

RAG 파이프라인 구축 프로세스 보고서

작성일 : 2024-12-30

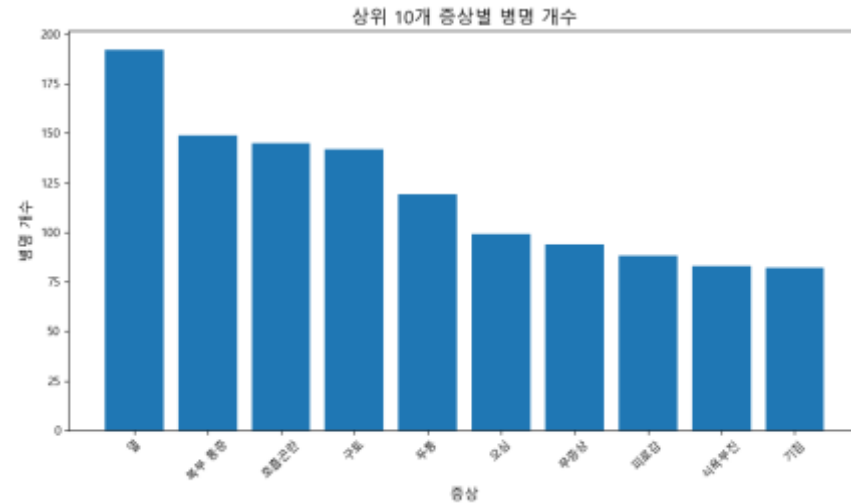
팀 : 광주 1팀

팀원	임동성 1217223
	지용석 1215630
	최유정 1212350
	허현준 1218288
	김민철 1217942

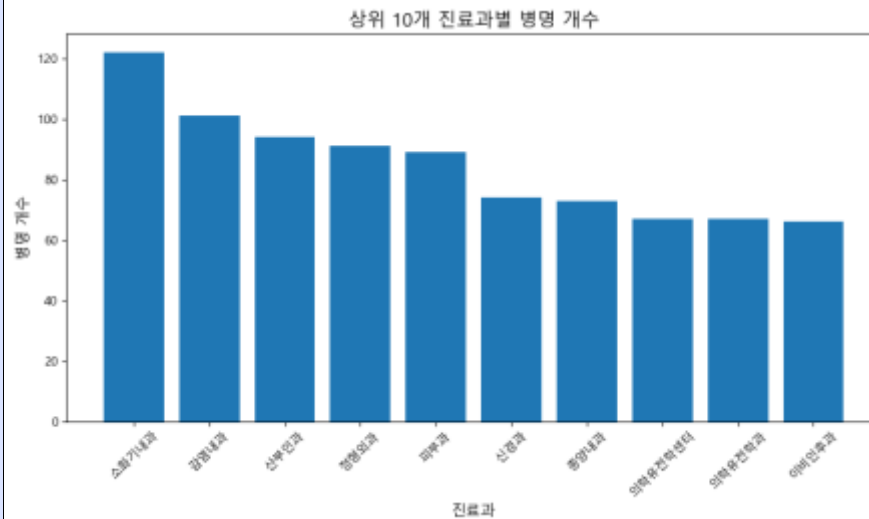
RAG 파이프라인 구축 프로세스 보고서

서비스 명 및 개요	<p>서비스명: 맞춤진료소 추천 서비스</p> <p>서비스 개요:</p> <p>많은 고객이 증상은 있는데, 해당 증상을 가지고 어느 병원을 가야 증상에 맞는 진료를 해주는지에 대한 정보를 얻는데 큰 어려움이 있습니다. 기존의 방식은 비효율적이고 직접 찾아야한다는 한계가 있습니다. 이를 해결하기 병에 따른 증상을 크롤링하고 데이터를 검색해 고객의 증상에 맞는 정확한 진료소를 추천해주는 챗봇 서비스를 설계했습니다. 이 서비스는 고객의 시간을 단축하고, 정확한 진료를 받을 수 있도록 하는데 기여합니다.</p> <p>몸에 아픈 증상이 있는 사람들이 주로 사용하며, 고객이 직접 챗봇을 이용해 알맞은 진료소를 찾을 수 있는 QA 도구로 사용할 수 있습니다.</p>
타겟 사용자 및 시장 분석	<p>타겟 사용자:</p> <ul style="list-style-type: none"> - 예상 사용자 유형: 증상은 있는데 어디를 가야할지 모르는 환자. - 주요 요구사항 및 사용 목적: <p>고객 : 빠르고 정확한 증상에 따른 진료소를 추천받을 수 있다.</p> <p>시장 분석:</p> <p>이 서비스는 단순 검색엔진과는 달리, 질병과 증상에 대한 정보를 가져오기 때문에 사용자가 보다 정확하고, 빠르게 정보를 제공한다는 차별성을 가집니다.</p> <p>고객의 증상에 적합한 정보를 이미 한번 분류하여 db를 만들고 이를 활용해 빠르게 제공하기에 고객의 시간을 단축해주고, 효율적인 답변 생성해준다는 점에서 강한 경쟁력을 가집니다.</p>
목표 및 기대효과	<p>서비스 목표:</p> <ul style="list-style-type: none"> - 사용자의 증상에 신속하고 정확한 진료소를 제공해 고객이 진료소에 가기 전까지의 시간을 단축합니다. <p>기대효과:</p> <ul style="list-style-type: none"> - 고객은 자신의 증상을 입력하면, 직접 검색엔진을 사용해 해당 증상에 따른 진료소를 얻을 수 있습니다. 이에 따라 고객은 만족스러운 진료 서비스를 받을 수 있습니다.
데이터 구성 및 활용	<p>원천데이터 소스 : 서울 아산 병원 건강정보 웹페이지</p> <p>원천 데이터 형식: 웹페이지 형태 -> JSON</p> <p>데이터 처리 방법:</p> <ul style="list-style-type: none"> - 데이터 수집: requests와 BeautifulSoup를 이용해 서울아산병원 건강정보 페이지의 질병 리스트를 확보 - 데이터 전처리: <ul style="list-style-type: none"> - BeautifulSoup를 활용하여 HTML 형태의 텍스트 데이터로 변환 - 질병명, 증상, 권장 진료과 등을 파싱한 뒤, 필요없는 중복 요소나 공백 제거 - 정제된 텍스트를 JSON형태로 가공하여 파일에 저장 <p>데이터 분포</p>

