAPI와 DB를 연결하는 작업을 하였다. 우선 MariaDB-FastAPI를 연결하였다.



(b) MariaDB-FastAPI 연결

이후 ElasticSearch-FastAPI를 연결하였다. 참고자료[xiv]를 활용하여 ElasticSearch query 이용해 유사도 분석을 진행하였다. 각각의 임베딩 값을 단독으로 사용해 유사도 분석을 진행하는 경우 이미지가 전반적으로 좀 더 높은 유사도를 보였다. 최종적으로 구축된 서버에서 인스턴스 연결 문제가 종종 발생하는 것을 발견하였다. 개발 중 메모리 과부하 등으로 서버 접근이 불가능한 경우가 발생하였고 이를 해결하기 위해 재부팅을 할 수 밖에 없었다. 그러나 인스턴스 재부팅 후에는 모든 도커 컨테이너를 restart 해줘야 하는 시간적 소요가

발생하였다.

UI 구성에 있어서는 웹 접속 시 유사문항을 찾고 싶은 문제 이미지를 업로드 할수 있도록 하였다. 유사문항 검색 성능을 높이기 위해 과목 필터링을 추가하였고 우측 상단 홈버튼을 누르면 메인페이지로 돌아올 수 있게 하였다. 이 때, 홈버튼을 누를 경우 웹 출력을 위해 서버에 저장된 이미지를 삭제해 서버 부하를 줄였다. 또한, 데이터 버전관리를 위해 문제지를 pdf파일로, 해당하는 답안지를 이미지 파일로 업로드할 수 있는 폼을 생성하였다. 과목, 학년, 월, 영역을 선택하고 년도를 입력하면 그에 맞춰 파일명이 자동적으로 생성되도록 하였다. 업로드한 파일은 임시저장 후 데이터베이스에 업로드하였고 업로드된 파일은 임시저장소에서 삭제되도록 하였다.

7. 산출물

문항 추천 시스템의 산출물은 크게 버전 관리를 위한 기능과 사용자가 이용을 위한 기능으로 구성되어 있다.

(1) 데이터 버전 관리 기능

본 프로젝트의 데이터셋은 고등학교 1, 2, 3학년 학생들이 치룬 수능 및 모의고사 문항들로 구성되어 있다. 그렇기에 월, 년도마다 새로운 데이터가 업데이트되고 있기에 이를 기존 데이터셋에 반영해주어야 한다고 보았다. FastAPI를 통해서 pdf 형식의 문제지와 이미지 형식의 답안지를 넣으면 이미지 crop, 파싱의 전처리 과정을 거친 후 최종적으로 S3에 png형식의 문항들과 답안지, json형식으로 파싱된 문항들이 저장된다. 이 때, 텍스트 데이터는 S3에서 문항 정보 테이블로 MariaDB에 저장된다. S3에서 가져온 이미지와 텍스트 데이터에서 기본 정보를 추출하고 임베딩 모델을 통해 임베딩 벡터값을 산출하여 Elasticsearch에 저장한다. 마지막으로, Elasticsearch에서 적재된 데이터를 대상으로 코사인 유사도를 실시하여 유사도의 벡터값을 계산한다. 이는

추후 외부문제와의 비교를 통해 유사 문항 추천을 위해 사용된다. 이를 정리하면 다음과 같다.

1. FastApI를 통해 PDF문제지와 이미지 답안지를 업로드.
2. FastApI에서 외부 OpenAI API를 호출하여 이미지 crop 및 파싱 전처리를 수행.
3. 전처리된 데이터를 S3에 저장(PNG 문제 및 답안지, JSON 형식의 파싱 데이터)
4. 텍스트 데이터는 s3에서 가져와 문항의 기본 정보 테이블로 MariaDB에 저장.
5. S3에서 가져온 이미지와 텍스트 데이터에서 기본 정보 추출 및 병합
6. 임베딩 모델을 통해 텍스트 임베딩 벡터값 산출
7. 산출된 임베딩 벡터값과 문항 기본 정보를 Elasticsearch에 저장
8. Elasticsearch에서 적재된 데이터를 대상으로 코사인 유사도를 실시하여 유사도의 벡터값을 계산 및 MariaDB에 저장. 추후 외부 문제와의 비교를 통해 유사 문항 추천에 사용.

