

나의 학습을 똑똑하게 도와주는

Study-mate

파이썬기반 딥러닝

- 인공지능학과 이재욱

Contents

- 1. 프로젝트 개요
- 2. 프로젝트 구성도
- 3. 데이터셋
- 4. 데이터 전처리

- 5. 모델 소개
- 6. 성능 평가 방안
- 7. 개발 일정
- 8. 활용 방향

Project - overview

프로젝트 배경 및 목적

과거에는 전공책과 노트를 들고 다니며 공부했지만, 이제는 대부분의 학습 자료가 PPT 형식의 디지털 강의자료로 대체되었다.

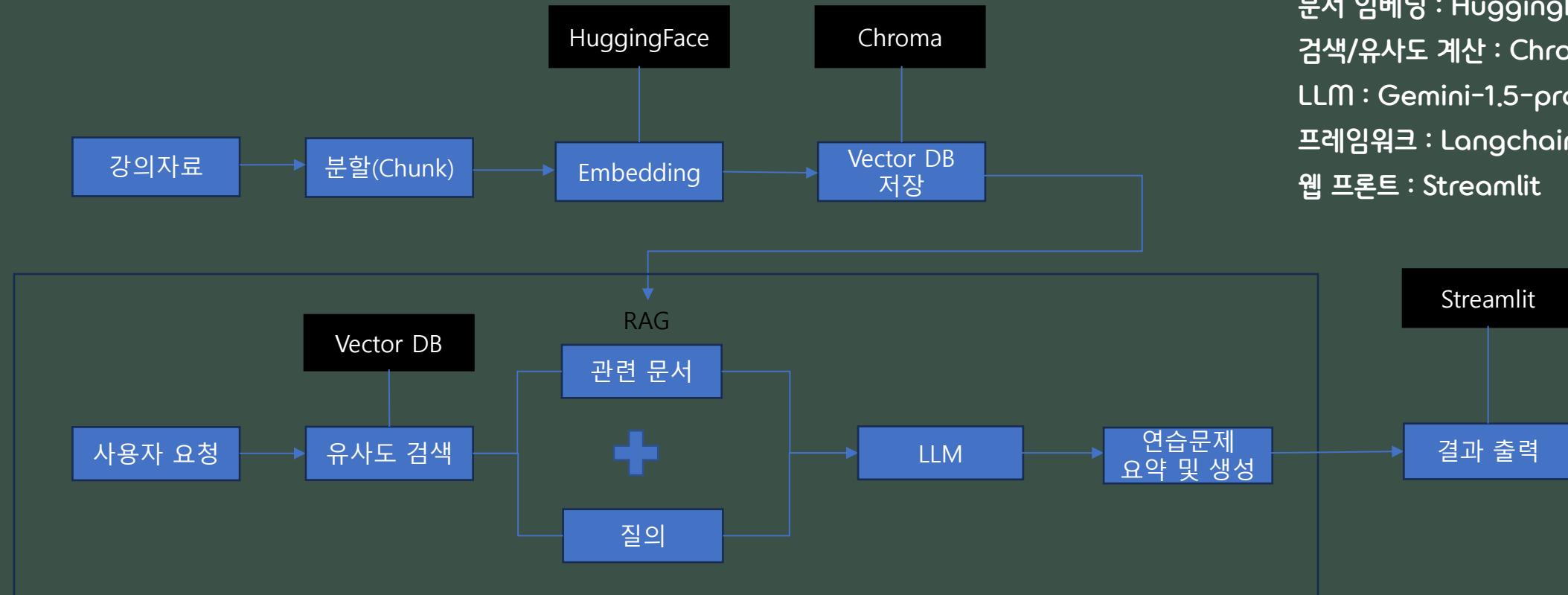
이러한 흐름 속에서, 파일을 업로드하기만 하면 강의자료를 자동으로 요약하고, 연습문제를 생성해주는 챗봇형 학습 도우미가 있다면 학생들의 학업 효율성과 자기주도적 학습 역량을 크게 높일 수 있을 것이다.

study-mate는 바로 이러한 필요에서 출발했다. Study-mate는 강의자료(PPT, PDF 등)를 분석해 핵심 개념을 간결하게 요약하고, 주요 내용을 기반으로 퀴즈나 연습문제를 자동 생성하는 시스템이다.



