



커널 아카데미 : AI 부트캠프

Information Retrieval Seminar

2025. 11. 27(목)

목차

- 01. 팀 소개
- 02. 경진대회 수행 절차 및 방법
- 03. 분석 인사이트 및 결과
- 04. 회고

01

팀 소개

팀장/팀원 소개
협업 방식

* [AD-1 팀] 스타일은 달라도, 우리는 한 팀!



팀원
김시진
AI / 정보 과학
IR 성능 개선, 프롬프트
엔지니어링



팀원
임예슬
클라우드/관광경영&컴공
사회자 역할 & 코드 개선



팀원
김상윤
컴퓨터 공학
프롬프트



팀원
장윤정
컴퓨터 공학
프롬프트 엔지니어링

경진대회 협업 방식

Information Retrieval [대회] Scientific Knowledge Question Answering

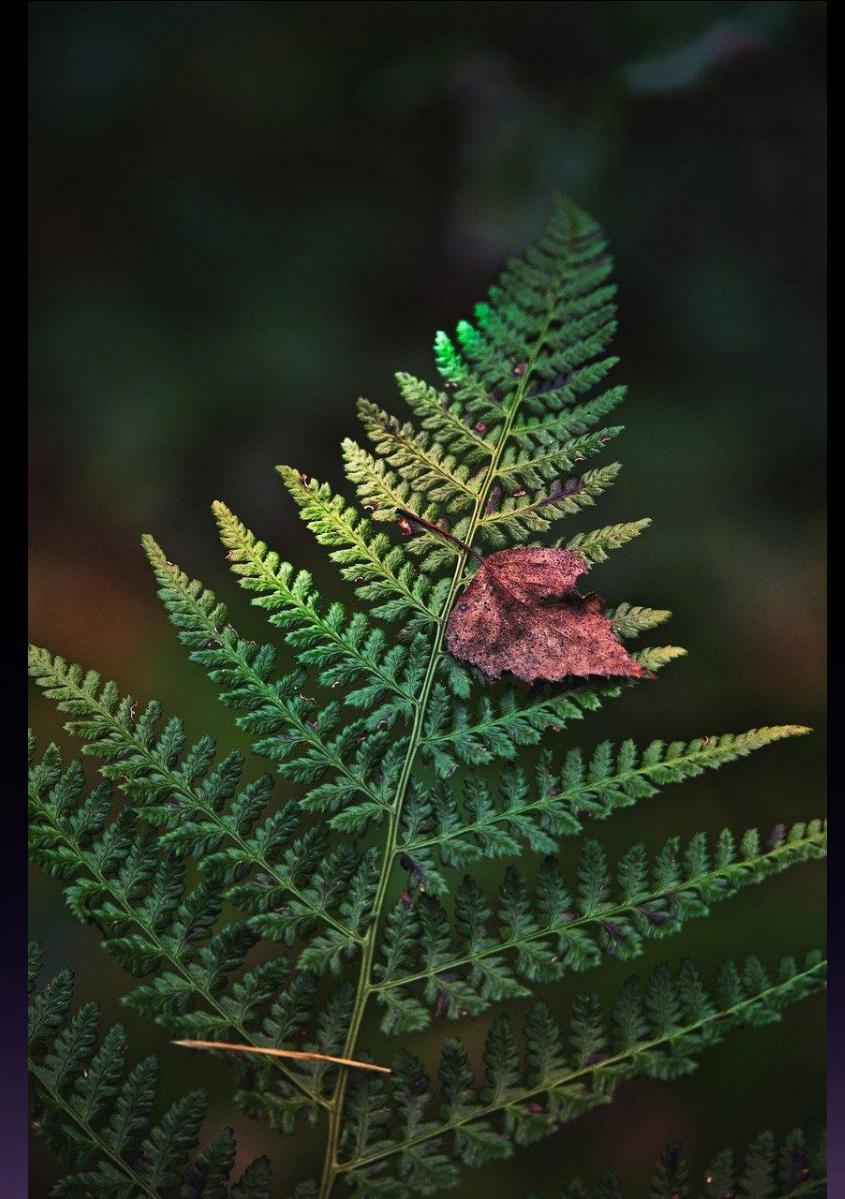
자연스러운 성장과 협업(*Leaf & Flow*)

협업 마인드셋 : 병렬 작업, 모르는 건 공유. 매일 아침 회의로 속도는 달라도 한 팀.

협업 진행 횟수 및 일정 : 매일 오전 10시, 데일리 미팅 및 인사이트 공유.

