캡스톤디자인I 중간보고서

<table>
<tr><td rowspan="2">제 목</td><td>국문</td><td>AI 챗봇을 활용한 온라인 강의 웹 서비스</td></tr>
<tr><td>영문</td><td>Online lecture web service using AI chatbot</td></tr>
<tr><td rowspan="2">진 행
상 황</td><td>중요마일스톤</td><td>1. 기능 요구 사항 정의 - 질문 분석기, 임베딩 검색기, 문서 검색기, 언어모델, 기계독해, 메인 페이지, 로그인 관리, 내강의목록 페이지, 강의 대시보드 페이지, 강의 동영상 페이지, 챗봇 채팅, 강의 관리자 페이지, 회원 및 강사 데이터 처리
2. 성능 요구사항 정의 - 챗봇 서버 동시 접속 수, 질문에 대한 응답 시간, 언어모델 fine tuning
3. 사용자 인터페이스 요구사항 정의 - 웹 구성, 웹 디자인, 메인 페이지, 강의 대시보드 구성
4. 시스템 인터페이스 요구사항 정의 - 강의 업로드, 데이터베이스와 임베딩 검색기의 연계, 임베딩 검색기와 문서 검색기의 연계, 문서 검색기와 언어모델의 연계, 데이터베이스와 언어모델의 연계, 챗봇UI와 질문 분석기의 연계, 질문 분석기와 언어모델의 연계, 언어모델과 기계독해기의 연계, 기계독해기와 챗봇UI의 연계
5. 데이터 요구사항 정의 - 초기자료 구축, 임베딩 파일, 임베딩 vocab, 언어모델 vocab, 데이터베이스 관리
6. 테스트 요구사항 정의 - 언어모델 임베딩 성능 테스트, 임베딩 길이 테스트, 질문 분석기 성능 테스트, 기계독해기 성능 테스트, 사용자 웹 페이지와 챗봇 서버 연결 테스트, 사용자 웹 페이지 회원 관리, 강의 플레이어와 챗봇 채팅 테스트
7. 보안 요구사항 - 응용 및 DB보안, 웹 페이지 보안
8. 품질 요구사항 - 챗봇의 답변 정확도, 챗봇의 답변 속도, 챗봇 서버 오류 처리, 챗봇 UI의 가독성, 언어모델의 이식성
9. 제약 사항 정의 - 언어모델 변경 제약 사항, 질문 분석기 변경 제약 사항, 웹 클라이언트-챗봇 서버 간 통신 데이터 제약 사항, 시스템 구조 설계
10. 프로젝트 관리 요구사항 - 프로젝트 진행 방법론</td></tr>
<tr><td>진행상황</td><td>- 핵심 기능 파악 완료
- 기능 요구 사항 정의 완료
- 사용자 인터페이스 요구사항 정의 완료
- 시스템 인터페이스 요구사항 정의 완료
- 데이터 요구사항 정의 완료
- 보안 요구사항 정의 완료
- 품질 요구사항 정의 완료
- 프로젝트 관리 요구사항 정의 완료
- 테스트 계획 수립 완료
- 웹페이지 중 강의영상과 챗봇 채팅 페이지 구현완료
- 클라이언트, 서버 메시지 송수신 구현 중
- 프로토 타입 모델 학습 완료
- E/R 다이어그램 설계완료, Entity set, Relation set을 Relation으로 전환한 데이터베이스 설계 완료.</td></tr>
<tr><td>산출물</td><td colspan="2">요구사항 정의서(별첨 1), 중간보고서(별첨 2)</td></tr>
</table>

팀 구성원	학년	학 번	이 름	연락처(전화번호/이메일)
	4	20171612	박상현	010-5212-8903 / 20171612@edu.hanbat.ac.kr
	4	20197132	주준하	010-9221-1305 / 20197132@edu.hanbat.ac.kr
	4	20171768	남승완	010-4015-0181 / nsw05138@naver.com

컴퓨터공학과의 프로젝트 관리규정에 따라 다음과 같이 요구사항 정의서와 중간보고서를 제출합니다

2022 년 4월 29일

책임자 : 박상현(인)

지도교수 : 임경대(인)

[별첨1]

프로젝트명 : AI 챗봇을 활용한 온라인 강의 웹 서비스

소프트웨어 요구사항 정의서

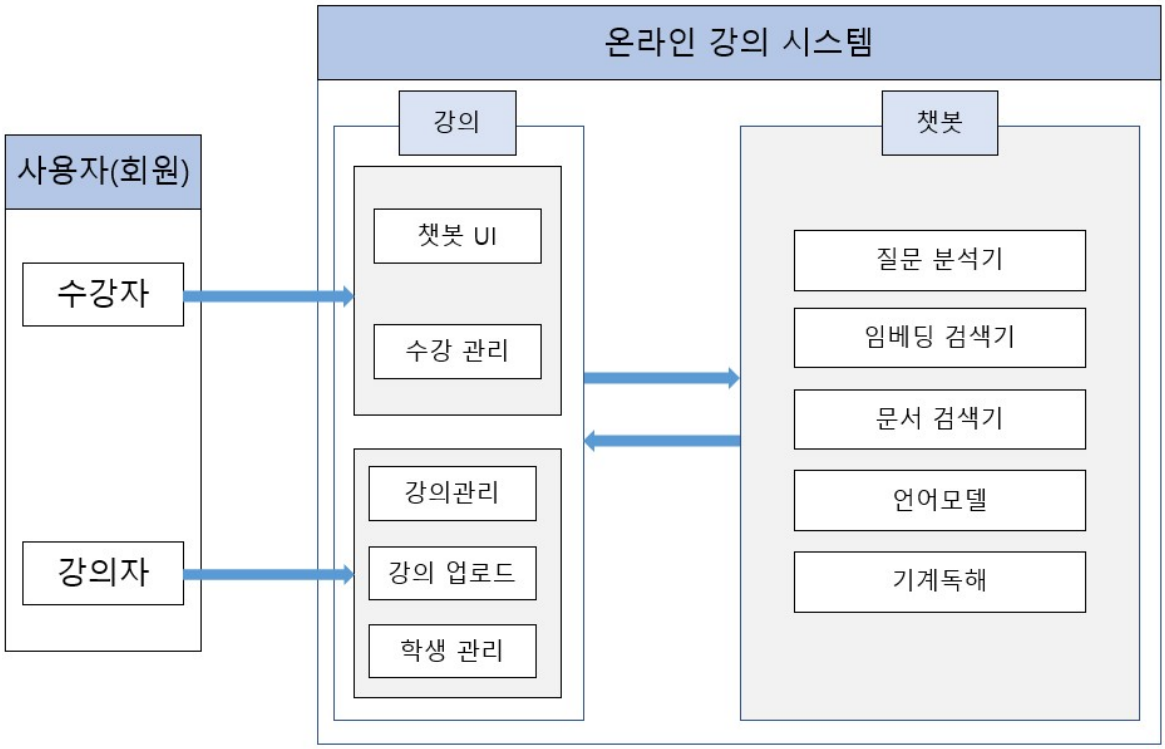
Version 1.0

개발 팀원 명(팀리더):박상현
주준하
남승완

대표 연락처: 010-5212-8903
e-mail: 20171612@edu.hanbat.ac.kr

목차

1. 시스템 개요


온라인 강의 시스템
강의
챗봇
사용자(회원)
수강자
강의자
챗봇 UI
수강 관리
강의관리
강의 업로드
학생 관리
질문 분석기
임베딩 검색기
문서 검색기
언어모델
기계독해

2. 시스템 장비 구성요구사항

요구사항 고유번호		ECR-001		
요구사항 명칭		웹 개발장비		
요구사항 분류		시스템 장비구성 요구사항	응락수준	필수
요구사항 상세설명	정의	웹 어플리케이션		
	세부내용	- 장비 품목 : Asus Laptop - 장비 수량 : 1개 - 장비 기능 : 챗봇 기능과 동영상 플레이어. - 장비 성능 및 특징 : 사용자는 챗봇을 이용하여 질의응답 서비스를 이용하며 온라인 강의 수강할 수 있는 UI를 개발. 관리자는 챗봇에 이용할 수 있는 키워드를 제공하고 사용자에게 온라인 강의를 제공 할 수 있도록 관리자 페이지 개발		