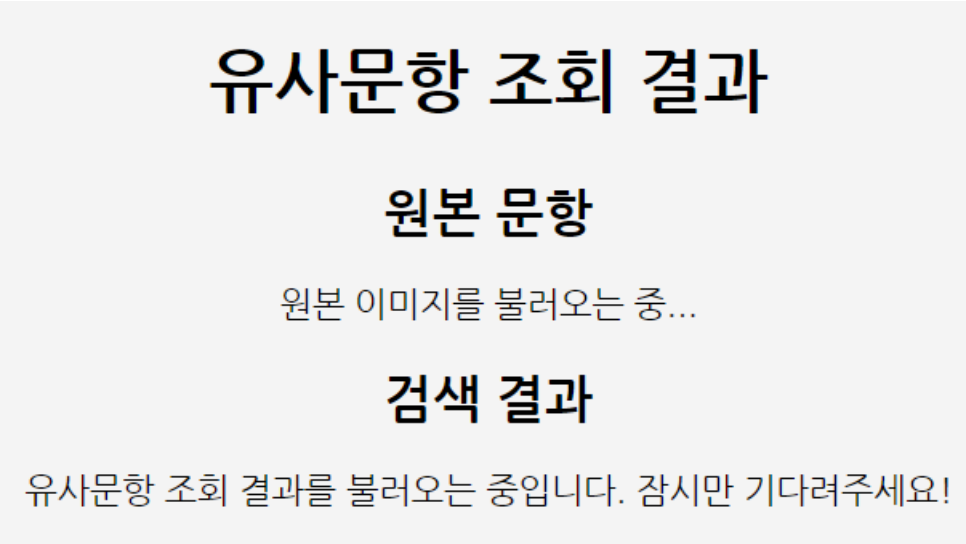
(2) 사용자 이용

실제 사용자가 이미지 파일을 웹에 업로드하면 FastAPI에서 해당 이미지에 대한 crop 및 파싱의 전처리를 진행한다. 전처리가 진행된 이미지와 텍스트 데이터는 S3에 적재된다. S3에 적재된 이미지와 텍스트 데이터를 웹상에서 불러와 임베딩 모델을 통해 텍스트 임베딩을 진행한 후 코사인 유사도를 통해 임베딩 벡터값과 유사도 벡터값을 산출한다. 이렇게 산출된 벡터값은 Elasticsearch에 저장되어 기존 데이터셋의 유사도 벡터값과의 비교를 통해 유사 문항을 상위 20개 산출한다. 이 때 과목에 따라 유사도 검증의 순서를 다르게 구성하였다. 국어와 수학의 경우 텍스트 유사도를 통해 상위 20개의 문항을 추출한 후 추출된 문항들을 대상으로 다시 이미지 유사도를 비교하여

	최종적으로 상위 5개의 문항을 유사 문항으로 추천하게 되어 웹 상에 추천 문항들이 나타나게 된다. 영어의 경우 이미지 유사도를 통해 상위 20개의 문항을 추출한 후 추출된 문항들을 대상으로 텍스트 유사도를 비교하여 최종적으로 상위 5개 문항을 유사 문항으로 추천하는 방식을 사용하였다. 이를 정리하면 다음과 같다. 1. 사용자가 웹에 이미지 파일 업로드. 2. FastAPI에서 업로드된 이미지에 대해 crop 및 파싱 전처리 진행. 3. 전처리된 이미지와 텍스트 데이터를 S3에 저장. 4. S3에 적재된 이미지와 텍스트 데이터를 웹상에 불러온 후 임베딩을 수행하여 임베딩 벡터값 산출, 코사인 유사도를 통해 유사도 벡터값 산출. 5. 산출된 벡터값들을 Elasticsearch에 저장. 6. 과목에 따라 유사도 검증 순서를 다르게 구성한 것을 바탕으로 기존 데이터셋과 비교하여 유사 문항 상위 5개 산출. 다음은 본 프로젝트 웹페이지에 대한 설명이다. 웹페이지는 메인 페이지와 문항 추천 페이지로 나뉜다. 메인 페이지에서는 사용자가 문제 이미지를 넣으면 유사도를 기반으로 문항을 추천해주는 문항추천시스템과 데이터 버전관리를 위한 PDF 업로드 시스템으로 구성되어 있다.

