LLM을 활용한 데이터 검색 및 분석

실무자를 위한 RAG(검색 증강 생성) 파이프라인 구축 가이드

Upstage Technical Content Creator

교육 개요

교육 목표

- **이해하기 :** RAG의 핵심 개념과 필요성
- **구현하기 :** 문서 수집부터 검색, LLM 답변 생성 파이프라인 구축
- **평가하기 :** Hit@k, MRR 지표를 활용한 정량적 평가

대상 학습자

- 사내 문서를 다루지만 RAG 구축 경험이 없는 실무자
- Python 기본 문법과 데이터 처리(Pandas) 기초가 있는 분
- 데이터 기반의 정확한 답변 시스템이 필요한 엔지니어

목표: RAG 이해, 구현, 평가 | 대상: Python 기초가 있는 실무자

왜 RAG(검색 증강 생성)인가?

문제 1: 환각

LLM은 모르는 내용도 사실인 것처럼 그럴듯하게 지어내는 (Hallucination) 치명적인 단점이 있습니다.

문제 2: 최신 정보 부재

학습 시점 이후의 데이터나 사내 비공개 정보(Private Data)에 대해서는 답변할 수 없습니다.

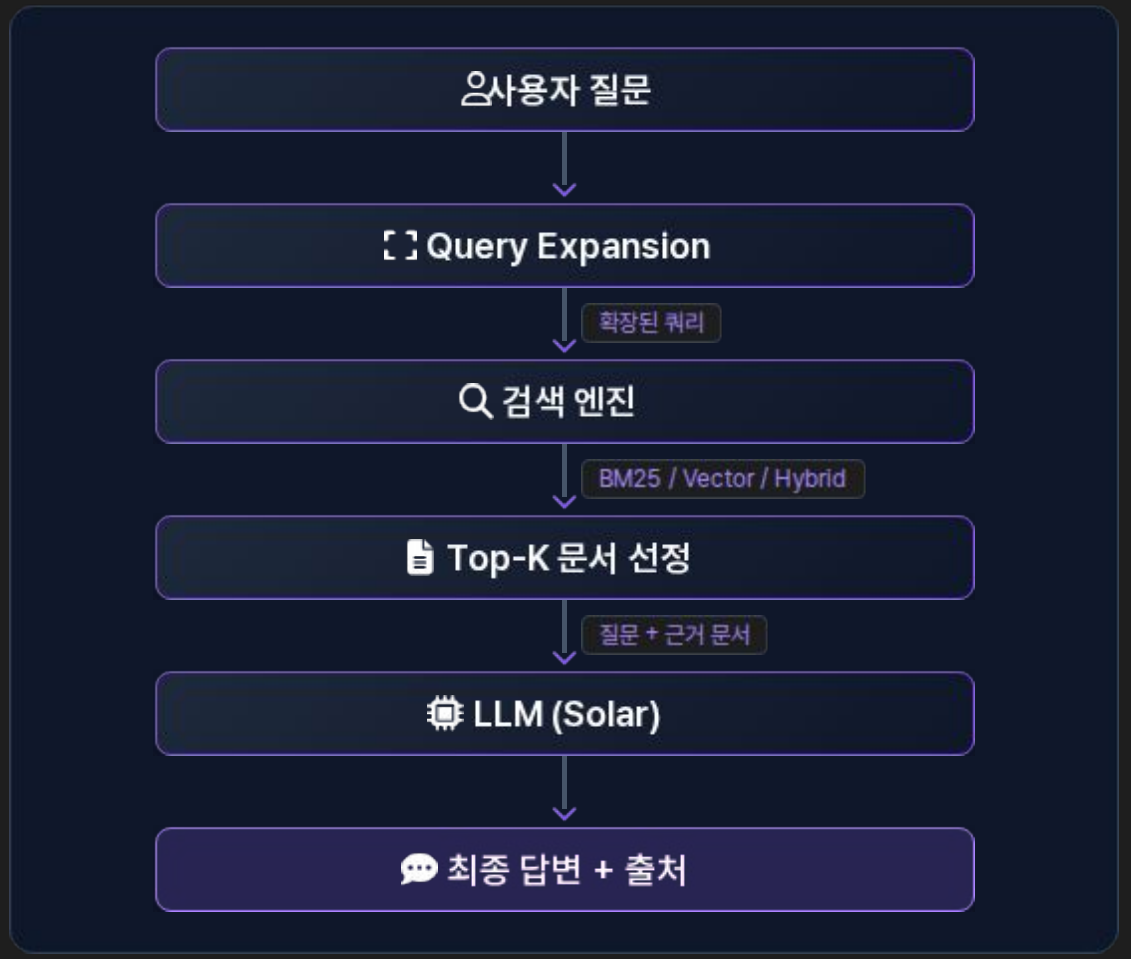
해결책: RAG

"오픈북 시험"과 같습니다. 질문에 답하기 전에 관련 문서를 먼저 찾아보고(Retrieval) 그 내용을 근거로 생성합니다.

핵심: 환각과 정보 부재 해결책으로서의 RAG (오픈북 시험 비유)

RAG 파이프라인 구조

- **Step 1: User Query**
  사용자의 질문 입력 및 확장
- **Step 2: Retrieval**
  검색 엔진에서 Top-K 문서 선정
- **Step 3: Augmentation**
  질문 + 근거 문서 프롬프트 결합
- **Step 4: Generation**
  LLM(Solar)이 최종 답변 생성



흐름: 질문 -> 확장 -> 검색 -> 결합 -> 생성 (Mermaid 다이어그램 시각화)

실습 데이터셋 : AGORA

Global AI Governance Documents

전 세계 1,500개 이상의 AI 법률, 규제, 가이드라인을 포함한 대규모 데이터셋입니다.