협업 진행 시 생긴 문제점 : 경진대회 진행하며, 막히는 부분 발생

문제 해결 방법 : 멘토링 1차, 2차 적극 활용



02

경진대회 수행 절차 및 방법

목표 수립
수행 내용 / 수행 결과

경진대회 목표 수립

Information Retrieval [대회] Scientific Knowledge Question Answering

주제

Scientific Knowledge Question Answering | 과학 지식 질의 응답 시스템 구축

질문과 이전 대화 히스토리를 보고 참고할 문서를 검색엔진에서 추출 후 이를 활용하여 질문에 적합한 대답을 생성하는 태스크입니다.

목표

목표

문서를 임베딩하고 문서 DB에서 검색하는 과정을 경험하며 RAG
파이프라인 학습

개요

소개 및 배경 설명

과학 상식 문서를 기반으로 RAG(검색 증강 생성) 구조를
적용하고, LLM이 standalone query를 생성해 핵심 질문을 요약한
뒤 답변을 생성

기간

2025. 11. 14 ~ 2025. 11. 27

경진대회 수행 내용

Information Retrieval [대회] Scientific Knowledge Question Answering

1
* 개발 환경 구축

```
python 3.10.13
sentence-transformers==5.1.2
transformers==4.57.3
elasticsearch==8.8.0
openai==1.7.2
```



2
* 데이터 분석

- 과학 상식 문서 데이터베이스 건수
- 4,200 건
- 평가 데이터 전체 건수 : 220
- 멀티턴 대화 : 20건
- 칫챗 질문 : 20건
- 멀티턴 포함 과학 상식 질문 : 200건



3
* 사용한 임베딩 모델

- Bi encoder
- Cross Encoder



경진대회 수행 결과

Information Retrieval [대회] Scientific Knowledge Question Answering

- 임베딩 모델
 - snunlp/KR-SBERT-V40K-klueNLI-augSTS
 - Qwen/Qwen3-Reranker-4B
- 과학 상식 문서 데이터베이스 건수 : 4,200 건
- 평가 데이터 전체 건수 : 220
 - 멀티턴 대화 : 20건
 - 칫챗 질문 : 20건
 - 과학 상식 질문 : 200건
- Bi encoder + Cross encoder & Reranker 임베딩 모델 적용 결과

Submitter	MAP MAP (Final)	MRR MRR (Final)	Created at	Phase	
	0.8227 -	0.8227 -	2025.11.26 15:32	Complete	

03

분석 인사이트 및 결과

문제 및 인사이트 도출
해결 방법 및 결과

경진대회 인사이트 공유

Information Retrieval [대회] Scientific Knowledge Question Answering

01. 문제 발생 배경 및 원인 분석

- 평가 데이터 220 건 중에 멀티턴 20건, 칫챗 대화 주로 감정 표현 20건 등의 데이터와 과학 상식 대화 200건의 LLM 프롬프트 엔지니어링을 통한 *Standalone query* 생성하는 부분에서 어려움이 있었음

LLM API를 통한 *Standalone query* 생성 결과

```
# 'standalone_query'가 비어있는 행의 개수 카운트 (빈 문자열 ''인 경우)
empty_query_count = (df['standalone_query'] == '').sum()

# 'topk'가 비어있는 행의 개수 카운트 (빈 리스트 []인 경우)
empty_topk_count = df['topk'].apply(lambda x: len(x) == 0).sum()

print(f"'standalone_query'에 데이터가 없는 행의 수: {empty_query_count}")
print(f"'topk'에 데이터가 없는 행의 수: {empty_topk_count}")

✓ 0.0s
```