요구사항 고유번호		ECR-002		
요구사항 명칭		GPU		
요구사항 분류		시스템 장비구성 요구사항	응락수준	필수
요구사항 상세설명	정의	언어모델(BERT) 학습 및 fine tuning을 위한 GPU		
	세부내용	- 장비 품목 : NVIDIA Tesla P100 GPU - 장비 수량 : 1개 - 장비 기능 : 언어모델 학습 및 fine tuning을 위한 GPU - 장비 성능 및 특징: google colab pro에서 제공하는 GPU mezzanine(NVLink): 3584 * 1328/1480 MHz, 9519-10609 GFLOPS, 16G memory		

요구사항 고유번호		ECR-003		
요구사항 명칭		챗봇 서버		
요구사항 분류		시스템 장비구성 요구사항	응락수준	필수
요구사항 상세설명	정의	질의응답 서비스 제공 서버		
	세부내용	- 장비 품목 : 노트북 - 장비 수량 : 1개 - 장비 기능 : 질의응답 서비스 제공 - 장비 성능 및 특징: 질의응답 서비스 제공을 위한 python 패키지 설치 필요(websocket, pytorch, transformers)		

요구사항 고유번호		ECR-004		
요구사항 명칭		웹 운영 서버		
요구사항 분류		시스템 장비구성 요구사항	응락수준	필수
요구사항 상세설명	정의	- 웹 서비스 제공		
	세부내용	- 장비 품목 : - 장비 수량 : 1 - 장비 기능 : 강의동영상 재생 플랫폼, 데이터 로직 처리 - 장비 성능 및 특징: 1. 클라이언트로부터 요청받은 로그인 또는 회원가입 정보를 받아서 DB에 저장 또는 로그인에 대해 성공여부 보냄 2. 로그인 여부에 따라 웹서비스 이용 3. 질의응답 모델을 호출하여 사용 가능해야함. 4. 사용자에게 보여지는 프론트부분과 로직을 처리하는 백엔드가 분리되어 존재.		

요구사항 고유번호		ECR-004		
요구사항 명칭		DataBase		
요구사항 분류		시스템 장비구성 요구사항	응락수준	필수
요구사항 상세설명	정의	- 데이터저장 및 관리		
	세부내용	- 장비 품목 : - 장비 수량 : 1 - 장비 기능 : 회원정보 저장, 조회, 수정 - 장비 성능 및 특징: 사용자 정보 저장, 요청에 따른 조회 또는 수정		

3. 기능 요구사항

요구사항 고유번호		SFR-001		
요구사항 명칭		질문 분석기		
요구사항 분류		기능 요구사항	응락수준	필수
요구사항 상세설명	정의	사용자 질문 분석, 주요 키워드 추출		
	세부내용	1. 사용자 질문 분석 및 키워드(일반, 고유 명사) 추출 2. 핵심적인 단어 몇 가지를 선정 3. 임베딩 검색기로 전달		

요구사항 고유번호		SFR-002		
요구사항 명칭		임베딩 검색기		
요구사항 분류		기능 요구사항	응락수준	필수
요구사항 상세설명	정의	서버 내 저장된 키워드에 대한 임베딩 검색		
	세부내용	1. 서버 내에 키워드에 대한 임베딩 파일 검색 2. 파일 경로 - 키워드명/문단주제.embedding -예시 : 챗봇/정의.embedding 3. 키워드 디렉터리가 없으면 문서 검색기로 전달 4. 키워드 디렉터리가 있으면 질문과 파일명의 유사도 측정 후 답변을 검색할 문단 영역 결정		

요구사항 고유번호		SFR-003		
요구사항 명칭		문서 검색기		
요구사항 분류		기능 요구사항	응락수준	필수
요구사항 상세설명	정의	wikipedia에서 키워드에 대한 정보를 검색		
	세부내용	1. 동음이의어에 의한 복수 검색 키워드에서 필요한 키워드 선정 2. 키워드에 대한 문서 검색 3. 문서를 문단 단위로 분할 4. 문단 단위로 언어모델에 입력		

요구사항 고유번호		SFR-004		
요구사항 명칭		언어모델		
요구사항 분류		기능 요구사항	응락수준	필수
요구사항 상세설명	정의	BERT 기반의 자연어 임베딩 생성 모델		
	세부내용	1. 질문에 대한 임베딩 생성 2. 문서 검색기로부터 얻은 정보는 문단 별로 임베딩하여 '키워드 명'/문단주제.embedding으로 저장 3. 기본 모델로 구글의 다국어 BERT 모델을 사용하며 서비스 성능 향상을 위해서 한국어 BERT 모델인 ETRI의 KorBERT를 사용		

요구사항 고유번호		SFR-005		
요구사항 명칭		기계독해		
요구사항 분류		기능 요구사항	응락수준	필수
요구사항 상세설명	정의	사용자 질문에 대한 임베딩과 문단 임베딩을 이용한 답변 영역 예측		
	세부내용	1. FFNN, softmax를 이용한 문단 내 답변의 시작과 종료 영역 결정 2. 시작 임베딩 토큰, 종료 임베딩 토큰의 인덱스를 이용해서 자연어 답변영역 추출		

요구사항 고유번호		SFR-006		
요구사항 명칭		메인 페이지		
요구사항 분류		기능	응락수준	필수
요구사항 상세설명	정의	웹 사이트의 메인 페이지.		
	세부내용	- 사이트에 등록된 강의 목록과 최근 학습 강의 확인. - 로그인 비활성화시 nav의 전체강의 버튼을 제외한 모든 버튼 클릭 시 로그인 창으로 이동 - 관리자로 인증되지 않은 일반 사용자는 관리자페이지 비활성화 안내.		

요구사항 고유번호		SFR-007		
요구사항 명칭		로그인 관리		
요구사항 분류		기능	응락수준	선택
요구사항 상세설명	정의	인증된 사용자에 한하여 로그인.		
	세부내용	- 학교계정의 사용자에 한하여 회원가입 가능. @edu.hanbat.ac.kr로 끝나는 이메일을 통해 인증 - 자동 로그인 사용 체크하면 자동으로 로그인 가능. 쿠키를 통해 관리 쿠키 삭제시 자동 로그인이 되지 않음. - 로그인을 하지 않고 메인 강의 목록 페이지로 이동 가능. - email, password 형식의 로그인		

요구사항 고유번호		SFR-008		
요구사항 명칭		내 강의목록 페이지		
요구사항 분류		기능	응락수준	필수
요구사항 상세설명	정의	전체 강의목록 확인 가능		
	세부내용	- 로그인된 사용자가 등록한 강의 목록확인 - 강의목록 section 클릭 시 강의창으로 이동. - 강의의 진도율 표시 - 로그인되지 않은 사용자가 접근 시 로그인 페이지로 이동. - 스크롤을 통해 강의 목록 확인.		

요구사항 고유번호		SFR-009		
요구사항 명칭		강의 대시보드 페이지		
요구사항 분류		기능	응락수준	필수
요구사항 상세설명	정의	강의 대시보드를 통해 강의 학습		
	세부내용	- 로그인된 사용자는 자신이 등록한 강의 대시보드를 확인 할 수 있다.		

요구사항 고유번호		SFR-010		
요구사항 명칭		강의 동영상 페이지		
요구사항 분류		기능	응락수준	필수
요구사항 상세설명	정의	동영상으로 이루어진 강의를 수강 할 수 있고 챗봇을 이용할 수 있다.		
	세부내용	- 동영상 영역과 챗봇 채팅 영역으로 이루어져 있다. - 챗봇 채팅 영역은 오른쪽 nav를 통해 강의 목차로 바꿀 수 있다. - 챗봇 채팅은 강의 목차로 바뀌어도 기록이 저장된다. - 동영상은 유튜브플레이어를 통해 재생된다. - 강의 목록은 스크롤을 통해 볼 수 있다.		

요구사항 고유번호		SFR-011		
요구사항 명칭		챗봇 채팅		
요구사항 분류		기능	응락수준	필수
요구사항 상세설명	정의	챗봇을 통해 사용자는 질의 응답을 할 수 있다.		
	세부내용	- 텍스트를 통해 챗봇에게 질문을 할 수 있으며 챗봇은 적절한 답변을 전달한다. - 챗봇은 부적절한 사용자의 질문이 있을 시 사용자에게 응답이 어려움을 전달한다.		

