RAG기법 데이터 학습 자동화 구현을 이용한 클라우드 환경의 행정·공공기관용 chat 웹서비스 기능 개발

김형식, 이재진

숭실대학교

nabbyy@naver.com, zlee@ssu.ac.kr

Development of chat web service functions for administrative and public institutions in the cloud environment using the implementation of RAG technology data learning automation

Kim Hyeong Sik, Lee Jae Jin

Soongsil Univ.

요 약

Open AI의 초거대 언어 모델(LLM) ChatGPT의 등장으로 최근 인공지능 기술 분야는 새로운 국면을 맞이하고 있다. 국내외 기업 모두 생성형AI를 활용한 기술과 혁신적인 서비스 개발을 위해 분주하게 노력하고 있다. 하지만 생성형AI에는 잘못된 정보를 마치 정답처럼 답변하는 환각 현상 문제가 있는데, 해당 현상을 완화할 수 있는 여러 기법에 대한 연구가 진행되고 있으며, 대표적으로는 Fine-tunning과 RAG가 있다. 본 논문에서는 클라우드 기술을 통해 정형 데이터 전처리 과정과 RAG기법의 vector embedding 과정을 자동화하여, 비전문가도 손쉽게 생성형AI의 LLM 학습을 가능하게 하였다. 특히, 국내 민간 공공클라우드 인프라(CSAP) 기반으로 Chat서비스를 구현하여 행정·공공기관에서도 도입할 수 있는 환경을 마련하였다.

주제어 : RAG(Retrieval Augmented Generation), 생성형 AI(Generative AI), LLM(Large Language Model), 환각 현상(Hallucination), HyperCLOVA

Ⅰ. 서 론

2022년 ChatGPT의 등장으로 초거대 인공지능(AI)서비스 즉, 생성형 AI의 연구와 개발이 열풍이다. OpenAI의 ChatGPT, Google의 Bard가 대표적이고, 현재 국내에서도 NAVERCLOUD, KT, LGCNS, KAKAO 등에서 서비스를 공개하였거나 준비가 한창이다.

하지만 ChatGPT와 같은 생성형 AI에는 잘못된 정보를 마치 정답인 것처럼 답변하는 "환각(Hallucination) 현상"이 발생하는 치명적인 문제가 있는데, 이를 보완하기 위해 생성형 AI의 뇌 역할을 할 수 있는 RAG (Retrieval Augmented Generation)기법 연구가 급부상하고 있다.

RAG서비스는 질문에 관련된 정보를 외부 문서를 참조한 Vector database에서 검색한 후 이를 기반으로 답변을 생성하는 방법을 일컫는다. 기존에는 Retrieval model과 Generation 모델을 함께 학습시키는 것을 RAG라고 하였지만, LLM이 등장하며 통칭하여 불리기 시작하였다.

본 연구에서는 네이버클라우드의 Cloud Functions 기능을 활용하여 추출된 원본 데이터를 텍스트데이터 형태로 변환하고, TEXT Chunk 및 Vector Embedding까지의 과정을 자동으로 처리할 수 있도록 프로세스를 구현하였으며, 이를 통해 기관이 보유한 최신의 데이터를 RAG기법을 통해 학습할 수 있도록 하였다.

추가로, 국내 행정·공공 기관은 [클라우드 컴퓨팅 발전 및 이용자 보호에 관한 법률]에 따라 주요 정보(데이터)가 포함된 정보시스템은 CSAP(클라우드컴퓨팅서비스 보안인증)을 획득한 클라우드를 우선 사용해야 한다. 또한, 행정안전부는 지난 5월 [공무원을 위한 '챗GPT 활용 방법 및 주의사항 안내서' 배포] 등을 통해 정보탐색 능력, 언어능력 활용, 컴퓨터능력 활용 등의 업무 생산성 증가 부분을 제외하고 의사결정이 완료되지 않거나, 공표되지 않은 정보, 외부 반출이 허용되지 않은 정보 등은 질문 자체를 금지 하고 있다.

따라서, 이번 구현에서는 ChatGPT가 네이버클라우드의 공공클라우드 인프라 환경과 HyperCLOVA의 LLM과 아닌 Embedding API를 활용하여 구현하여 국내 행정·공공 기관에서 구축에 보안 및 행정적인 이슈가 없도록 구현하였다.

Ⅱ. 관련 연구

LLM 서비스에 기존 학습 과정에 포함되지 않은 지식(데이터)을 주입하는 방법은 크게 Fine-Tuning과 RAG 크게 2가지를 주로 사용한다.

