김민지, LLMOps Engineer

minji.sql@gmail.com

010-5742-7697

Blog : medium.com

Summary

- 약 **10000건의 문서**를 전처리하고, Embedding 하여, **기본 RAG 파이프라인 구축**을 한 경험이 있습니다.
- 대규모 언어 모델의 성능을 지속적으로 모니터링하고, 사용자 피드백을 수집하여 모델을 **지속적으로 향상**시키는 것을 즐깁니다.
- **Neo4j** DB를 활용하여, 자율적이고 관계가 복잡한 데이터의 사이를 효과적으로 표현하고 관리한 경험이 있습니다.
- end-to-end 프로젝트를 리드한 경험이 있으며, 문제 해결과정을 **글로 정리하며 학습**합니다. (5000 view 달성 블로그)

Experience

2024.03 ~ 2024.08

R&D DX 팀 / AI Developer

생성형 AI 기반 기술문서 QA 시스템 개발

- **Information**
  - **한국타이어 연구소의 약 1만건의 문서를 기반으로 질의응답이 가능한 Chat Bot 시스템입니다.**
  - **주니어 직원의 질문에 시니어가 답변하는 과정 또는 업무 관련 문서를 찾는 과정을 도와 휴먼 리소스를 절감하는 데 목적이 있습니다.**

사용 기술 | Python, FastAPI, Docker, OpenSearch, Bedrock Agent, Langchain, Huggingface, Claude 3

- 문서 데이터 전처리시, **중요한 metadata 정보는 태깅**을 하여, 검색시 필터링을 할 수 있게끔 했습니다.

[RAG 성능 향상]

- test data set(100개)를 만들어 성능평가를 진행했습니다. (성능평가 지표 52/100 달성)
- 효율적인 chunk 개수, overlap 값을 알아내기 위해, 1n번의 인덱싱을 진행했습니다. (성능평가 지표 74/100 달성)
- **ReRanking과 Ensemble Retriever**를 활용하여 기본 RAG 파이프라인의 검색 성능을 향상시켰습니다. 특히, 다양한 Ensemble 파라미터 값을 실험하여 retriever 성능을 최적화하고, 더 나은 검색 결과를 도출할 수 있는 최적의 설정을 발견했습니다. (성능평가 지표 94/100 달성)

[부하테스트]

- 기존 서버는 부하테스트(Load Testing)시 **53.5초**가 걸리는 문제가 있었습니다. 여러 요청을 효율적으로 분산 처리하기 위해 uvicorn을 사용했고, 또한 **CPU 집약적인 작업은 효율적인 처리를 위해 내부 로직을 멀티프로세싱**로 작업할 수 있게 함으로써 10초 까지 줄였습니다. (53.5s →10s)

[OpenSearch 데이터 Migration] (블로그)

- 서비스 유저 수 증가 및 저장되는 문서 데이터의 양이 많아지면서, 검색 서버와 인덱싱 서버를 분리 운영할 필요성이 증가했습니다.
- 데이터 손실 없이 이전하기 위해 Docker 볼륨을 통한 스냅샷 저장소를 설정하고, SCP를 활용하여 서버 간 데이터를 안전하게 전송했습니다.
- OpenSearch의 Snapshot 기능을 자동화된 파이프라인으로 설정하여, 파일 전송 및 복원 과정을 효율화하고, 오류를 최소화했습니다.
- 대규모 문서 데이터 운영의 안정성을 확보하고, 향후 서비스 확장 시에도 유연하게 대응할 수 있는 구조를 마련했습니다.

음성 파일 기반 회의록 자동 작성 시스템 PoC

- **Information**
  - **라디오 및 회의 음성 데이터를 자동으로 텍스트로 변환하고, 화자를 식별하는 시스템입니다.**
  - **음성 데이터를 활용한 정보 검색과 분석이 가능하도록 PoC 단계에서 프로토타입을 구축했습니다.**

사용 기술 | Python, ffmpeg, Faster-Whisper, Pyannote, Huggingface

- 화자 정보가 'unknown'으로 분류될 경우, 앞뒤 대화 내용 문맥을 파악해 보정했습니다.
- 발화 간격을 기준으로 동일 화자의 연속 발화를 묶어 불필요한 화자 전환을 줄였습니다.
- 정규 표현식을 사용해 타임스탬프 및 문장 구조를 정리했습니다.
- 최종 대화록을 정제된 형태로 저장하고, 검색 및 분석이 가능하도록 구조화된 데이터를 제공했습니다.
  화자 분리 및 텍스트 변환 성능을 평가한 결과, 초기 모델 대비 오류율을 크게 줄이고, 회의록 자동 생성의 정확도를 개선했습니다.

