RAG 파이프라인 구축 프로세스 보고서

서비스 명 및 개요	서비스명: 마트 쇼핑 도우미 챗봇 (SmartShopper) 서비스 개요 **1.** 서비스가 해결하려는 문제점**(Problem Statement)** 소비자가 마트에서 상품의 가격 정보를 비교하고 합리적인 결정을 내리는 데 드는 시간과 노력을 절감. **2.** 제공하고자 하는 주요 기능 및 가치 실시간 가격 비교 및 가성비 평가 할인 정보 제공 개인 맞춤형 추천 **3.** 타겟 사용자와 예상 사용 시나리오 타겟 사용자: 가격 민감도가 높은 일반 소비자 및 주부 예상 사용 시나리오: 사용자가 특정 상품의 가격을 입력하거나 검색하면, 최적의 가격과 가성비 분석 결과를 제공
타겟 사용자 및 시장 분석	타겟 사용자와 예상 사용 시나리오 예상 사용자 유형: 일반 소비자, 쇼핑 애호가, 주부, 대학생 등. 주요 요구사항 및 사용 목적 - 실시간 가격 비교 - 합리적인 구매 결정을 위한 가성비 분석
목표 및 기대효과	서비스 목표**:** 1. 사용자 질문에 신속하고 정확한 답변을 제공하여 고객 문제 해결 시간을 단축합니다. 2. 제시된된 가격보다 저렴하게 구매할 수 있는 추천 구매처를 제공합니다. 기대효과**:** 1. 고객은 간단한 구매정보 입력만으로 전체 매매가를 탐색하지 않고도 필요한 정보를 얻을 수 있습니다. 이에 따라 고객의 현명한 소비를 가능하게 합니다. 2. 최저가인 추천 구매처를 소개하여 사용자의 구매를 촉진합니다.
데이터 구성 및 활용	공공데이터 [한국소비자원_생필품 가격 정보_20241220] 활용 웹크롤링을 활용한 풍부한 데이터셋 구축

RAG 파이프라인 구축 프로세스 보고서

RAG 파이프라인 설계	데이터 최적화 : - Chunk Size: 100 - > 가격을 물어보는거라 큰 사이즈가 필요없음 - > 작은 청크 사이즈 : 높은 세밀도, 느린 검색 성능 - > 큰 청크 사이즈 : 낮은 세밀도, 빠른 검색 성능 - Overlap: 100 벡터 데이터베이스 구축 및 임베딩: - 벡터 DB : Pinecone - 임베딩 모델 : Open API `text-embedding-3-large` **Retriever** 및 **Reranker** 구현: - VectorStore Retriever 이용 - 하이퍼파라미터 튜닝 : - 반환할 문서 수(k) = 5 -> 성능 분석 후 변경될 수 있음 -> 보통 질문이 단순한 경우 작은 값, 복잡한 경우 큰 값을 사용함 - 유사도 임계값 = 0.7 -> 성능 분석 후 변경될 수 있음 - (optional) Reranker : - CrossEncoderReranker 를 활용해 Retriever 의 후보 문서를 재정렬 - 적용 문서 수: 상위 10 개 문서 **LLM** 프롬프트 설계 및 답변 생성, 평가 : 1/ Task : QA / 챗봇 2/ 프롬프트 예시 `prompt = ChatPromptTemplate.from_messages( [ ( "system", """ 너는 인공지능 챗봇으로, 주어진 데이터를 분석해서 소비자가 구매하고 싶은 물품을 검색해서 현재 판매가격과 할인 또는 원플러스원 물품으로 파는곳을 제공해줘. 데이터에 있는 내용으로만 답하고 내용이 없다면, 잘 모르겠다고 답변해. --- CONTEXT: {context} """, ), ("human", "{input}"), ] )` **3/** 답변 생성 **: Open API Assistant** **4/** 답변 사후 평가 **: Upstage Groundness Check API** - Upstage Groundness Check API 로 3 번 검증하여 Hallucination 방지
RAG 파이프라인 평가 및 결과	평가방법 - 정량 평가 : R A G A S 평가 지표 - context_precision: 검색한 문서 중에서 진짜로

RAG 파이프라인 구축 프로세스 보고서	
	관련된 문서가 차지하는 비율 - o context_recall: 실제로 관련된 문서 중에서 얼마나 많이 검색에 성공했는지 - o faithfulness: 생성된 답변이 가지고 있는 지식으로 얼마나 뒷받침 되는 지에 대한 비율 - o answer_relevancy: 생성된 답변이 주어진 질문과 얼마나 관련성이 있는 지 - 정성 평가 - o 샘플링 방식 - 무작위로 10 개의 질문을 선택하여 챗봇의 답변을 평가합니다. - 질문은 제품 매뉴얼의 다양한 섹션에서 추출된 내용을 기반으로 구성합니다. - o 평가 항목 - 정확성: 생성된 답변이 제품 매뉴얼의 내용과 얼마나 일치하는가? - 관련성: 답변이 검색된 문서/데이터와 관련이 있는가? - 명확성: 답변이 쉽게 이해되고 논리적으로 명확한가? - o평가 절차 - 각 질문에 대해 생성된 답변을 매뉴얼의 실제 내용을 기준으로 비교 검토합니다. - 관련성이 낮거나 잘못된 답변은 피드백을 기록하여 개선 방안을 도출합니다. - 평가 결과를 바탕으로 생성된 답변의 장단점을 정리하고, 추가 최적화 방향을 제시합니다.
결론 및 향후 발전 방향	향후 발전 방향 - 위치 기반 챗봇 서비스 : 사용자 위치 정보를 받아 가까운 오프라인 매장들 위주로 분석해 검색 물품을 추천해준다. - 마트 야간 할인 적용 - 원산지별 가격 반영