요구사항 고유번호		SFR-012		
요구사항 명칭		강의 관리자 페이지		
요구사항 분류		기능	응락수준	필수
요구사항 상세설명	정의	등록된 강의 관리자 전용 강의 관리.		
	세부내용	- 관리자는 동영상 링크 또는 파일을 통해 업로드. - 강의에 관련한 키워드 입력. - 강의 관련 내용을 작성.		

요구사항 고유번호		SFR-013		
요구사항 명칭		회원 데이터 처리		
요구사항 분류		기능 요구사항	응락수준	필수
요구사항 상세설명	정의	- 회원가입		
	세부내용	- 회원가입을 요구하는 사용자에게 사용자가 정한 id와 password를 등록 - 회원은 학부생과 강사로 분류. - DB에 암호화하여 저장		

요구사항 고유번호		SFR-014		
요구사항 명칭		강사 데이터 처리		
요구사항 분류		기능 요구사항	응락수준	필수
요구사항 상세설명	정의	- 키워드 등록		
	세부내용	- 강사로 분류된 회원은 자신의 강의의 핵심 주제 키워드를 등록. - 강사로 분류된 회원은 자신의 유튜브 영상을 플로팅하여 한 강의의 목차로 업로드 가능.		

4. 성능 요구사항

요구사항 고유번호		PER-001		
요구사항 명칭		챗봇 서버 동시 접속 수		
요구사항 분류		성능	응락수준	필수
요구사항 상세 설명	정의	- 챗봇 서버에 대한 동시 접속 제한		
	세부 내용	- 다수 사용자를 위한 멀티 스레드 기반의 웹 소켓 서버 - 일정 시간 미사용 세션 종료		

요구사항 고유번호		PER-002		
요구사항 명칭		질문에 대한 응답 시간		
요구사항 분류		성능	응락수준	필수
요구사항 상세 설명	정의	챗봇 응답 시간		
	세부 내용	- 응답 시간 최소화를 위해서 강의 업로드 시 키워드를 제공해서 문서에 대한 임베딩을 미리 생성 - 최소 응답 시간(질문 임베딩 생성 시간 + 문서 임베딩 검색 시간 + 답변 예측 시간) - 최대 응답 시간(질문 임베딩 생성 시간 + 문서 임베딩 검색 시간 + 문서 임베딩 생성 시간 + 답변 예측 시간) - 질문 임베딩 생성 시간 : 1초 이내 - 문서 임베딩 검색 시간 : 1초 이내 - 문서 임베딩 생성 시간 : 단문 기준 3초 이내(문서 길이 비례) - 답변 예측 시간 : 2초 이내		

요구사항 고유번호		PER-003		
요구사항 명칭		언어모델 fine tuning		
요구사항 분류		성능	응락수준	필수
요구사항 상세 설명	정의	언어모델을 질의응답 데이터 셋으로 fine tuning할 때 소요되는 자원		
	세부 내용	- gpu 메모리는 12GB 이상 필요하고 60,407개 질의응답 쌍(36MB 분량)의 데이터 셋을 학습시키는데 5시간 이상 소요(GPU 성능에 따라 차이 존재)		

5. 인터페이스 요구사항

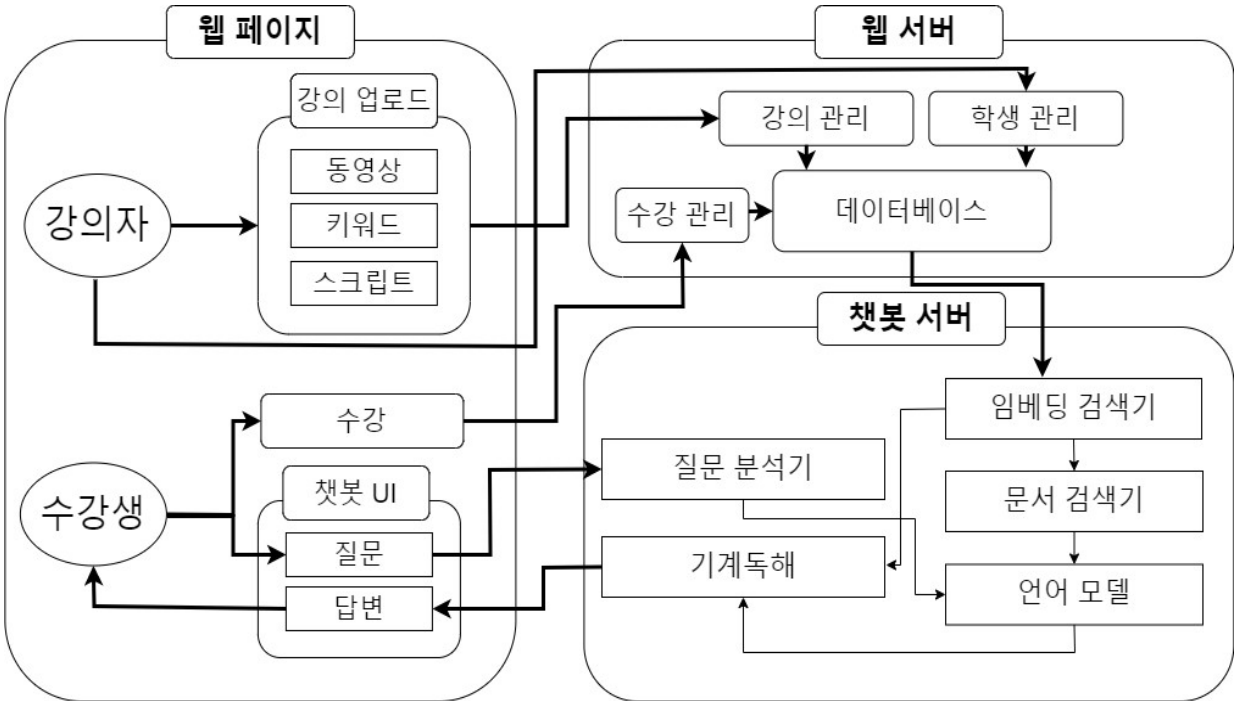
요구사항 고유번호		SIR-001	
요구사항 명칭	웹 구성		
요구사항 분류	인터페이스 요구사항	응락수준	필수
요구사항 상세설명	정의	웹 사이트의 메인 페이지.	
	세부 내용	- nav 영역에는 전체강의버튼, 내강의버튼, 관리자버튼, 최근강의버튼, 내정보버튼으로 이루어짐 - nav 컴포넌트를 재사용. - 로그인 상태가 아닐 때 회원만 이용가능한 항목은 로그인이나 회원가입을 유도. - 강의목록 section 클릭 시 강의 대시보드로 이동.	
주석	•		
요구사항 출처	•		

요구사항 고유번호		SIR-002	
요구사항 명칭	웹 디자인		
요구사항 분류	인터페이스 요구사항	응락수준	필수
요구사항 상세설명	정의	사용자 편리를 위한 웹 디자인	
	세부 내용	- 상하 스크롤을 통해 화면 이동 - 사용자 피로도 최소화와 가독성을 고려한 색상과 글씨 크기를 구성 - 버튼과 텍스트의 구분을 확실하게 하여 사용자가 사용하기 직관적이게 설계	
주석	•		
요구사항 출처	•		

요구사항 고유번호		SIR-003	
요구사항 명칭	메인 페이지		
요구사항 분류	인터페이스 요구사항	응락수준	필수
요구사항 상세설명	정의	웹 사이트의 메인 페이지.	
	세부 내용	- main 영역은 전체강의 목록영역과 최근 학습 강의 목록 영역으로 나뉘어짐 - 전체 강의 목록영역과 최근 학습 강의 목록 영역은 최대 3개의 강의를 보여주며 강의 이미지, 제목, 설명으로 하나의 section을 이룬다.	
주석	•		
요구사항 출처	•		

요구사항 고유번호		SIR-004	
요구사항 명칭	강의 대시보드 구성		
요구사항 분류	인터페이스 요구사항	응락수준	필수
요구사항 상세설명	정의	사용자 강의 대시보드 구성	
	세부 내용	- 강의 진도율과 강의 커리큘럼을 확인 할 수 있다. - 강의 목차를 클릭하여 수강 할 수 있다. - 전에 학습했던 구간을 이어서 수강 할 수 있다.	
주석	•		
요구사항 출처	•		