(a) 메인페이지 구성

문항추천시스템의 경우 사용자가 입력한 원본문항과 비교하여 유사도가 높은 상위 5개의 이미지가 산출되도록 UI를 구성하였다. 문항 추천과정에서 시간 소요시 로딩 화면이 나타나도록 설정되었다.



(b) 로딩페이지 구성

로딩페이지 후 유사문항 조회가 완료되면 다음과 같이 조회 결과가 나타난다. 조회 결과는 사용자가 넣은 이미지의 원본 문항과 유사도가 높은 순으로 유사문항을 5개 추천해주는 방식으로 이루어져 있다.

(c) 유사문항 조회 결과 – 원본 문항

검색 결과

2020년도 1학년 국어 11월 공통 12번 0.90447

12. ㉠, ㉡에 해당하는 예로 적절하지 않은 것은? [3점]

① ㉠: 관객이 많으니 미리 줄을 서라.
㉡: 돌아오는 기차표는 네 것만 끊어라.

② ㉠: 눈을 떠 보니 다음날 아침이었다.
㉡: 네가 집에 빨리 가서 아쉬웠다.

③ ㉠: 체육 시간에는 교실 불을 꺼 두자.
㉡: 오늘은 새 신발을 신고 학교에 가자.

④ ㉠: 지금 마는 김밥은 어머니께 드릴 점심이다.
㉡: 독서로 쌓은 지식은 삶의 자양분이 될 것이다.

⑤ ㉠: 아버지 대신 빨래를 너는 모습이 보기 좋다.
㉡: 가을빛을 담고 있는 감나무 열매를 본다.

(d) 유사문항 조회 결과 – 추천 문항1

2019년도 3학년 국어 03월 공통 11번 0.87131

11. ㉠~㉤에 대한 이해로 적절한 것은?

① '한몫[한목]'을 발음할 때, ㉠이 일어난다.
② '놓기[노키]'를 발음할 때, ㉤이 일어난다.
③ '끊지[끌치]'를 발음할 때, ㉡과 ㉢이 일어난다.
④ '값할[가팔]'을 발음할 때, ㉡과 ㉤이 일어난다.
⑤ '맞힌[마친]'을 발음할 때, ㉢과 ㉤이 일어난다.

(e) 유사문항 조회 결과 – 추천 문항2

데이터 버전 관리용 웹 기능의 경우 버전관리에 해당되는 수능 및 모의고사의 문제지 pdf 파일과 답안지 이미지(png, jpg, jpeg) 파일을 넣고 파일의 정보에 맞는 과목, 학년, 월, 영역을 선택하고 년도를 입력하면 일정 로딩이 지난 후

작업이 완료된다.

8. 개선점

1) 텍스트 파싱할 당시 ocr기반으로 파싱 작업을 시작하고 추후에 프롬프트 엔지니어링을 통해 텍스트 파싱이 가능하다는 것을 깨닫고 수학 이미지를 파싱하는 작업을 수행하였는데 논문, 스택오버플로우, 깃허브를 찾아서 보다 양한 방법을 시도해본 후 파싱 작업에 들어갔다면 프롬프트 엔지니어링을 활용하여 시간손실을 줄일 수 있거나 다른 수학공식 파싱 방법을 발견하여 비용 손실을 줄일 수 있었을 것이다.

2) 웹 상에서 사용자가 결과에 만족하지 못할 경우 다른 문제들도 보이게 하거나 자동 채점 기능을 구현하지 못한 점이 개선점이라고 생각한다. 시간의 부족으로 인해 구현하지 못한 부분이기에 추후 기회가 된다면 디벨롭하고자 한다.

3) 문항 분류 결과 기반으로 유사문제 추천/생성을 기존에 개발 목표로 잡았으나 생성의 경우 시간 및 인력 부족으로 인해 자동태깅 및 유사도를 정교화하는 방향으로 가게 되었다. 특히, AWS SageMaker를 통해 배포 및 관리할 수 있는 모델 중 gpt4o, LLama 등의 모델을 사용한다면 문항 생성도 가능할 것이라 보여진다.