Python

```
'standalone_query'에 데이터가 없는 행의 수: 41
'topk'에 데이터가 없는 행의 수: 43
```

경진대회 인사이트 공유

Information Retrieval [대회] Scientific Knowledge Question Answering

02. 인사이트 도출

🔍 프롬프트 엔지니어링

* 일상 대화를 제외한 과학 상식 질문에 대해 **standalone_query**를 생성할 수 있도록 프롬프트 작성

# eval_id	└ standalone_query	└ msg
276		[{"role": "user", "content": "요새 너무 힘들다."}]
14		[{"role": "user", "content": "세제의 거품이 만들어지는 원리는?"}]
270		[{"role": "user", "content": "나무가 생태계에서 하는 역할에 대해 설명해줘."}]
238	전류 흐름 극대화 배터리 저항 연결 방법	[{"role": "user", "content": "전류의 흐름을 극대화 하려면 배터리와 저항을 어떻게"}]
269	식물이 높이 자랄 수 있게 하는 메커니즘	[{"role": "user", "content": "식물이 높이 자랄 수 있게 하는 메커니즘이 궁금해."}]
43	왜 달은 항상 같은 면만 보이나	[{"role": "user", "content": "달을 보면 항상 같은 면만 보이더라구?"}, {"role": "assistant", "content": "달은 항상 같은 면만 보이는 이유는 달의 운동軌道 때문이다. 달은 지구를 공전하는 동안 같은 면을 지향하는 경향이 있어서 우리는 그 면만 볼 수 있다."}]
65		[{"role": "user", "content": "식초와 베이킹 소다를 섞어주면 어떤 일이 일어나나?"}]
97		[{"role": "user", "content": "세균이 나쁜줄 알았는데 그게 아니야?", {"role": "assistant", "content": "세균은 유익한 면도 있고 나쁜 면도 있다. 예를 들어, 일부 세균은 미생물 치료에 사용되며 다른 일부는 감염을引き起す。"}]
206		[{"role": "user", "content": "오토마톤의 특징에 대해 알려줘."}]
21		[{"role": "user", "content": "다양한 책을 catalog화 하는 코드에서 class 정의 방법은?"}]
221		[{"role": "user", "content": "전구가 병렬로 연결될 때 전류가 줄어드는 원인은?"}]
71	물속에서 침전이 발생하는 원리	[{"role": "user", "content": "물속에서 침전이 발생하는 원리에 대해 알려줘."}]
254	화산 폭발 후 새로운 생물 군집이 생겨나는 것	[{"role": "user", "content": "화산 폭발이 발생한 후 새로운 생물 군집이 생겨나는 것에 대해 설명해줘."}]
226		[{"role": "user", "content": "축전기를 병렬로 이어주면 전체 용량이 어떻게 되는가?"}]
241	정육면체가 물에 떠 있을 때 수면 윗부분의	[{"role": "user", "content": "정육면체가 가라앉지 않고 물 위에 떠 있을 때 수면 윗부분의 물 속에 들어온 공기의 역할은?"}]



# eval_id	└ standalone_query	└ msg
276		[{"role": "user", "content": "요새 너무 힘들다."}]
14	세제의 거품이 만들어지는 원리	[{"role": "user", "content": "세제의 거품이 만들어지는 원리는?"}]
270	나무가 생태계에서 하는 역할	[{"role": "user", "content": "나무가 생태계에서 하는 역할에 대해 설명해줘."}]
238	전류의 흐름을 극대화하기 위한 배터리와 저항	[{"role": "user", "content": "전류의 흐름을 극대화 하려면 배터리와 저항을 어떻게"}]
269	식물이 높이 자랄 수 있게 하는 메커니즘	[{"role": "user", "content": "식물이 높이 자랄 수 있게 하는 메커니즘이 궁금해."}]
43	달이 항상 같은 면만 보이는 이유	[{"role": "user", "content": "달을 보면 항상 같은 면만 보이더라구?"}, {"role": "assistant", "content": "달은 지구를 공전하는 동안 같은 면을 지향하는 경향이 있어서 우리는 그 면만 볼 수 있다."}]
65	식초와 베이킹 소다를 섞으면 어떤 화학 반응이 일어나나?	[{"role": "user", "content": "식초와 베이킹 소다를 섞어주면 어떤 일이 일어나나?"}]
97	세균의 순기능에 대해 설명해 주세요.	[{"role": "user", "content": "세균이 나쁜줄 알았는데 그게 아니야?", {"role": "assistant", "content": "세균은 유익한 면도 있고 나쁜 면도 있다. 예를 들어, 일부 세균은 미생물 치료에 사용되며 다른 일부는 감염을引き起す。"}]
206	오토마톤의 특징	[{"role": "user", "content": "오토마톤의 특징에 대해 알려줘."}]
21	다양한 책을 catalog화 하는 코드에서 class 정의 방법은?	[{"role": "user", "content": "다양한 책을 catalog화 하는 코드에서 class 정의 방법은?"}]
221	전구가 병렬로 연결될 때 전류가 줄어드는 원인은?	[{"role": "user", "content": "전구가 병렬로 연결될 때 전류가 줄어드는 원인은?"}]
71	물속에서 침전이 발생하는 원리	[{"role": "user", "content": "물속에서 침전이 발생하는 원리에 대해 알려줘."}]
254	화산 폭발 후 새로운 생물 군집이 생겨나는 것	[{"role": "user", "content": "화산 폭발이 발생한 후 새로운 생물 군집이 생겨나는 것에 대해 설명해줘."}]
226	축전기를 병렬로 연결했을 때 전체 용량이 어떻게 되는가?	[{"role": "user", "content": "축전기를 병렬로 이어주면 전체 용량이 어떻게 되는가?"}]
241	정육면체가 물에 떠 있을 때 수면 윗부분의	[{"role": "user", "content": "정육면체가 가라앉지 않고 물 위에 떠 있을 때 수면 윗부분의 물 속에 들어온 공기의 역할은?"}]