- 특정 증상을 포함하는 병의 개수 (상위 10개)



- 특정 과에 포함되는 병의 개수 (상위 10개)



서울아산병원 건강정보 웹페이지에서 질병 및 증상 데이터를 크롤링하여 JSON 형식으로 변환하였습니다. BeautifulSoup를 활용해 웹페이지의 HTML 구조를 분석하고, 질병명, 주요 증상, 권장 진료과 정보를 추출하였습니다. 크롤링한 데이터는 "질병명" - "주요 증상 및 진료과"의 pair로 구성되었으며, 이를 JSONL 형식으로 저장하여 QA 데이터 셋 생성을 위한 기초 데이터를 준비하였습니다.

RAG
파이프라인
설계

데이터 최적화:

- **Chunk Size:** 1500
- **Overlap:** 200

벡터 데이터베이스 구축 및 임베딩:

- 벡터 DB : Pinecone
- 임베딩 모델 : Upstage Embeddings (**embedding-passage**)
- 벡터 차원 : 자동 계산 (Upstage API 활용)

Pinecone 설정:

- **metric:** Cosine
- **서버리스 사양:** AWS(us-east-1)

Retriever 및 Reranker 구현:

Retriever

방식: Dense Retriever (**Maximum Marginal Relevance** 방식)

구현: PineconeVectorStore

하이퍼파라미터 튜닝 : 반환할 문서 수(k) = 3

검색방식 : MMR (최대 여백 기준 검색 방식)

LLM 프롬프트 설계 및 답변 생성, 평가 :

1/ Task : 맞춤진료소 추천 QA

2/ 프롬프트 예시

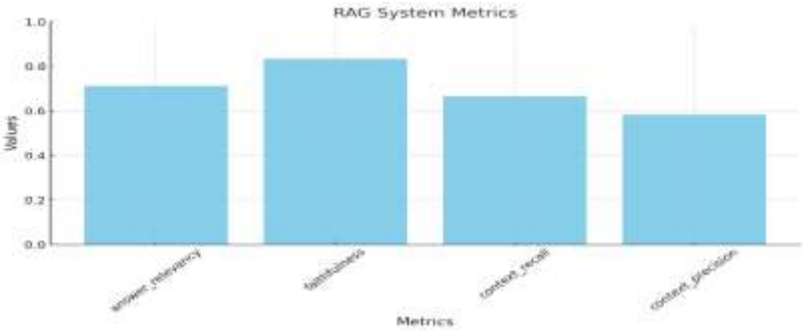
```
prompt = ChatPromptTemplate.from_messages([
    # system prompt
    ("system", """다음 의료정보를 참고하여 답변해줘: {context}
    너는 의료 상담 챗봇이야. 다음 가이드라인을 따라 응답해줘:
    1. 증상 분석: 사용자가 제시한 증상을 체계적으로 분석해줘.
    2. 추천 진료과:
        - 증상을 바탕으로 병문하면 좋을 진료과를 최대 2가지를 추천해줘.
        - 진료과 우선순위를 제시해줘.
    3. 주의사항:
        - 이는 참고용 정보이며, 정확한 진단은 의사의 진찰이 필요함을 명시해줘.
        - 응급 상황으로 판단되면 즉시 응급실 방문을 권고해줘.
    4. 형식:
        - 응답은 [증상 분석], [추천 진료과], [주의사항] 세션으로 구분하여 제공해줘.
        - 전문 의학 용어는 일반인이 이해하기 쉽게 설명해줘.
    """),
    # few-shot prompting
    ("human", "가림과 열이 나고, 목이 간지럽고 마파."),
    ("ai", """[증상 분석]
    목이 아프고 가림이 나는 경우 호흡기 질환일 수 있습니다.

    [추천 진료과]
    1. 이비인후과
    2. 내과과 [이비인후과가 없음 경우]

    [주의사항]
    - 본 정보는 참고용이며, 정확한 진단을 위해서는 반드시 의사의 진찰이 필요합니다.
    - 목이 발열된 경우 조가 진단이 중요하므로 가능한 빨리 진료를 받으시기 바랍니다.
    - 분비물, 통증 등 다른 동반 증상이 있다면 의사에게 함께 말씀해주세요."""),
```