2) 시스템 인터페이스 요구사항 분석 및 도출



요구사항 고유번호		SIR-005	
요구사항 명칭	강의 업로드		
요구사항 분류	인터페이스	응락수준	필수
요구사항 상세설명	정의	강의 동영상을 웹 서버 데이터베이스에 저장하기 위한 인터페이스	
	세부 내용	- 동영상 파일 또는 유튜브 동영상 링크 업로드에 대한 인터페이스 - 강의 주제, 핵심 키워드 저장에 대한 인터페이스 - 강의 대본(스크립트) 저장에 대한 인터페이스	
주석	강의 대본 업로드는 선택 사항이나 챗봇의 답변 영역을 강의 내 발화 내용까지 확장할 수 있음		
요구사항 출처	•		

요구사항 고유번호		SIR-006		
요구사항 명칭		데이터베이스와 임베딩 검색기의 연계		
요구사항 분류		인터페이스	응락수준	필수
요구사항 상세 설명	정의	웹 서버 데이터베이스와 챗봇 서버 임베딩 검색기간의 인터페이스		
	세부 내용	- 강의 업로드를 챗봇 서버의 임베딩 검색기에 통보하고 강의 키워드를 전달하여 임베딩이 존재하는 지 검색 - 이미 존재하는 임베딩 파일을 중복해서 만들지 않기 위한 작업		
산출정보		키워드에 대한 임베딩 파일 유무		

요구사항 고유번호		SIR-007		
요구사항 명칭		임베딩 검색기와 문서 검색기의 연계		
요구사항 분류		인터페이스	응락수준	필수
요구사항 상세 설명	정의	임베딩 검색기와 문서 검색기간의 인터페이스		
	세부 내용	- 임베딩 파일이 존재하지 않은 키워드를 문서 검색기에 전달하고 위키피디아에 검색하여 정보를 획득 - 문서 검색기에서는 동음이의어에 대한 처리 절차가 필요		
산출정보		키워드에 대한 위키피디아 정보		

요구사항 고유번호		SIR-008		
요구사항 명칭		문서 검색기와 언어모델의 연계		
요구사항 분류		인터페이스	응락수준	필수
요구사항 상세 설명	정의	문서 검색기와 언어모델간의 인터페이스		
	세부 내용	- 문서 검색기를 통해 얻은 키워드에 대한 위키피디아 문서를 문단 단위로 분할하여 언어모델에 입력 - 생성된 임베딩들을 '키워드 명'/'문단주제'.embedding으로 저장 - 임베딩이 생성된 키워드를 vocab 파일에 추가 - 언어모델로 입력되는 토큰의 수는 최대 512개로 제한		
산출정보		키워드에 대한 임베딩 정보		

요구사항 고유번호		SIR-009		
요구사항 명칭		데이터베이스와 언어모델의 연계		
요구사항 분류		인터페이스	응락수준	필수
요구사항 상세 설명	정의	웹 서버의 데이터베이스와 챗봇 서버의 언어모델간의 인터페이스		
	세부 내용	- 강의 스크립트에 대한 임베딩을 생성하기 위해서 데이터베이스에 저장된 스크립트를 언어모델로 전달하여 '강의명'.embedding 파일로 저장 - 언어모델로 입력되는 토큰의 수는 최대 512개로 제한		
산출정보		강의 스크립트에 대한 임베딩		

요구사항 고유번호		SIR-009		
요구사항 명칭		챗봇UI와 질문 분석기의 연계		
요구사항 분류		인터페이스	응락수준	필수
요구사항 상세 설명	정의	챗봇UI와 질문 분석기간의 인터페이스		
	세부 내용	- 사용자로부터 받은 질문을 챗봇 서버의 질문 분석기로 전달하기 위한 인터페이스 - 질문 분석기는 받은 질문을 분석해서 핵심적인 키워드를 추출 - 정확한 키워드 추출을 위해서 vocab 파일 조회		
산출정보		질문에 대한 핵심 키워드		

요구사항 고유번호		SIR-010		
요구사항 명칭		질문 분석기와 언어모델의 연계		
요구사항 분류		인터페이스	응락수준	필수
요구사항 상세 설명	정의	질문 분석기와 언어모델간의 인터페이스		
	세부 내용	- 사용자 질문을 언어모델로 전달하기 위한 인터페이스 - 질문에 대한 임베딩을 생성 - 언어모델로 입력되는 토큰의 수는 최대 512개로 제한		
산출정보		질문에 대한 임베딩		

요구사항 고유번호		SIR-011		
요구사항 명칭		언어모델과 기계독해기의 연계		
요구사항 분류		인터페이스	응락수준	필수
요구사항 상세 설명	정의	언어모델과 기계독해기간의 인터페이스		
	세부 내용	- 언어모델에서 생성한 질문 임베딩, 문서 임베딩을 기계독해기로 전달하는 인터페이스 - 기계독해기에서 문서 임베딩 내에서 질문에 대한 답변 영역을 예측		
산출정보		답변 영역		

요구사항 고유번호		SIR-012		
요구사항 명칭		기계독해기와 챗봇UI의 연계		
요구사항 분류		인터페이스	응락수준	필수
요구사항 상세 설명	정의	기계독해기와 챗봇UI간의 인터페이스		
	세부 내용	- 예측한 답변영역을 사용자에게 전달하기위한 인터페이스 - 챗봇 UI에서 답변의 가독성을 위한 길이 조절 또는 출력 형식 변화 필요		
산출정보		챗봇UI에 답변 출력		

6. 데이터 요구사항

요구사항 고유번호	DAR-001		
요구사항 명칭	초기자료 구축		
요구사항 분류	데이터	응락수준	필수
요구사항 세부내용	- 강의 동영상을 업로드하고 분야에 맞게 분류 - BERT를 fine tuning 질의응답 테스크에 맞춰 fine tuning하기 위해서 질문-답변 쌍의 데이터 구축(KorQuAD로 대체)		

요구사항 고유번호	DAR-002		
요구사항 명칭	임베딩 파일		
요구사항 분류	데이터	응락수준	필수
요구사항 세부내용	- 키워드 임베딩 파일의 저장 경로는 '키워드 명'/'문단 주제'.embedding - 강의 스크립트 임베딩 파일의 저장 경로는 '강의명'.embedding - 질문과 문단 주제의 연관성을 분석하여 적절한 임베딩 파일을 선택할 수 있도록 문단 주제를 파일 명에 명시 - 임베딩 가능한 토큰의 수가 512개로 제한되므로, 한 강의에 대한 복수의 embedding 파일 존재 가능		

요구사항 고유번호	DAR-003		
요구사항 명칭	임베딩 vocab		
요구사항 분류	데이터	응락수준	필수
요구사항 세부내용	- 임베딩이 생성된 키워드들을 저장한 파일 - 임베딩이 생성될 때마다 업데이트되며, 키워드 외에 별도 정보를 저장하지 않음.		

요구사항 고유번호	DAR-004		
요구사항 명칭	언어모델 vocab		
요구사항 분류	데이터	응락수준	필수
요구사항 세부내용	- 언어모델에서 문장 tokenize 및 인덱스 부여에 사용하는 파일 - 언어모델을 교체하지 않는 한 변동되지 않음.		

요구사항 고유번호	DAR-005		
요구사항 명칭	데이터베이스 관리		
요구사항 분류	데이터	응락수준	필수
요구사항 세부내용	- 테이블은 유저 정보 테이블, 강의명과 사용자 질문의 쌍 테이블로 설정 - redundancy를 최소화 해야 함. - null값이 없도록 해야 함.		