경진대회 인사이트 공유

Information Retrieval [대회] Scientific Knowledge Question Answering

02. 인사이트 도출

🔍 검색 모델 심화 (Bi-encoder vs Cross-encoder)

구분	Bi-encoder	Cross-encoder
속도	매우 빠름 ⚡	느림 🐢
정확도	보통	매우 높음 ⚪
계산 횟수	Query 100개 + Doc 1,000개 = 1,100번	$100 \times 1,000 = 100,000$ 번
용도	초기 검색	재순위화(re-ranking)

경진대회 인사이트 공유

Information Retrieval [대회] Scientific Knowledge Question Answering

02. 인사이트 도출



Hackathon 문제

기존 방식의 한계

- 문서 4,200개 × 쿼리 1개 = 4,200번 유사도 계산
- 문서가 수백만 개라면? → 매우 느림

해결책: ANN (Approximate Nearest Neighbor)

- 모든 벡터를 검색하지 않고, 유사할 것 같은 일부만 검색
- 정확도는 약간 희생, 속도는 10~100배 향상

알고리즘	원리	장점	단점
전체 검색	모든 벡터 비교	100% 정확	매우 느림
FAISS	클러스터링 (IVF)	범용적, GPU 지원	설정 복잡
HNSW	계층 그래프	정확도 최고	메모리 많이 사용
Milvus	다양한 인덱스	프로덕션 완성도	복잡한 설정
pgvector	HNSW (PostgreSQL)	기존 DB 활용	PostgreSQL 의존
ScaNN	벡터 양자화	속도/정확도 균형	설정 어려움

경진대회 인사이트 공유

Information Retrieval [대회] Scientific Knowledge Question Answering

03. 해결방법

- Bi 를 통해 100개의 유사도 높은 문서 추출
- Cross encoder & Reranker 를 통해 100개의 문서와 검색 쿼리와 짹을 지어 유사도 높은 k개의 문서를 추출

04. 결과

- Cross encoder 모델을 사용했을 때 보다 훨씬 속도가 빨라지고 효율적임
- 베이스라인 코드의 LLM 프롬프트를 통한 Standalon query 생성 보다 검색 성능이 개선됨
- LLM 프롬프트를 통한 MAP 는 0.73~0.79점

서브미션 결과

Submitter	MAP MAP (Final)	MRR MRR (Final)
	0.8227 -	0.8227 -

04

회고

우리 팀의 목표 달성도
느낀점 및 향후 계획

경진대회 회고

Information Retrieval [대회] Scientific Knowledge Question Answering

Point 1

우리 팀의 처음 목표에서 어디까지 도달했는가

* 강의 수강을 통해 기본적인 IR 기본 구조와 개념을 학습하고, Baseline 코드를 실행하며 기본 파이프라인 이해

Point 2

우리 팀이 잘했던 점

* 매일 회의를 통해 점수 향상 방법에 대해 공유, 그 외 도움될만한 팁 공유

Point 3

협업하면서 아쉬웠던 점

* 인덱싱 방식과 임베딩 방법을 다양하게 실험해보고 싶었지만, 충분히 시도하지 못함

- 향후 계획 : 추가 학습을 통해 다양한 임베딩과 인덱싱 방법을 이해하고 실험해보기

경진대회 진행 소감

Information Retrieval [대회] Scientific Knowledge Question Answering



* 김시진 이번 IR 경진대회를 통해 문장 간의 유사도를 구하는 *Bi encoder, Cross encoder & Reranker* 알고리즘 등과 검색 성능을 개선하는 방법 중에 하나인 *HyDE*를 배우게 되어 좋은 경험이었습니다.

* 임예슬 여러가지 방법론을 배울수 있어 좋았습니다. 작업은 병렬로 진행되었지만 여러가지 인사이트를 팀원들과 나눌 수 있어 좋았습니다.
감사합니다.

* 김상윤 이번 프로젝트는 개인적인 일이 있어서 많은 시간투자를 하지 못해서 아쉬웠고, 프롬프트 작성만으로 많은 효과를 볼 수 있다는게 재미있었습니다.

* 장윤정 RAG실습을 해볼 수 있어서 좋았고, 멘토링을 진행하면서 많은 인사이트를 얻을 수 있었습니다. 멘토님께서 말씀해주신 방법을 이번에 다 해보지 못했지만 추후 시간이 된다면 해보고 싶습니다.

Life-Changing Education

감사합니다.