RAG(Retrieval Augmented Generation) 기법은 새로운 정보가 있는 텍스트 데이터 정보를 Embedding하여 Vector database에 숫자 형태로 저장하고, 프롬프트 구성을 진행할 때 그림[1]과 같이 Vector database에 저장된 정보를 먼저 참고하여 프롬프트를 구성한 뒤 LLM으로 부터 최종 답변을 얻어내는 구조이다. Vector Database는 방대한 양의 고차원 데이터를 Vector 형태로 저장하고 처리하기 위한 특수 데이터베이스 이다.

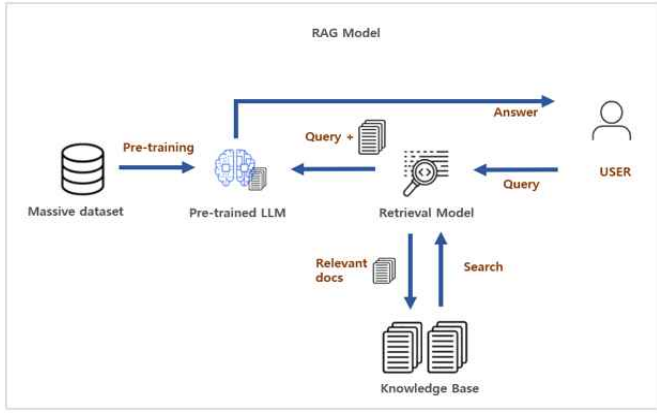


그림 [1]

Vector database 중 오픈소스로 사용할 수 있는 종류는 Chroma, Milvus, Vespa, Vald 등이 있다.

Ⅲ. 연구 방안

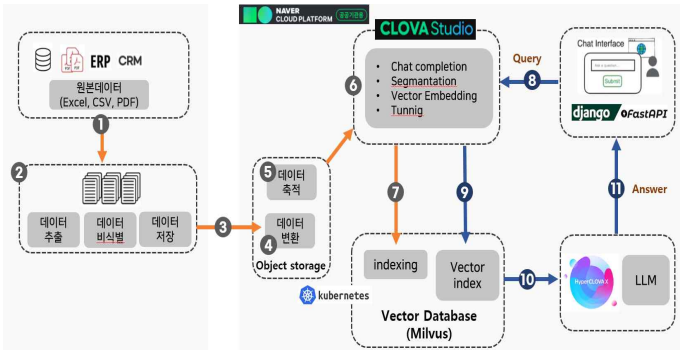
본 연구에서 구현되는 Chat Application 서비스의 특장점 및 활용 기술은 다음과 같다.

1) 그림[2] 1번~7번까지 데이터 추출, 변환, 분할, 벡터 저장까지 일련의 과정을 클라우드의 Serverless 기술(Cloud Functions)과 HyperCLOVAX의 API를 활용하여 자동화 처리하였다.

2) 자동화 작업을 통해 최신의 데이터를 Vector Database에 학습이 가능하며, 최신 데이터에 대한 질의에 답변할 수 있다.

3) 주요 시스템 영역이 클라우드 기술로 구축되어 관리 영역 축소, 인건비 절감, 개발 속도, 운영 비용 절감 효과가 증가한다.

4) 네이버 공공클라우드(CSAP) 기반 구축을 통해 행정·공공기관에서 시스템 구현에 필요한 인프라 보안성 문제를 해결할 수 있다.



그림[2]

데이터 전처리 및 Vector Index 자동화 프로세스는 다음 그림[3]과 같이 구현하였다.

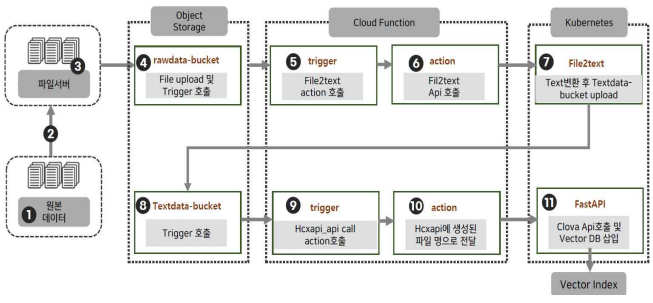
1) 원천 데이터를 보유한 기관의 ERP, CRM, SCM과 같은 정보시스템에서 주기적으로 데이터를 추출한다.