美대학, 교육에 생성형 AI 접목… “챗GPT” 명지대학교, 학생 주도 RAG 기반 AI 정적 측면 활용 챗봇 ‘마루에그봇’ 개발로 입학처 상담 효율화

System Architecture Diagram



사용 기술 스택

문서 임베딩 : HuggingFace
검색/유사도 계산 : Chroma Vector DB
LLM : Gemini-1.5-pro
프레임워크 : Langchain
웹 프론트 : Streamlit

Experiment Process

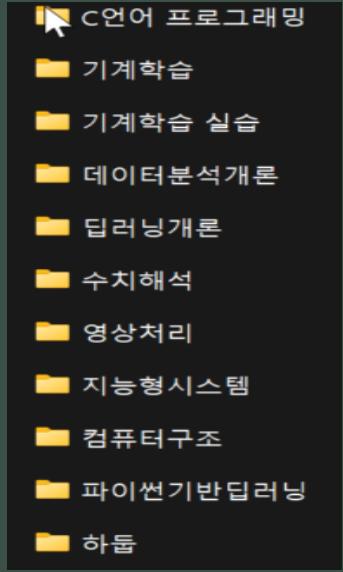
실험 과정 예시

1. 강의자료 업로드(Input)
- PDF
2. 텍스트 추출(Text Extraction)
- PyMuPDF를 이용하여 슬라이드별 텍스트 추출
3. 청크화(Chunking)
- 텍스트를 문맥 단위로 분할
- 중첩(overlap) 적용으로 문맥 손실 최소화
4. 임베딩(Embedding)
- HuggingFace 모델로 각 Chunk를 벡터화
- 의미 기반 수치 벡터 생성
5. 벡터 저장(Vector Storage)
- Chroma Vector DB에 임베딩 벡터 및 원문 텍스트 저장
6. 사용자 질의(User Query)
- 사용자가 질문 또는 요청 문장 입력
7. LLM 입력 구성(Prompting)
- 검색된 문맥을 프롬프트에 포함하여 LLM에 전달
8. 결과 생성
- LLM이 요약문 및 연습문제 자동 생성

입력	Chunk	Embedding	Vector DB	사용자 요청	응답
PDF	["AI는 인공지능을 의미하며, 최근 Transformer 모델이 각광받고 있다...", "Transformer는 Attention 메커니즘을 사용해 문맥을 이해한다...", ...]	[[0.012, -0.234, 0.456, ..., 0.089], [0.034, -0.111, 0.520, ..., 0.100], ...]	벡터 리스트	"강의내용을 요약하고 연습문제 생성 해줘."	요약: 본 강의는 AI 발전과 RAG 기반 요약 시스템을 다룹니다. 문제 1: RAG의 의미는 무엇인가? 1) Retrieval-Augmented Generation 2) Random Access Generator

Dataset

Data: 실제 다양한 전공 분야의 대학교(대학원) 강의자료(PPT, PDF)



Aa	과목명	교수	강의계획서
	사이버보안개론	도인실	사이버보안개...
	정보시스템보안	양대현	정보시스템보...
	소프트웨어공학	오소연	소프트웨어공...
	블록체인응용	김중길	블록체인응용...
	빅데이터응용	이민수	빅데이터응용...
	빅데이터보안	애호	빅데이터보안...
	정수론	황지현	정수론_황지현...

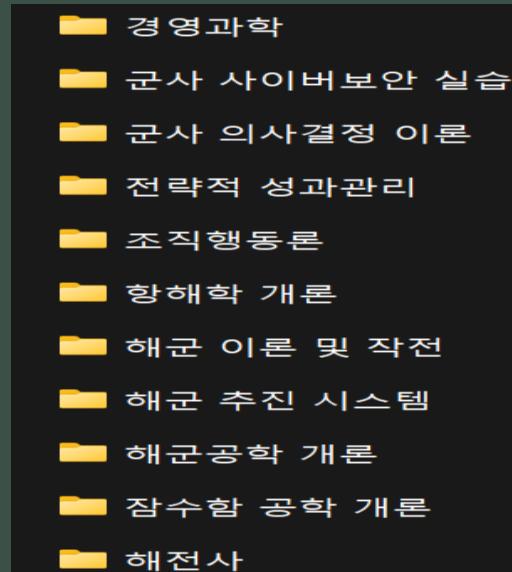
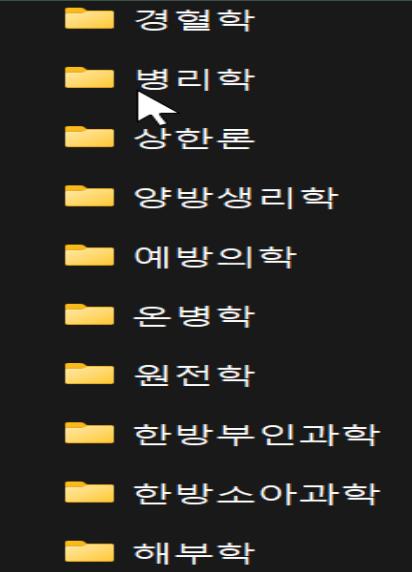
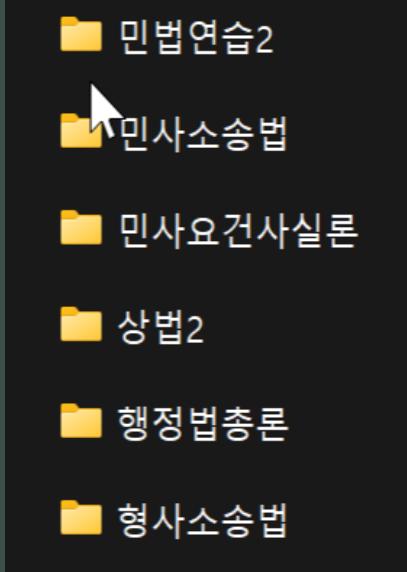
세종대학교
인공지능학과

