SK네트웍스 **Family AI**과정 **6**기

데이터 수집 및 저장 프로젝트 기획서

◘ 개요

- 산출물 단계 : 데이터 수집 및 저장
- 평가 산출물 : 프로젝트 기획서
- 제출 일자 : 1/31
- 깃허브 경로 : https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN06-FINAL-3Team
- 작성 팀원 : 박서윤, 박유나, 유경상, 장예린

프로젝트 주제	**화장품 회사 내부정보 검색 및 경쟁사 정보 제공 시스템**
문제정의	**문제점** ● 화장품 관련 사내문서가 방대하고 필요한 정보를 신속하게 찾기 어려움. ● 기존 검색 시스템이 특정 요구사항(영어 및 전문용어)을 충족하지 못하는 경우 검색 불가. ● 사내 제품들에 대한 성분 및 효과, 유사한 성분을 가진 제품들을 한번에 검색 및 파악이 어려움. **목표** ● 내부문서 검색시스템을 구축하여 화장품회사직원들(특히 비전문 부서, ex: 영업/마케팅팀, 회계팀, 법무팀)이 쉽게 정보를 검색하고 활용할 수 있도록 함. ● 시스템이 화장품 성분용어에 특화된 도메인 생성 ● 한글 검색으로도 검색 용이 ● 화장품의 성분 및 효과, 유사한 성분을 가진 제품들을 한번에 검색 ● 회사 내부문서(재무정보, 매출, 광고, 브랜드개요 등)을 제공 ● 화장품 관련 규제 및 원료 검색 및 정보제공

시장조사	**국내외 화장품회사의 내부문서 검색 시스템 조사** ● 국내 화장품회사의 AI기반 내부문서 검색시스템 부재 ● 현재 내부문서 검색 기능이 갖춰진 시스템이 아니라, 문서 보안과 관리에 초점이 맞춰진 시스템만 운영되고 있음. **국내 화장품 회사에서 활용하는 검색시스템의 한계점 조사.** ● 전문용어에 취약한 부서 직원들은 검색 및 제품이해에 취약 ● 외부 API연동을 이용한 검색시스템으로 보안에 취약 ● 화장품별 성분별 화장품 및 성분 목록을 한번에 제공함으로써 가시성과 업무 효율을 높임. ● 브랜드 가치, 제품 정보, 재무 정보 및 실적 등 다양한 내부 정보를 통합하여 제공하는 시스템이 부재 ● 대부분의 기업이 내부 문서를 활용한 검색 시스템이 아닌, GPT 기반의 단순 질의응답 기능을 이용하는 수준에 머물러 있어, 심층적인 정보 탐색 및 활용이 제한적임.
시스템 구성	**기능** ● 최근 산업동향 및 뉴스 제공 ● 자연어처리 검색(한글/영문 검색 키워드 지원) ● 성분 용어에 대한 설명 및 효과 제공. ● 성분별 제품 목록, 제품별 성분 목록 제공 ● 재무/회계 정보 제공 ● 회사 브랜드 및 주력 상품 소개 ● 경쟁사 제품 설명 및 요약 ● 개인별 제품, 성분 즐겨찾기 기능 ● 시스템 내 사원관리 및 사원추가 기능 ● 회사-외부 업무 보고서 및 이메일 서포트 기능 **구성 요소** ● 데이터 수집 및 저장 시스템. ● 검색 및 NLP 기반 질의응답 엔진. ● 사용자 인터페이스(UI) 및 시각화 대시보드.