2) 추출된 원본 데이터는 민감정보 포함 여부(개인정보, 위치정보)를 확인 후 별도의 비식별 솔루션을 통해 가공 후 파일서버에 적재한다. 민감 정보가 없는 바로 파일서버에 복제한다.

3) 파일서버에 적재된 데이터는 파일서버에 배치 스크립트를 통해 클라우드 오브젝트 스토리지의 지정된 버킷에 자동 업로드 한다.

4) 클라우드 오브젝트 스토리지에 파일이 업로드 되면 그림[3]의 4번째 와 같이 Cloud function의 코드가 자동 실행된다.

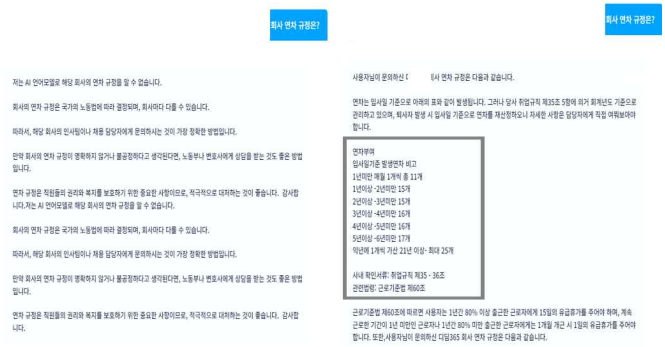
5) Cloud Function은 action 기능을 통해 javascript, swift, java, python, php 언어를 지원하며, 이번 연구에서는 python코드를 통해 file2data 전처리 로직을 생성하였다.



그림[3]

Ⅳ. 연구 결과

기관 내부 규정 및 행정민원(QnA)의 데이터 등을 hyperCLOVA Default 문의와 Vector DB를 참고 후 데이터에 비교 질의할 수 있도록 Chat WEB 서비스 기능 구현 후 질의 결과에 대한 그림[4]이다. 학습된 데이터에 대하여 구체적인 답변 확인이 가능하다.



그림[4]

Ⅴ. 결론

본 연구에서는 행정·공공기관이 네이버클라우드 HyperCLOVA에서 제공하는 Vector embedding API와 CLOVA의 LLM을 활용하여 Chat WEB Service를 구현하고, 실제 사용 기관의 보유한 데이터 학습을 통해 질의한 내용에 대하여 정상적인 결과가 응답하는지 확인하였다.

또한, 물리적인 시스템과 일반적인 민간클라우드가 아닌 공공이용 민간클라우드(CSAP) 기반의 서비스를 구축하여 정부 클라우드 도입 및 전환에 대한 정책을 준수한 정보시스템을 구축하는 방안을 도출하였으며, 자동화된 데이터 전처리 프로세스와 사용한 만큼 과금 되는 클라우드 서비스의 요금 방식을 통해 레거시 기반 시스템을 구축하는 것보다 관리 및 운영 비용의 절감 효과가 발생하였다.

추가적인 연구를 통해 생성형 AI 서비스의 고질적인 문제인 환각(할루시네이션) 현상을 추가 완화할 수 있도록 지속적인 학습이 필요하며, 이미지, 영상, 음성과 같은 고도화된 데이터에 대한 데이터 변환, 축적 기술 등의 전처리 과정에서 자동화할 수 있는 기술에 대한 연구가 필요하다.

참 고 문 헌

[1] Joong-Min Shin., Seung-Ryul Park., Hye-Rin Kim., & Jung-hun Lee. (2023). “Search-based Generation Techniques for Enhancing LLM Responses: A Comparative Study of GPT3.5 and GPT4 in Zero-shot and RAG”. 한국정보통신학회 종합학술대회 논문집. 27(2):350-352

[2] Hyun-Seung Lee., Jae-Beom Kim. (2023). “Case study on mitigating hallucinations in generative AI for game content generation”. 한국게임학회 논문지. 323(5):121-129.

[3] Naver cloud, “cloud function”, (2023.10)
https://guide-gov.ncloud-docs.com/docs/ko/cloudfunctions-overview

[4] Naver cloud, “hyperclovax, vector embedding”, (2023.11)
https://guide.ncloud-docs.com/docs/ko/clovastudio-overview

[5] Microsoft, “retrieval augmented generation”, (2023.09.25.)
https://learn.microsoft.com/ko-kr/azure/machine-learning/concept-retrieval-augmented-generation?view=azureml-api-2