이화여자대학교
사이버보안학과

인하대학교
법전문대학원

경희대학교
한의학과

해군사관학교



Data Preprocessing

데이터 전처리

1. 파일 입력 및 텍스트 추출

입력 데이터 : .ppt, .pdf 등 강의자료 파일

Python-pptx 또는 PyMuPDF 라이브러리를 활용하여 슬라이드 / 페이지별 텍스트 추출

슬라이드 제목, 본문 텍스트, 표나 리스트 등 다양한 형식을 모두 텍스트화

4. 불용어(Stopwords) 처리 & 토큰화

- 의미 없는 단어 제거("the", "is", "그리고" 등)
- 필요시 형태소 분석(KoNLPy 등)을 활용해 한국어 명사 중심으로 정리
- 영어 PPT의 경우 spaCy 또는 NLTK 활용

2. 텍스트 정제(Cleaning)

추출된 텍스트에서 불필요한 요소를 제거

- 공백, 특수문자 정리
- "Slide 1", "Page 1"과 같은 정보 제거
- 중복 문장 제거

정제된 문장은 모델이 읽기 쉽도록 문단 단위로 정돈

5. 임베딩(Embedding) 전 준비

- 전처리된 텍스트를 문서단위로 리스트화
 - 각 문서에 슬라이드 제목, 순서, 주제 등 메타데이터 추가
- * 추후 유사도 검색 시 문맥 정보까지 함께 검색 가능

3. 문단 분할(Chunking)

- 긴 텍스트를 LLM이 처리 가능한 길이로 분할
- 일반적으로 300~500 토큰으로 나눔
- 문맥 손실을 줄이기 위해 중첩(overlap) 적용

models

LLM(Large Language Model)

사용 모델 : Gemini 1.5pro

역할

- 사용자가 업로드한 강의자료 내용을 기반으로 요약 생성
- 핵심 개념을 분석해 연습문제(객관식/주관식) 자동 생성
- 검색된 문맥을 참조하여 정확하고 일관된 응답 제공

Vector Database

사용 모델 : Chroma

역할

- 임베딩된 벡터를 저장하고, 사용자의 질문과 유사도가 가장 높은 문단 검색
- RAG 구조에서 검색 단계를 담당

선택 이유

- 빠른 검색 속도
- LangChain과 완벽히 호환

Embedding Model

사용 모델 : HuggingFace Sentence Transformer

역할

- 강의자료의 문장/문단을 의미 벡터로 변환
 - 변환 후 사용자의 질문과 의미적으로 가까운 문서 검색에 활용
- 선택 이유
- 가벼우면서도 빠르고 정확한 임베딩 성능
 - 영어와 한국어 혼용 데이터에 대응 가능

Framework : LangChain

역할

- Embedding → VectorDB → LLM단계를 자동으로 연결
- RAG 구현의 핵심 프레임워크
(데이터 로딩 → 전처리 → 검색 → 생성)의 과정을 하나의 체인으로 관리

선택 이유

- 간결한 코드로 복잡한 RAG 파이프라인 구축 가능
- 다양한 LLM과 손쉽게 연결

Web Framework : Streamlit

역할

- StudyMate의 사용자 인터페이스 구현
- 파일 업로드, 요약 결과, 문제 출력 화면 구성

Evaluation Plan

평가 목적

Study-Mate가

1. 강의자료를 얼마나 정확하게 요약하는지
2. 생성된 연습문제가 얼마나 적절하고 유용한지
3. 전체 시스템이 얼마나 빠르고 안정적으로 동작하는지
4. 다양한 분야의 전문용어, 다양한 언어(한문, 영어, 한글)를 정확하게 표현하는지