특징 1: 전문 용어
법률적 정의와 약어가 다수 포함되어 키워드 매칭이 중요함

특징 2: 복잡한 구조
문서 길이가 길고 문장 구조가 복잡하여 적절한 청킹(Chunking)이 필수

AGORA 데이터셋: 법률 용어가 많고 긴 문서 -> 검색 난이도가 높음

핵심 검색 기법 비교

BM25 (Keyword)

- **원리:** 단어의 빈도(TF-IDF) 기반
- **장점:** 정확한 용어, 고유명사 검색에 매우 강력
- **단점:** '자동차'로 검색하면 '차량'을 못 찾음 (동의어 불가)

Vector (Semantic)

- **원리:** 의미론적 유사도(Embedding) 기반
- **장점:** 문맥 파악 가능, 자연어 질문 처리에 능숙
- **단점:** 아주 구체적인 키워드 매칭에는 약할 수 있음

BM25(단어 퍼즐) vs Vector(의미 구름)의 장단점 비교

고도화 전략: Hybrid & Query Expansion

Hybrid Search (RRF)

BM25와 Vector 검색 결과를 결합하여 상호 보완합니다.

키워드 정확도 + 의미적 유연성
= 최적의 검색 품질 확보

Query Expansion

사용자의 짧고 모호한 질문을 LLM을 이용해 풍부하게 확장합니다.

입력: "AI 규제"
⬇ 확장 (LLM)
"AI 규제 인공지능 법률 거버넌스 컴플라이언스"

Hybrid: 두 기법의 장점 결합 | QE: 모호한 질문을 구체화하여 검색 성능 향상

검색 품질 평가 지표

Hit@k (적중률)

상위 k개 검색 결과 안에 정답 문서가 포함되어 있는가?

- Hit@5 = 1 (성공)
- Hit@5 = 0 (실패)

MRR (Mean Reciprocal Rank)

정답 문서가 몇 번째 순위에 등장했는가?

- 1위 등장: 1.0점
- 2위 등장: 0.5점
- 순위가 높을수록 점수가 큼

Hit@k: 포함 여부 | MRR: 등장 순위의 정확도

AGORA 데이터셋 평가 결과

순위	검색 기법	특징	비고
1위 WINNER	**BM25**	정확한 키워드 매칭	법률 용어/고유명사 검색에 강세
2위	Hybrid + QE	LLM 확장 + 결합	모호한 질문 보완에 효과적
3위	Hybrid	키워드 + 의미 결합	범용적인 성능 우수
4위	Vector	의미적 유사성	전문 용어 구분이 어려울 수 있음

💡 **Insight:** 무조건 최신 기술(Vector)이 좋은 것은 아닙니다. **전문 용어가 많은 데이터**에서는 전통적인 **BM25**가 더 강력할 수 있습니다.

결과: 전문 용어 특성상 BM25 1위 -> 데이터 특성에 맞는 기법 선정 중요

환각(Hallucination) 방지 전략

신뢰할 수 있는 답변 만들기

1. 시스템 프롬프트 제어

"제공된 문서 내용으로만 답변하세요. 정보가 없으면 '알 수 없음'이라고 답하세요."라고 명시적으로 지시합니다.

2. 근거 제시 (Citation)

답변 끝에 참조한 문서의 제목이나 페이지를 반드시 남기도록 하여 사용자가 검증할 수 있게 합니다.

전략: 시스템 프롬프트로 제한 + 출처 표기를 통한 검증 유도

실습 데모: Python + Upstage Solar

```python
def answer_with_context(question):
# 1. 문서 검색 (Hybrid Mode)
results = retrieve(question, mode='hybrid')
# 2. 프롬프트 구성
context = format_docs(results)
prompt = f"문서: {context}\n질문: {question}"
# 3. LLM 호출 (Upstage Solar)
return solar.chat.completions.create(prompt)
```

실행 결과 예시

Q: 미국 AI 권리 장전의 핵심 원칙은?

"미국 AI 권리 장전의 5가지 핵심 원칙은 다음과 같습니다: 안전하고 효과적인 시스템, 알고리즘 차별 방지, 데이터 프라이버시..."

[근거: Blueprint-for-an-AI-Bill-of-Rights.pdf]

데모: 검색 -> 프롬프트 -> 생성 과정의 코드 구현 및 결과 확인

향후 발전 과제 (Next Steps)

Reranker (재순위화)

1차 검색된 50~100개의 후보군을 정밀한 모델로 다시 정렬하여 정확도 극대화

Chunking 최적화

단순 길이 기준이 아닌, 의미 단위(Semantic) 또는 문단 단위 청킹 실험

메타데이터 필터링

연도, 국가, 문서 유형 등의 메타데이터를 활용하여 검색 범위 좁히기

고도화: Reranker 도입, 청킹 전략 수정, 메타데이터 활용

Q & A

강의 내용에 대해 궁금한 점이 있으신가요?

※ Summary

- RAG는 **환각을 줄이고 신뢰도**를 높이는 핵심 기술
- 데이터 특성에 맞는 **검색 기법(BM25 vs Vector)** 선택 중요
- **정량적 평가(Hit@k)**를 통한 지속적인 파이프라인 개선 필요