7. 테스트 요구사항

요구사항 고유번호	TER-001		
요구사항 명칭	언어모델 임베딩 성능 테스트		
요구사항 분류	테스트	응락수준	필수
요구사항 세부내용	- 복수의 언어모델에 대해서 동일한 검증 데이터 셋을 준비하고 가장 높은 성능을 보이는 모델을 선택		

요구사항 고유번호	TER-002		
요구사항 명칭	임베딩 길이 테스트		
요구사항 분류	테스트	응락수준	필수
요구사항 세부내용	- 모델에서 한 번에 생성 가능한 토큰의 임베딩 수는 최대 512개로 모델에 입력하기 전에 길이를 적절하게 분할하는 로직이 정상작동하는 지 테스트		

요구사항 고유번호	TER-003		
요구사항 명칭	질문 분석기 성능 테스트		
요구사항 분류	테스트	응락수준	필수
요구사항 세부내용	- 질문 분석기가 키워드를 추출할 때 위키피디아에서 검색 가능한 형태로 추출하는 지에 대한 테스트		

요구사항 고유번호	TER-004		
요구사항 명칭	기계독해기 성능 테스트		
요구사항 분류	테스트	응락수준	필수
요구사항 세부내용	- 기계독해기에서 사용하는 FFNN가 학습 데이터 셋에 과적합 되지 않고 다양한 입력 케이스에서 적절한 답을 출력하는 지에 대한 성능 테스트		

요구사항 고유번호	TER-005		
요구사항 명칭	사용자 웹 페이지와 챗봇 서버 연결 테스트		
요구사항 분류	테스트	응락수준	필수
요구사항 세부내용	- 사용자가 챗봇 UI에 입력하는 데이터를 챗봇 서버로 정확하게 적절한 시간 내에 전달하는 지와 다수의 사용자가 동시 접속했을 때 이를 효과적으로 처리할 수 있는 지에 대한 성능 테스트		

요구사항 고유번호	TER-006		
요구사항 명칭	사용자 웹 페이지 회원 관리		
요구사항 분류	테스트	응락수준	필수
요구사항 세부내용	- 로그인 상태에 활성화된 페이지가 정상적으로 작동 하는지 테스트		

요구사항 고유번호		TER-007		
요구사항 명칭		강의 플레이어와 챗봇 채팅 테스트		
요구사항 분류		품질(기술관점)	응락수준	필수
요구사항 상세 설명	정의	다른 브라우저와 사용자 장치 환경에서 잘 작동하는지 테스트		
	세부 내용	- 사용자마다 사용하는 브라우져와 장치 환경이 다를때 오류없이 정상적으로 작동 하는지 테스트		

8. 보안 요구사항

요구사항 고유번호	SER-001		
요구사항 분류	응용 및 DB보안		
요구사항 분류	보안	응락수준	필수
요구사항 세부 내용	- 일반 사용자는 직접적으로 DB접근을 할 수 없음. - 사용자 개인정보(아이디, 이름, 학번, 비밀번호 등)는 소스코드에 직접 하드코딩 하지 않음. - 해싱을 통해 내부 관리자도 패스워드를 알아볼 수 없게 함.		

요구사항 고유번호	SER-002		
요구사항 분류	웹페이지 보안		
요구사항 분류	보안	응락수준	필수
요구사항 세부 내용	- 편의를 위한 쿠키의 이용에서 쿠키는 사용자 개개인에 따라 구분되어야 함. - 쿠키의 이름에 개인정보가 들어가지 않도록 쿠키명을 무작위 또는 암호화하여 생성해야 함.		

9. 품질 요구사항

요구사항 고유번호		QUR-001		
요구사항 명칭		챗봇의 답변 정확도		
요구사항 분류		품질(기술관점)	응락수준	필수
요구사항 상세 설명	정의	언어모델 성능과 기계독해기의 성능과 관련된 챗봇의 답변 정확도		
	세부 내용	- 답변 정확도가 0.9 이하라면 언어모델과 기계독해기 성능의 향상이 필요 - 사용자가 답변의 내용적 오류를 발견하지 못하거나 오인할 가능성이 있는 답변을 제시하는 경우 서비스 구조의 전반적인 검토가 필요		

요구사항 고유번호		QUR-002		
요구사항 명칭		챗봇의 답변 속도		
요구사항 분류		품질(기술관점)	응락수준	필수
요구사항 상세 설명	정의	사용자가 질문을 한 시점부터 답변을 받을 때까지 걸리는 시간		
	세부 내용	- 답변을 생성할 때 가장 오래 걸리는 케이스는 질문 임베딩 생성, 문서 임베딩 검색, 문서 임베딩 생성, 답변 예측을 모두 포함할 때이며 속도 향상을 위해서 cpu, gpu 등의 성능 향상이 필요		

요구사항 고유번호		QUR-003		
요구사항 명칭		챗봇 서버 오류 처리		
요구사항 분류		품질(기술관점)	응락수준	필수
요구사항 상세 설명	정의	챗봇 서버에서 발생하는 다양한 오류에 대한 처리		
	세부 내용	- 다수 사용자의 동시 접속에 의한 성능 저하, 오류에 대해서 적절한 오류 발생 메시지를 사용자에게 전달 - 답변 생성 과정에서 발생하는 오류에 대해서 복구가 불가능할 시 사용자에게 적절한 메시지를 전달		

요구사항 고유번호		QUR-004		
요구사항 명칭		챗봇 UI의 가독성		
요구사항 분류		품질(기술관점)	응락수준	필수
요구사항 상세 설명	정의	챗봇 UI가 출력하는 답변의 가독성		
	세부 내용	- 답변이 긴 경우 가독성이 떨어지기 때문에 적절한 길이 조절 또는 출력 형태 수정이 필요		

요구사항 고유번호		QUR-005		
요구사항 명칭		언어모델의 이식성		
요구사항 분류		품질(기술관점)	응락수준	필수
요구사항 상세 설명	정의	언어모델의 기존 서버와 다른 운영체제에서 동작하는 지에 대한 여부		
	세부 내용	- 개발 완료 후 웹 서버와 통합 또는 서버 이전을 할 때 새로운 환경에서 정상작동을 하기위해서 운용 운영체제와 호환되는 파이썬 패키지 설치가 필요 - pytorch, transformers, tensorflow 등		

10. 제약 사항

요구사항 고유번호	COR-001		
요구사항 명칭	언어모델 변경 제약 사항		
요구사항 분류	제약사항	응락수준	필수
요구사항 세부내용	- 언어모델은 기본적으로 BERT 기반의 모델을 사용하며, 성능 향상, 안전성 등의 목적으로 교체할 때 pytorch로 작성된 모델을 사용한다. 또한 문장 tokenize에 사용하는 방법을 언어모델에서 사용하는 방식과 일치 시켜야한다.		

요구사항 고유번호	COR-002		
요구사항 명칭	질문 분석기 변경 제약 사항		
요구사항 분류	제약사항	응락수준	필수
요구사항 세부내용	- 질문 분석기의 정확도 향상을 위해서 사용되는 알고리즘, 도구 등을 변경할 때 선정되는 키워드의 품사는 명사, 고유명사가 되도록 한다.		

요구사항 고유번호	COR-003		
요구사항 명칭	웹 클라이언트, 챗봇 서버 간 통신 데이터 제약 사항		
요구사항 분류	제약사항	응락수준	필수
요구사항 세부내용	- 클라이언트의 사용자와 챗봇 서버 간에 교환(질문, 답변)하는 데이터는 확장성을 위해서 Json을 사용한다.		

요구사항 고유번호	COR-004		
요구사항 명칭	시스템 구조 설계		
요구사항 분류	제약사항	응락수준	필수
요구사항 세부내용	- 오류 수정, 성능 향상, 안정성 등을 목적으로 기존 모듈을 교체를 유연하게 하기 위해서 모듈 간 인터페이스를 독립적으로 구성하고 데이터 입출력 방식이 변경되지 않도록 설계한다.		