요약 품질

평가 지표 : ROUGE-L

- 원문과 요약문 간의 문장 유사도, 핵심정보 포함도 측정

평가 기준

- 0.7 이상이면 좋은 요약 품질

검색 정확도

평가 지표 : Precision@k, Recall@k

- 질의 시 상위 k개의 문단이 관련 문서인지 평가

응답 속도

평가 지표 : Processing Time

- PDF 업로드 → 결과 출력까지 평균 처리 시간

Development Schedule

1 week

Preprocessing

- 텍스트 추출 코드구현(PyMuPDF)
- 불필요한 텍스트 제거 및 청크화 수행
- 임베딩 테스트 및 Chroma Vector DB 구축

2 ~ 3 week

Model Intergration

- HuggingFace 임베딩 모델 적용 및 파라미터 튜닝
- RAG 구조 완성
- Gemini 1.5 pro API 연동
- Prompt Template 설계

4 week

웹 인터페이스 개발 (Frontend / Streamlit)

- Streamlit 기반 UI 개발
- 파일 업로드 / 요약결과 출력 / 문제 표시 기능 구현
- LLM 호출 및 결과 표시 연결 테스트
- 사용자 입력 품 추가

5 week

Evaluation & Optimization

- 평가 지표 (ROUGE-L, Precision@ k , Recall@ k) 계산
- 처리 속도 평균 분석
- 결과 피드백 기반 하이퍼파라미터 튜닝
- Prompt 개선 및 사용자 의견을 반영하여 인터페이스 개선

시스템의 확장성 및 기대효과

1. 시스템의 확장성(응용 가능성)

1-1) 데이터 확장 측면

다양한 입력 형식 지원 :

현재의 계획은 PDF 중심이지만, 추후 음성 녹음 파일과 같은 다양한 입력형식으로 확장
(음성 강의를 자동 텍스트 변환(SST) 후 요약까지 지원)

멀티모달 확장 가능성 :

- 이미지와 도표가 포함된 슬라이드에서도 OCR을 이용하여 텍스트 추출
- 수식과 그래프까지 인식 가능한 모델(LayoutLM)로 확장 가능

1-2) 기능 확장

맞춤형 문제 생성 기능:

사용자의 수준에 맞춰 난이도 조절된 문제 자동 생성

대화형 복습 모드:

사용자가 LLM과 대화하면 양방향적 학습 가능

1-3) 기술의 확장 측면

모델 교체 유연성:

LangChain 구조 덕분에 Gemini 뿐만이 아닌 헷-5, Claude 등으로 쉽게 교체 가능

사용자 데이터 기반 개인화:

- 사용자의 이전 학습 기록 기반으로 자동 요약 범위 조정
- 사용자가 약한 슬라이드(범위) 자동 추천 기능

일정이 허락하는 기간 내에 가능한 확장은 이번 프로젝트에서 구현 예정

시스템의 확장성 및 기대효과

2. 기대효과

2-1) 학습 효율 향상

- 사용자들이 긴 강의자료를 짧은 시간에 핵심만 복습 가능
- 수동 요약/정리 시간을 70%이상 절감
- 학업 부담 완화

2-2) 자기주도 학습 강화

- 사용자가 직접 질문 -> 요약/문제 생성 과정을 반복하며 스스로 학습을 설계하는 능력 향상
- 학습 참여도와 이해도 부분에서 향상 효과 기대

2-3) 교육자 지원

- 강의자료를 자동 요약하여 수업 준비 시간 단축
- 자동 생성된 연습문제 활용으로 평가 및 쿠즈 제작 효율 향상

2-4) 확장 가능성

- 자격증, 공무원 시험, 기업 직무교육 등 다양한 도메인 학습 자동화 가능
- RAG 구조를 그대로 유지하면서 도메인 데이터셋만 교체하면 새로운 분야로 확장

2-5) 비대면 온라인 교육 환경 대응

- COVID-19 팬데믹과 같은 비대면 온라인 교육환경에서도 큰 효과

reference

- [1] M. Lee and W. Jeon, "A Study on System Development Using LangChain and Attribute-Based Sentiment Analysis Model: Focusing on the Real Estate Domain," J. Korean IT Service Society, vol. 24, no. 1, pp. 41-59, 2025.
- [2] E. Lee and H. Bae, "A Survey on the Latest Research Trends of Retrieval-Augmented Generation (RAG) Technology," Journal of the Korean Institute of Information Processing Systems, vol. 13, no. 9, pp. 429-436, 2024.
- [3] C. Jung, "Implementation of Generative AI Services Using LLM Application Architecture: Based on RAG Model and LangChain Framework," Journal of Intelligence and Information, vol. 29, no. 4, pp. 129-164, 2023.
- [4] G.-M. Choi, "Building and Utilizing a Retrieval-Augmented Generation (RAG) Based Customized Chatbot System for Enterprises," The Journal of the Korea institute of electronic communication sciences, vol. 10, no. 6, pp. 1281-1292, Dec. 2024.



감사합니다