논문 : 한국어 텍스트 분류를 위한 임베딩 모델 및 Advanced RAG 방법론 비교 (on writing)

연구 배경 및 필요성

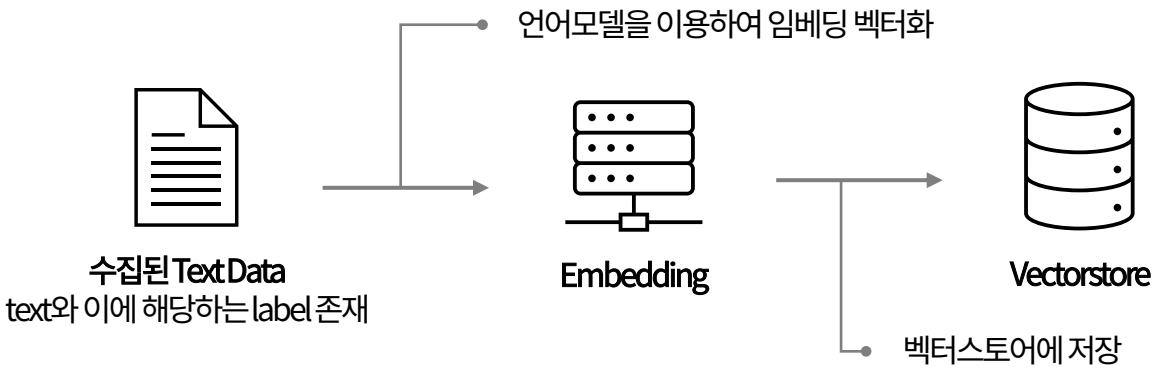
- 업무의 자동화와 효율성을 위한 텍스트 분류 모델의 필요성
- 기존의 언어모델은 fine-tuning으로 인한 컴퓨팅 자원 및 시간의 소모가 큼
- RAG를 적용한 LLM을 사용한 텍스트 분류

LLM의 NLP 성능에 실시간으로 새로운 유형의 데이터 반영하여 분류 성능 개선 및 효율성 제고

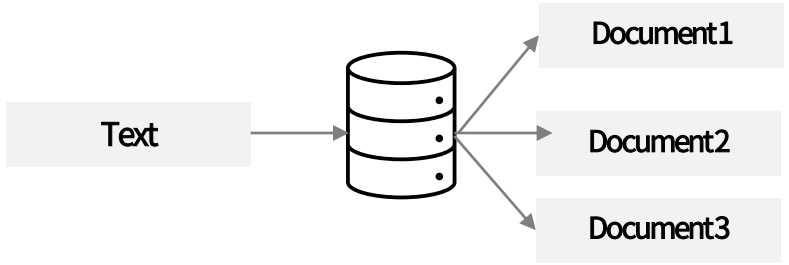
사용 데이터셋

AI-HUB에서 제공하는
민원, 윤리, 법률, 쇼핑몰
네 개 도메인의 텍스트 분류 데이터

RAG 개요



RAG : Retrieval Augmented Generation



LLM에 text가 입력으로 들어오면
벡터스토어로부터 유사한 문장을 검색(Retrieval)하여
답변을 생성(Generate)
=> Fine-tuning 없이도 새로운 데이터 반영 가능

논문 : 한국어 텍스트 분류를 위한 임베딩 모델 및 Advanced RAG 방법론 비교 (on writing)

임베딩 모델

- train data을 각 모델별로 임베딩하여 벡터스토어 구축 후 최적의 임베딩 모델 선정
- 한국어 특화 임베딩 언어모델들을 사용

openai의 3-small와 ada-002, kobert, koelectra-v3-discriminate, jina-embedding-v3, KoE5, KURE-v1

- 평가지표는 Recall@k와 MRR@k를 사용
- 50개의 seed에 대한 평균 및 표준오차를 기준으로 평가

Models	Recall@1	Recall@3	Recall@5	Recall@10
text-embedding-3-small	0.5700 (0.0015)	0.6789 (0.0012)	0.7301 (0.0012)	0.7903 (0.0009)
text-embedding-ada-002	0.5761 (0.0014)	0.6862 (0.0013)	0.7379 (0.0013)	0.7980 (0.0012)
kobert	0.3385 (0.0023)	0.5272 (0.0024)	0.6250 (0.0020)	0.7389 (0.0016)
koelectra-base-v3-discriminator	0.2420 (0.0018)	0.4533 (0.0020)	0.5776 (0.0023)	0.7293 (0.0021)
Jina-embeddings-v3	0.5834 (0.0016)	0.6898 (0.0014)	0.7424 (0.0015)	0.8012 (0.0010)
KoE5	0.5932 (0.0015)	0.7045 (0.0013)	0.7530 (0.0010)	0.8059 (0.0009)
KURE-v1	0.5935 (0.0016)	0.7019 (0.0013)	0.7514 (0.0013)	0.8044 (0.0008)

Models	MRR@1	MRR@3	MRR@5	MRR@10
text-embedding-3-small	0.5700 (0.0015)	0.6177 (0.0012)	0.6294 (0.0012)	0.6375 (0.0011)
text-embedding-ada-002	0.5761 (0.0014)	0.6239 (0.0012)	0.6356 (0.0012)	0.6438 (0.0011)
kobert	0.3385 (0.0023)	0.4204 (0.0022)	0.4427 (0.0021)	0.4581 (0.0019)
koelectra-base-v3-discriminator	0.2420 (0.0018)	0.3327 (0.0017)	0.3610 (0.0016)	0.3816 (0.0015)
Jina-embeddings-v3	0.5834 (0.0016)	0.6297 (0.0014)	0.6417 (0.0013)	0.6497 (0.0012)
KoE5	0.5932 (0.0015)	0.6420 (0.0013)	0.6531 (0.0012)	0.6603 (0.0011)
KURE-v1	0.5935 (0.0016)	0.6410 (0.0013)	0.6523 (0.0012)	0.6596 (0.0011)

Advanced RAG

Pre-retrieval	Multi-Query, Hyde
Retrieval	Hybrid Search
Post-retrieval	Reranker, Modular RAG

- 50개의 seed에 대해 validation data를 이용해 최적의 문서 개수 선정
- 각 방법론을 적용하여 50개의 seed에 대해 test data의 성능 비교 및 평가
  - Binary Classification (윤리, 법률 데이터) : Accuracy, Precision, Recall, F1-score
  - Multicalss Classification (민원, 쇼핑몰 데이터)
    : Accuracy, Macro Precision & Recall & F1-score
- 최적의 Advanced RAG 방법론 선정