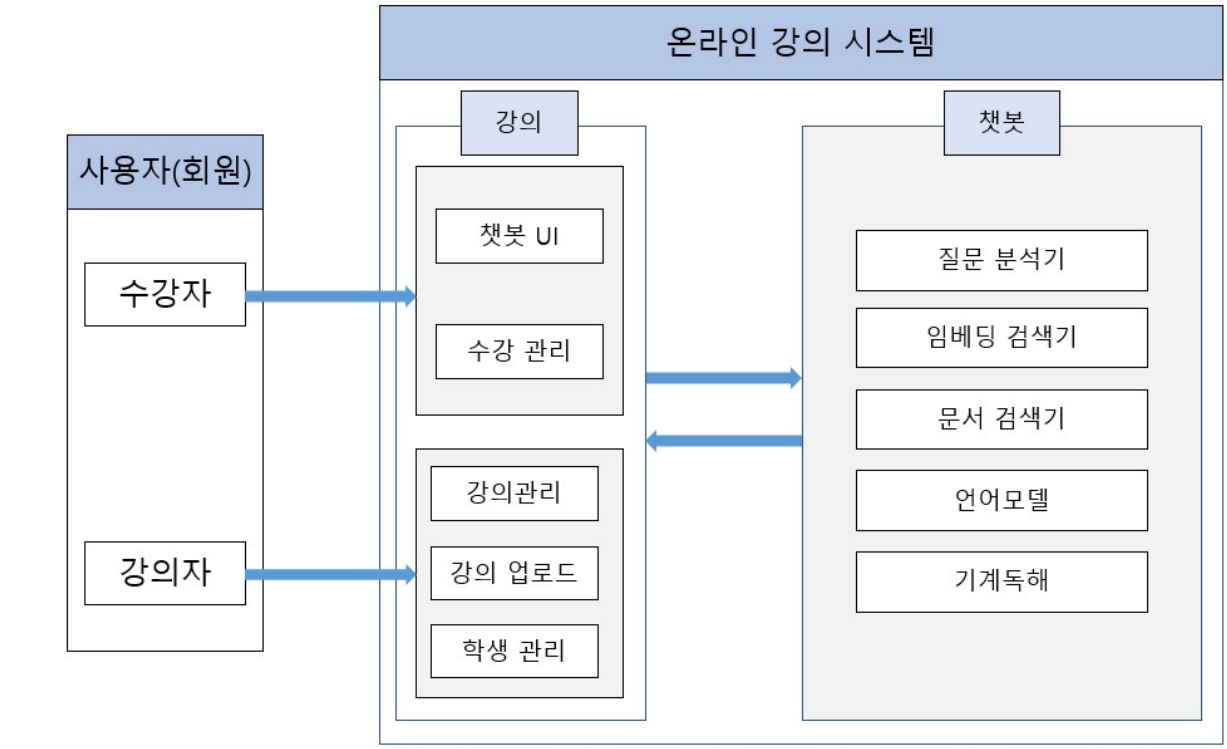
11. 프로젝트 관리 요구사항

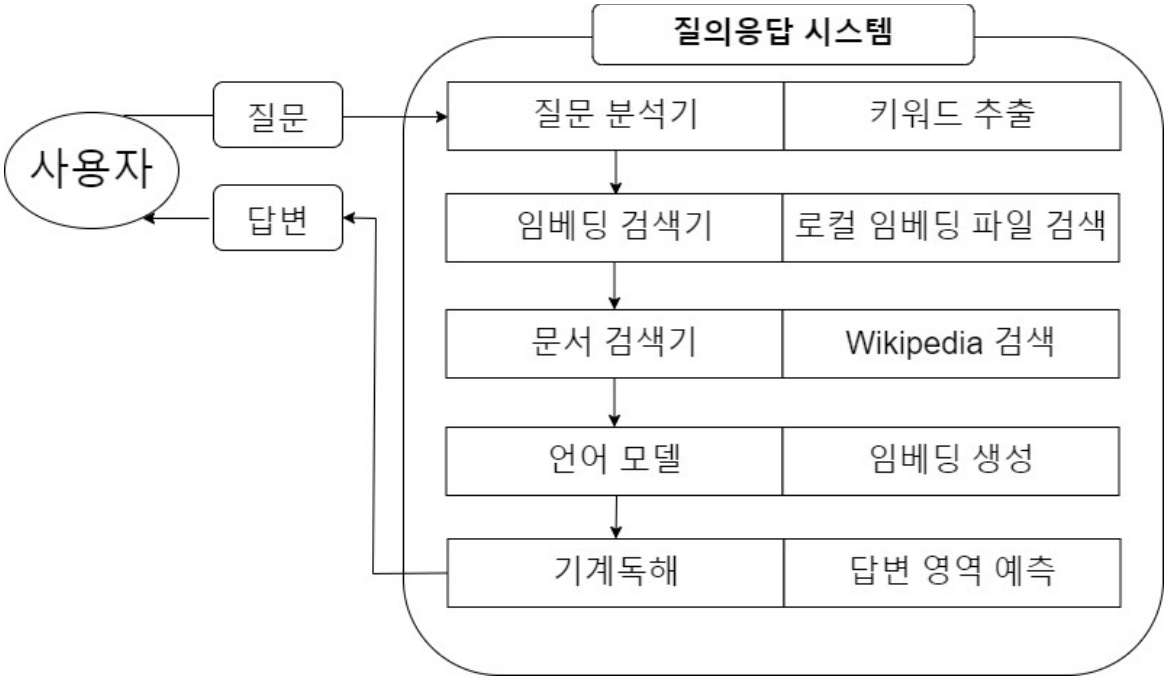
요구사항 고유번호	**PMR-001**		
요구사항 명칭	품질관리		
요구사항 분류	프로젝트 관리	응락수준	필수
요구사항 세부내용	- 웹서비스 개발과 챗봇 모델 개발로 나눈다. - 웹서비스는 UI개발, DB, 서버로 나뉜다. - UI는 프론트엔드가 맡으며 동작의 전반적인 구조를 설계해야한다. - DB, 서버는 서비스 배포, 아키텍쳐 전반적인 인프라를 설계해야한다. - 프로토타입을 공유하여 수정, 보완 사항을 의논한다. - 각 구성원들은 각자 맡은 분야에 대한 설계, 개발, 문서화를 진행하고 주기적인 검토와 회의를 통해서 전체적인 설계 내용을 구현하고 통합한다.		

[별첨2]

중간보고서

시스템 구성





1. 질의응답 시스템

- 질문 분석기 : 사용자의 질문을 분석해서 키워드를 추출하고 핵심적이라고 판단되는 단어를 임베딩 검색기로 전달한다.

- 임베딩 검색기 : 질문 분석기가 추출한 단어에 대한 임베딩 파일이 데이터베이스에 저장되어 있는 지 검색한다. 임베딩 파일이 있으면 질문 임베딩 후 키워드 임베딩 파일에서 답변 영역을 예측하며, 없으면 문서 검색기를 통해서 정보를 검색한다.

- 문서 검색기 : 위키피디아에서 키워드에 대한 정보를 검색하고 문단 단위로 분할하여 언어모델에 입력한다.

- 언어모델 : 사용자 질문과 위키피디아 문서에 대한 임베딩을 생성한다. 베이스 모델로 다국어 BERT 모델을 사용한다.

- 기계독해 : 사용자 질문에 대한 임베딩과 위키피디아 문서에 대한 임베딩을 이용해서 답변 영역을 예측하여 사용자에게 전달한다.

2. 강의 웹 서비스 서버

- 메인 페이지 : 사용자 로그인과 권한을 검증하고 사이트에 등록된 강의와 사용자의 최근 학습 강의를 목록화한다.

- 로그인 관리 : 이메일 검사를 통해 학교 관련자만 회원가입 가능하도록 하며 기본적인 로그인, 자동 로그인을 지원한다.

- 내 강의목록 페이지 : 사용자가 수강하는 전체 강의 목록을 확인하고 강의 진도 등 강의에 대한 정보를 표시한다.

- 강의 대시보드 페이지 : 강의 대시보드를 통해 사용자가 등록한 강의에 대한 정보를 확인하는 기능을 제공한다.

- 강의 동영상 페이지 : 동영상 강의를 재생하고 챗봇 UI를 제공한다.