4) 현재는 과목 분류만으로 필터링을 하고 있으나 태깅 정보를 활용하여 배점별(난이도 별), 학년 별로 필터링에 해당되는 문제만 볼 수 있게 세부적인 필터링 기능을 추가한다면 더 사용자 이용 시 용이한 서비스를 제공할 수 있을 것이라고 예상된다.

5) 자동 태깅 과목 중 수학 과목에 대해서는 온전한 자동태깅을 완료하지 못하였다. 현재는 비지도학습 방법만을 사용했기에 지도학습을 사용하고 모델을

	다양화하여 사용한다면 수학 과목에 대해서도자동태깅이 가능할 것으로 예상된다. **9. 소스코드** **깃허브 링크 참조: https://github.com/choijouneun/bigdata7-final_project-**
별첨	**1. AWS S3에서 데이터를 받아와 DB에 데이터를 적재하고 전처리 후 다시 S3에 올리는 과정에서의 소요 시간 및 결과 값.** 걸린 시간 / 국어 / 수학 / 영어 이미지 데이터 개수: 5394 / 4824 / 5130 이미지 임베딩: 1019.3782992362976 / 739.4697897434235 / 884.499279499054 텍스트 데이터 개수: 5333 / 4821 / 4952 (중복값 제거 후) 텍스트 임베딩: 871.4836752414703 / 758.7637872695923 / 4057.647351026535 총 임베딩: 1890.8975539207458 / 1498.2710933685303 / 4942.214681148529 전체 실행시간: 2603.3409831523895 / 2220.322898387909 / 5604.096158981323 **2. 소요 비용 산정** - AWS EC2: 75USD * 1 개월 - AWS S3: 6USD * 1 개월 - AWS SAGEMAKER: 272.16USD & 1 개월 - OpenAI: gpt4o 144.51USD * 1 개월 ➔ 총 프로젝트 1개월 동안 소요 비용: 500USD

	3. 테이블 정의서 - 별도로 첨부된 Excel 파일 참고

10. 참고문헌

[i] 한국경제) "내년 도입될 AI 디지털 교과서 선점하라"
https://n.news.naver.com/article/015/0004992468?sid=103

[ii] 피앤피뉴스) 킬러문항 없앤 첫 수능...적정 난이도로 변별력 갖췄다

https://www.gosiweek.com/article/1065582631806593

[iii] 내일신문) 비상교육 태블릿PC 전용 수능 학습 앱 '기출탭탭' 활용법
https://www.naeil.com/news/read/455315

[iv] 에듀동아) 프리윌린, '2024 인공지능 학습 플랫폼 매칭데이'에서 매쓰플랫과 풀리수학 선보여
http://m.edu.donga.com/news/view.php?at_no=20240223113015145830

[v] 에듀동아) 프리윌린, 학교 맞춤형 에듀테크 서비스 '스쿨플랫' 오픈...AI 기술로 교실에 '초개인화 교육' '학 습 격차 해소' 지원
http://m.edu.donga.com/news/view.php?at_no=20240516151302535017

[vi] 뉴스핌) 아티피셜소사이어티, 서울특별시교육청에 교육 콘텐츠 AI 솔루션 '젠큐' 공급
https://www.newspim.com/news/view/20240118000068

[vii] 전자신문) 비상교육 '비바샘', AI 기반 수학 문항 자동 생성 서비스
https://n.news.naver.com/article/030/0003078902?sid=102

[viii] QuestionWell 홈페이지 https://www.questionwell.org/

[ix] 천재교과서 지니아튜터, 서울시교육청 주최 에듀테크 교원연수 참가
https://www.it-b.co.kr/news/articleView.html?idxno=76613

[x] 천재교과서, 최신형 AI 엔진 탑재 수학문제은행 '닥터매쓰2.0' 그랜드 오픈
https://www.it-b.co.kr/news/articleView.html?idxno=69518

[xi] https://kagus2.tistory.com/50 문제지 header, body영역 분리 참고 블로그

[xii] https://iagreebut.tistory.com/74 각 문항 컨투어영역 참고 블로그

[xiii] https://lovflag.tistory.com/17 ssh 연결 오류 해결 참고 블로그

[xiv] https://medium.com/@pritam7798sonawane/building-a-text-search-application-with-elasticsearch-and-fastapi-14ea78cf1890 ElasticSearch-FastAPI 연결 참고 자료