DATA NETWORKS LAB

2024.03 ~ 2024.08

AI 데이터 네트워크 연구실 / 학부연구생

라디오 스트리밍 파일에서 텍스트 기반 광고 이벤트 탐지 방법

- **Information**
  - **기존 연구에서는 오디오 주파수 특징 분석 또는 지문 기반 기법을 활용하나, 계산 복잡도가 높고 실시간 적용이 어려운 한계가 있었습니다.**
  - **라디오 콘텐츠 속 광고 구간을 자동으로 탐지하고 제거하는 '광고 구간 탐지 파이프라인'을 설계하고 그 성능을 평가 했습니다.**

사용 기술 | Python, VAD(Voice Activity Detection), Faster-Whisper, GPT-4o-mini, inaSpeechSegmenter

- VAD(Voice Activity Detection) 알고리즘인 inaSpeechSegmenter로 오디오 데이터를 음성, 음악, 소음으로 분류했습니다.
- Whisper의 Robert 모델을 사용하여 음성을 텍스트로 변환 후, 광고 이벤트를 탐지하고, 시간 구간 및 광고 요약을 제공하도록 설계 했습니다.
- 광고는 보통 하나의 구간에 집중되어 여러 개의 광고가 몰려있다는 특징이 있어 분류에 있어 적합한 프롬프트를 설계 했습니다.

[실험 및 성능 평가]

- EBS, KBS, MBC에서 제공하는 6개의 라디오 프로그램을 대상으로 총 109개의 오디오 파일을 실험에 활용하였습니다.
- 광고 구간의 시간적 일치 여부와 상관없이 모델이 **실제 광고를 탐지한 정확도는 약 82.56%**입니다.
- 정밀도는 약 89.11%로, 재현율은 약 91.84%로 나타났습니다. 두 지표의 조화 평균인 **F1 스코어는 약 90.45%**로, 모델이 광고 탐지에 있어 정밀도와 재현율 간의 균형을 잘 유지하고 있음을 보였습니다.

북그북그

2023.12 ~ 2024.02

Techeer bootcamp / leader

AI 기반 동화책 제작 서비스

- **Information**
  - **아이들이 주인공이 되어 AI와 함께 이야기를 만들어가는 동화책 제작 서비스**

사용 기술 | Django, Python, Celery, RabbitMQ, Docker, AWS (EC2, S3), Nginx, Github Action, ChatGPT API, DALL-E api, Clova

- 실시간으로 여러 갯수의 이미지와 오디오 파일 생성 및 처리시, 30초 이상의 많은 시간이 소요됩니다. 사용자가 책을 클릭했을 때만 필요한 정보이기 때문에 **Celery, RabbitMQ를 사용해 작업을 비동기 처리 하여, 불필요한 기다림을 해결**했습니다.
- 구조화된 프롬프팅을 통해 원하는 답변의 복잡한 형식을 100% 지켰습니다. **프롬프트를 동화책 초기, 중기, 마무리로 나누어 복잡한 작업을 분해**하고, **응답결과의 예시를 제공**해 형식을 지킬 수 있게 하였습니다. 동화의 내용과 삽화의 퀄리티가 충요한 프로젝트이기에 **원하는 삽화 톤과 스타일**을 지정해, 퀄리티를 보장했습니다.

Activity

[대전 둔산경찰서] 시위 관리서비스 앱 기획 및 개발] (2023.05 ~ 2023.11)

[Techeer x D.camp x 티타임즈] 실리콘밸리 SW 부트캠프

- 실리콘밸리 Andrew Park (박상현) 엔지니어님이 주최한 부트캠프
- 1:1 면담을 통한 참여인원들의 고충 해결방법 제시
- 각 팀별 발생하는 트러블 슈팅 해결책 제시
- 실리콘밸리 2023 하계/동계 SW 부트캠프 참여 (2023.07 ~ 2023.08) / (2023.12 ~ 2024.02)
- 실리콘밸리 2024 하계 SW 부트캠프 기술멘토 (2024.06 ~ 2024.08)
  - RAG, Langchain 세션 진행
- 실리콘밸리 2024 동계 SW 부트캠프 멘토 (2024.12 ~ 2025.02)

테커 (Techeer) 6기 (2023.09 ~ 현재)

- 실리콘밸리 Andrew Park (박상현) 엔지니어님이 운영하는 대학생 코딩 공부모임
- 실리콘밸리 한달 살기, 기술 컨퍼런스, 해커톤, 네트워킹 행사 등 다수의 활동 주최 및 경험

[한국타이어] R&D DX팀 (2024.03 ~ 2024.08)

[DNLab] 학부연구생 (2024.08 ~ 2025.02)

- Text 기반 라디오 검색 웹 서비스 제작

Skills

Framework	Django, FastAPI
Data	MySQL, Neo4j
LLMs / Agent	Claude3, GPT, Langchain, Hugging Face AWS Bedrock Agent, Langchain Agent
etc.	RabbitMQ, Celery, AWS S3, OpenSearch

Education

충남대학교 컴퓨터융합학부 (2025.08 졸업 예정)