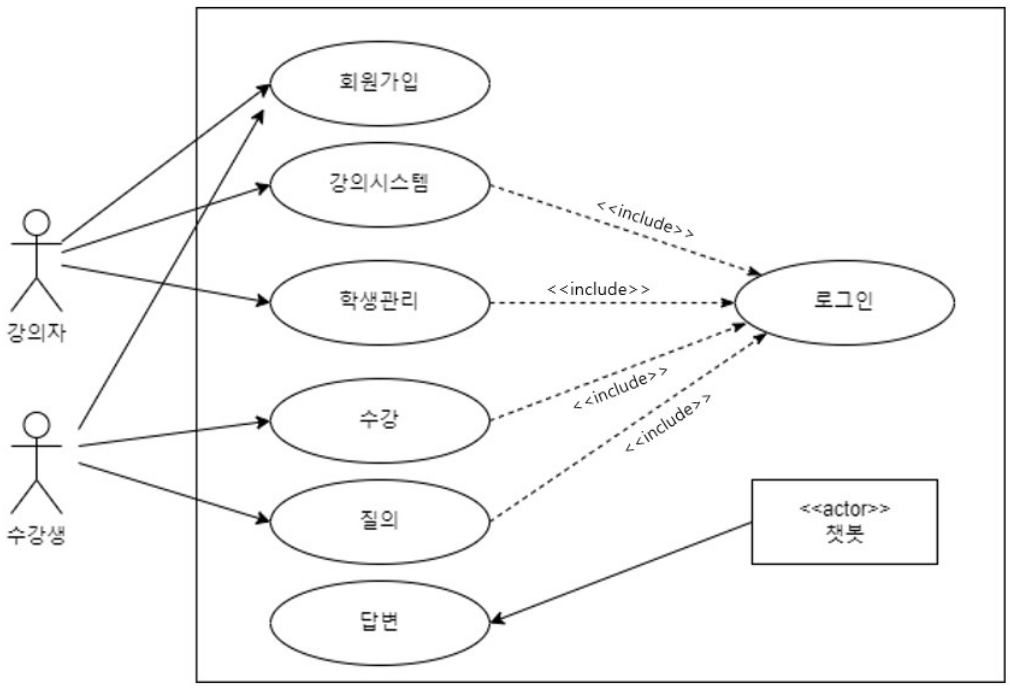
- 챗봇 채팅 : 사용자는 챗봇 UI를 통해서 질의응답을 수행할 수 있다.

- 강의 관리자 페이지 : 강의 관리자는 강의와 관련된 정보를 입력, 변경하거나 동영상을 업로드 할 수 있다.

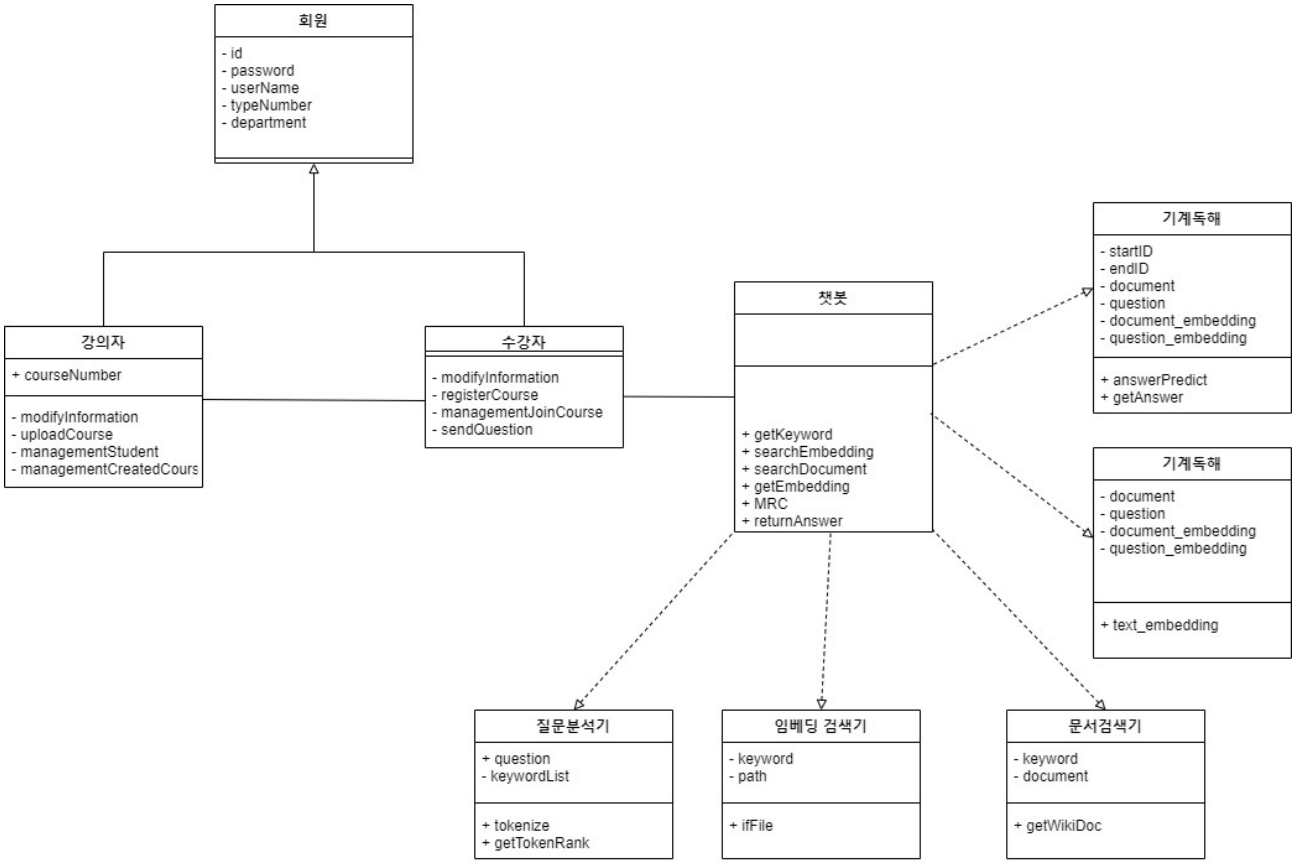
- 회원 데이터 처리 : 사용자가 회원가입을 할 때 입력하는 id, password를 데이터베이스에 암호화하여 저장하고 가입자의 소속을 분류한다.

- 강사 데이터 처리 : 강사는 자신의 강의를 업로드하고 키워드를 입력하여 질의응답 성능을 올릴 수 있다.