모델링 방안	**1. 사내 서버 기반 LLM 구축** ● BioBERT + LoRA,Quantization 적용 ○ 화장품 성분, 효과, 법규 등의 전문 용어를 반영한 BioBERT 기반 도메인 특화 모델 구축 ○ 사내 서버에서 경량화된 LoRA(Low-Rank Adaptation) 적용하여 효율적인 학습 수행 ● vLLM을 활용한 고속 Inference 최적화 ○ vLLM을 적용하여 대규모 언어 모델의 추론 속도를 최적화 ○ 메모리 효율적인 텐서 병렬 처리 및 동적 배치 관리로 빠른 응답 제공 ● NER 기반 검색 최적화 ○ 화장품 성분, 화학 용어, 사내 용어에 대한 Named Entity Recognition(NER) 파인튜닝 적용 **2. 검색 최적화 모델 (Hybrid Search)** ● Retriever 분리 설계 ○ 문서 유형에 따라 다중 리트리버 구조 적용 ■ 텍스트 문서: BM25 + Dense Retrieval (SBERT) ■ 이미지 및 멀티모달 데이터: CLIP 기반 검색 ■ 규제 및 원료 정보: 키워드 중심의 Rule-based Retrieval ○ 검색 질의 유형에 따라 최적의 리트리버 조합을 선택하여 검색 성능 개선 ● BM25 (통계적 검색) + Dense Retrieval (SBERT, CLIP) 결합 ○ 키워드 기반 BM25와 문맥을 이해하는 Dense Retrieval 모델(SBERT, CLIP) 혼합 ○ SBERT를 활용한 문맥 기반 검색 및 유사 문서 추천 ○ CLIP을 적용하여 이미지 및 멀티모달 검색 지원 **3. 데이터베이스 및 검색 엔진** ● 벡터스토어 분리 설계 ○ 데이터 유형별 벡터스토어 분리 적용 ■ 문서 벡터 DB: FAISS (사내 문서, 연구 자료) ■ 성분 및 제품 벡터 DB: FAISS (화장품 성분, 제품 정보) ■ 이미지 벡터 DB: CLIP (제품 패키징, 성분 구조 이미지) ○ 검색 질의에 따라 적절한 벡터스토어를 선택하여 검색 최적화 ● FAISS + MariaDB ○ FAISS(Vector DB): 문서 및 성분 데이터를 벡터화하여 빠른 검색 지원 ○ MariaDB: 검색 메타데이터 및 사내 문서 관리 ● LangChain + BioBERT + vLLM 기반 검색 엔진 개발 ○ LangChain을 활용한 자연어 검색 시스템 구축

	○ vLLM 기반 최적화된 BioBERT 추론을 통해 화장품 및 생명과학 분야 특화 검색 성능 개선 **4. 화장품 성분 및 문서 데이터 학습** ● 화장품 성분 및 효과 데이터셋 구축 ○ 기존 제품 및 경쟁사 제품의 성분 정보를 포함한 임베딩 DB 생성 ○ 성분 간 유사도를 벡터화하여 성분 추천 시스템 개발 ● 규제 및 원료 정보 자동화 업데이트 ○ 국내외 규제 데이터 API 연동 및 최신 원료 정보 반영
사용데이터	● 아모레퍼시픽 (기업 홈페이지에서 데이터 수집 및 자체 생성) ○ IR 및 기타 보고서 ○ 브랜드 ○ 화장품 ● 아모레퍼시픽몰: 자사 화장품 ● 올리브영: 자사 및 타사 화장품 ● 원료별 공식 명칭, 용도 주의사항 ● 식약처 ○ 원료성분 정보 및 기능성화장품 성분 리스트 ○ 화장품 사용제한 원료 정보 ○ 화장품 표시, 광고 관리 지침 ○ 기능성화장품 심사 관련 QnA ○ 2024년 화장품 제도 변경사항 ● 국가법령정보센터 ○ 화장품법 ○ 기능성화장품 성분 리스트 ○ 화장품 안전기준 등에 관한 규정 ○ 화장품 색소종류 기준 및 시험방법 ● 한국보건산업진흥원 화장품 관련 통계데이터 ● 뉴스/저널/학술지
R&R	● **PM:** 프로젝트 전반적인 일정 및 진행상황 관리 ● **데이터 엔지니어:** 데이터 수집,정제 및 전처리 ● **ML 엔지니어:** NLP 모델 설계 및 학습, 자체적인 검색 및 문서 요약 LLM 모델 개발 ● **프론트엔드 개발자:** UI 설계 및 개발 ● **백엔드 개발자:** 검색엔진 및 API 설계 및 구현, DB관리 ● **QA 엔지니어:** 시스템 안정성,정확성 테스트 및 결함 수정 지원