유스 케이스

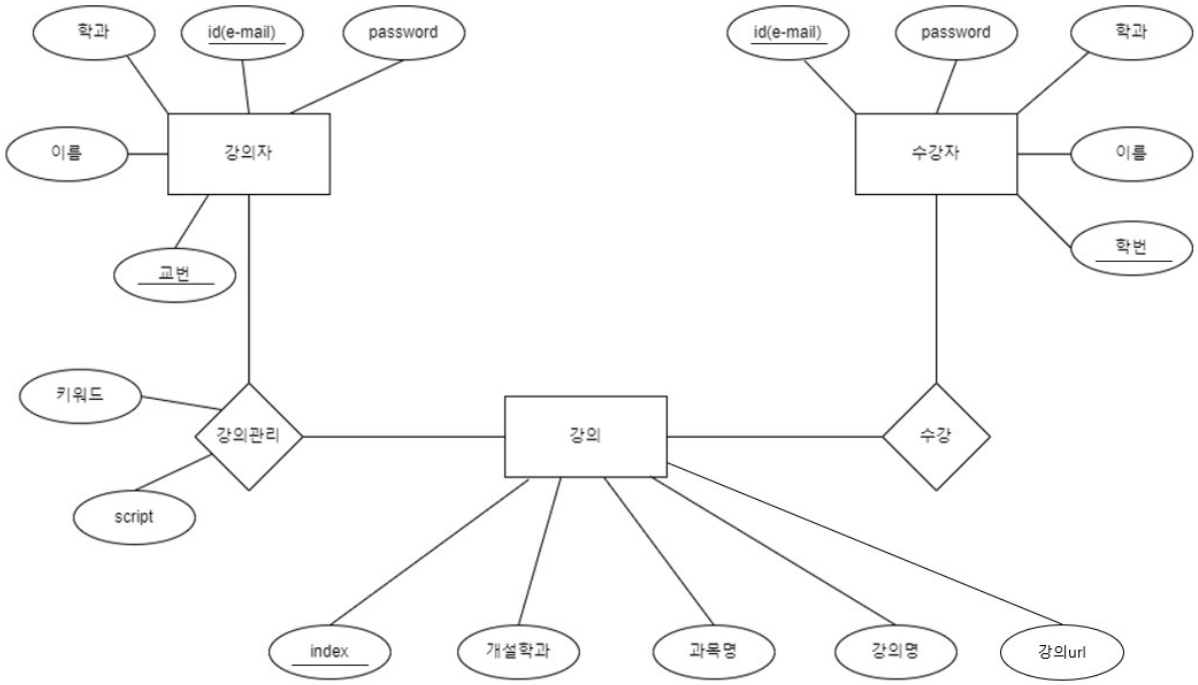


클래스 다이어그램



- UI 분석/설계 모델/프로토타입

E-R 다이어그램



테스트 계획서

NO	대분류	중분류	소분류	테스트 사항	Pass Result	비고
1	Front-end & Back-end	웹 페이지	메인화면	메인화면 접속 시 등록된 강의 목록이 정상적으로 출력되는 지 확인	디자인한 웹 페이지 동작	
2			로그인	로그인 화면 접속 후 id, pw 입력 시 로그인이 정상적으로 되는 지 확인	로그인 성공, 실패	
3			사용자 강의목록	로그인 후 내 강의 목록 페이지 접속 시 전체 강의 목록이 정확하게 표시되는 지 확인	사용자 강의 목록화	
4			강의 대시보드	강의 대시보드 접속 시 사용자가 등록한 강의에 대한 정보가 정상 출력되는 지 확인	수강한 강의 정보 표시	
5			강의 관리자 페이지	동영상 업로드와 키워드, 스크립트 입력이 정상적으로 되는 지 확인	동영상, 키워드, 스크립트 저장	
6		강의 동영상 페이지	동영상	강사가 업로드한 동영상, 링크 등이 동영상 플레이어에서 정상적으	동영상 재생	

				로 동작하는 지 확인		
7			챗봇UI	사용자 질문이 서버로 정상적으로 전달되어 답변을 출력하는 지 확인	질문 입력, 답변 출력	

NO	대분류	중분류	소분류	테스트 사항	Pass Result
1	챗봇	질문 분석기	문장 tokenize	문장을 tokenize할 때 검색 가능한 의미 단위로 분할하는 지 확인	위키피디아에서 검색 가능한 키워드 생성
2			임베딩 vocab 조회	임베딩 vocab을 조회해서 키워드에 대한 임베딩이 있는 지 정상적으로 검색하는 지 확인	임베딩 파일이 있는 키워드 추출
3		임베딩 검색기	강의 웹 서버 데이터베이스와의 연동	강의 웹 서버에 저장된 강의 스크립트를 정상적으로 가져올 수 있는 지 확인	강의 웹 서버 테이터베이스의 강의 스크립트 획득
4			키워드 검색	임베딩이 저장된 디렉터리에 접근해서 정상적으로 임베딩 목록을 가져올 수 있는 지 확인	키워드 임베딩 파일 검색
5		문서 검색기	동음이의어 처리	동음이의어에 대한 처리를 통해 관련된 키워드를 정확하게 추출하는 지 확인	복수의 동음이의어에서 단어 선정
6			위키피디아 검색	키워드를 위키피디아에서 정상적으로 검색을 할 수 있는 지 확인	키워드에 대한 위키피디아 문서
7			문단 분할	가져온 문서를 소주제, 문단에 따라 분할하는 지 확인	문서 소주제, 문단 단위로 분리
8		언어 모델	임베딩 길이 제한	토큰의 수가 모델에서 처리 가능한 수를 넘지 않는 지 확인	한 번에 모델에 입력되는 토큰 수를 512개 이하로 제한
9			사용자 질문 임베딩	사용자가 입력한 질문이 언어모델에서 임베딩을 생성하는 지 확인	질문 임베딩 생성
10			문서 임베딩	문서를 문단 단위로 임베딩을 생성하여 ‘키워드명’/‘문단주제’.embedding으로 저장되는 지 확인	문단 단위 임베딩 파일 생성
11		기계독해	답변 영역 예측	질문 임베딩, 문서 임베딩을 통해서 답변 영역을 찾을 수 있는 지 확인	문서 내 답변 영역의 시작, 끝 토큰 인덱스 예측
12			답변 길이 제한	답변을 챗봇 UI에 출력했을 때 가독성을 유지할 수 있는 길이인 지 확인	답변 길이 제한

2. 프로젝트 수행을 위해 적용된 추진전략, 수행 방법의 결과를 작성하고, 만일 적용과정에서 문제점이 도출되었다면 그 문제를 분석하고 해결방안을 기술하시오.

	문제해결을 위해 적용한 방법(또는 기법) 결과, (문제점, 해결방안)
	- 애자일 기법을 통한 유연하고 잦은 회의를 통해 속도감 있게 개발한다. 주 1회 회의를 통한 피드백과 개발
문제점	- 코로나19 감염증으로 인한 컨디션 저하로 인해 개발의 속도가 늦어졌다. 주 1회 회의를 하지 못하고 연락이 뜸해지는 상황. 타 과목 시험 기간으로 인해 캡스톤에 집중 할 수 없어 개발의 공백.
해결방안	- 집중이 덜 되더라도 화상회의를 통한 피드백, 잦은 화상회의를 통한 서로 동기부여.

	데이터 베이스 설계의 문제
문제점	- 데이터베이스의 설계에서 E/R 다이어그램 작도에서 강의자와 수강자, 강의 Entity set 사이에 루프가 생겨 redundancy의 원인이 될 가능성이 생김,
해결방안	- 집중이 덜 되더라도 화상회의를 통한 피드백, 잦은 화상회의를 통한 서로 동기부여.

문제점	- 따라서 이를 해결하기 위해 강의자와 수강자를 잇는 직접적인 relation을 삭제하고, 강의 Entity set를 사이에 둔 relation으로 강의자와 수강자의 relation을 표현해 루프, redundancy 발생을 예방함.

문제점	- 많은 BERT 기반의 언어모델은 기계독해 테스크에 대한 베이스라인 코드를 제공하지만 질문과 문서가 저장된 파일을 불러와서 처리한 후 다시 파일에 답변을 저장하기 때문에 실시간 채팅 서비스에 그대로 적용하기에 어려운 점이 있다. 따라서 모델 내에서 임베딩을 수행하는 BERT와 답변 예측을 하는 FFNN을 분리하거나 임베딩만 추출하는 부분만 사용하고 FFNN을 따로 만들어서 연결시켜야 하는데 소스가 복잡하고 관련 지식이 부족하여 필요한 기능에 대한 코드를 찾는 것에 많은 어려움이 있다.

문제점	- 기존에 학습한 모델이 tensorflow ckpt로 저장되었지만 작업을 원활하게 수행하기 위해서 keras h5 또는 pytorch 모델로 변경해야 한다. ckpt를 pytorch 모델로 변환하는 코드를 사용했지만 변환 과정에서 알 수 없는 오류가 발생하고 있으며, 오류를 예외처리하고 만들 경우 임베딩 성능이 매우 떨어지는 결과를 보이고 있다. 현재 h5모델로 저장하는 방법을 탐색하고 있다.

캡스톤 디자인 I 중간보고서 채점표

평가도구	평 가 항 목	평 가 점 수				
		1	2	3	4	5
중간 보고서 및 실행 결과	1. 요구사항 정의서(기능, 성능, 인터페이스 등)가 구체적으로 작성되었는가?					
	2. 요구분석, 설계 산출물(모델, 프로토타입 등)의 내용이 충실한가?					
	3. 설계 및 구현 문제를 위해 적용한 이론, 문제해결 방법이 제시되었으며 그 적용이 적합한가?					
	4. 구현된 소프트웨어(또는 이와 동등한 하드웨어 시스템)가 버그 없이 실행되었는가?					
	5. 구현된 소프트웨어(또는 이와 동등한 하드웨어 시스템)의 성능 요구사항은 충족되었는가?					
도구활용	6. 설계 및 구현을 위해 도구가 적절히 활용되었는가?					
	7. 도구의 활용수준(능숙도)은 프로젝트 수행에 적합한가?					
팀원의 업무 및 역할	8. 팀원의 업무분담에 따른 역할 및 협력이 충실히 이루어졌는가? (평가자에 의한 질의)					
	9. 프로젝트 중간 진척상황에 대해 팀원이 충분히 인지하고 있는가?(평가자에 의한 질의)					
합계						

***검토 의견(최종완료 때까지 보완해야할 점에 대해 작성 요망)**

심사위원(소속): **(이름)** **(인)**