3/ 답변 생성 :

- 생성 모델: ChatOpenAI (**gpt-4o-mini**)
- 온도(Temperature): 0.7 (적당한 창의성과 정보 정확성 유지)

<div>RAG 파이프라인 평가 및 결과</div>	<div><div>평가방법</div><div>정량 평가 : <u>RAGAS</u> 평가 지표</div><div><div>- context_precision: 검색된 의료 정보 중 실제로 문서가 차지하는 비율</div><div>- context_recall: 실제로 관련된 의료 정보를 얼마나 많이 검색했는지 평가</div><div>- faithfulness : 생성된 답변이 검색된 의료 데이터로 얼마나 잘 뒷받침되는지 비율</div><div>- answer_relevancy : 생성된 답변이 사용자 증상 질의와 얼마나 관련성이 있는지 평가가</div></div><div><div>정성 평가</div><div><div>샘플링 방식</div><div>- 무작위로 10 개의 사용자 증상 질문을 선택하여 챗봇의 답변을 평가합니다.</div><div>- 질문은 다양한 의료 시나리오 (호흡기 질환, 소화기 질환, 피부 질환 등)을 기반으로 구성합니다.</div></div><div><div>평가 항목</div><div><div>- 정확성: 생성된 답변이 저장된 의료 데이터와 얼마나 일치하는가?</div><div>- 관련성: 답변이 검색된 의료 데이터와 관련이 있는가?</div><div>- 명확성: 답변이 사용자가 이해하기 쉽고 논리적으로 명확한가?</div></div><div><div>평가 절차</div><div><div>- 각 질문에 대해 생성된 답변을 저장된 문서의 내용과 비교 검토합니다.</div><div>- 관련성이 낮거나 잘못된 답변은 피드백을 기록하여 개선 방안을 도출합니다.</div><div>- 평가 결과를 바탕으로 생성된 답변의 장단점을 정리하고, 추가 최적화 방향을 제시합니다.</div></div><div><div>평가 결과</div><div><div>정량 평가</div><div><div>- Context_precision : 0.5833</div><div>- Context_recall: 0.6667</div><div>- Faithfulness : 0.8333</div><div>- Answer_relevancy : 0.7112</div></div><div><div><div>RAG System Metrics</div><div><table><thead><tr><th>Metric</th><th>Value</th></tr></thead><tbody><tr><td>answer_relevancy</td><td>0.7112</td></tr><tr><td>faithfulness</td><td>0.8333</td></tr><tr><td>context_recall</td><td>0.6667</td></tr><tr><td>context_precision</td><td>0.5833</td></tr></tbody></table></div></div></div></div></div></div></div></div></div>	Metric	Value	answer_relevancy	0.7112	faithfulness	0.8333	context_recall	0.6667	context_precision	0.5833
Metric	Value										
answer_relevancy	0.7112										
faithfulness	0.8333										
context_recall	0.6667										
context_precision	0.5833										

	<p>정성 평가</p> <ul style="list-style-type: none"> - 정확성 : 10 개의 질문 중 10개는 내용과 일치하는 답변 생성. - 관련성 : 10 개 중 10 개가 검색된 문서와 밀접하게 관련 - 명확성 : 모든 답변이 문법적으로 정확하고, 이해하기 쉬운 표현으로 작성됨.
<p>결론 및 향후 발전 방향</p>	<p>결론</p> <p>OpenAI의 GPT-4-mini 모델과 Pinecone 기반 벡터 데이터베이스를 결합한 RAG(Retrieval Augmented Generation) 파이프라인으로 더욱 빠르고 정확한 맞춤 진료소 추천 서비스를 구현했습니다. GPT-4-mini의 뛰어난 추론 능력을 활용하여, 사용자의 질의를 벡터화하고 Pinecone에서 관련 의료 정보를 신속하게 검색한 후, 단순한 정보 제공을 넘어 증상 분석, 진료과 추천, 주의사항 등 맥락에 맞는 체계적인 답변을 제공합니다. 특히, GPT-4-mini의 빠른 처리 속도 덕분에 사용자들은 즉각적인 답변을 얻을 수 있으며, 자연스러운 대화 흐름 속에서 필요한 의료 정보를 효과적으로 파악할 수 있습니다. 이를 통해 사용자는 단순 키워드 검색보다 훨씬 정확하고 효율적인 의료 정보를 얻어, 의료 상담의 초기 단계를 간소화하고 의료 접근성을 높여 사용자 만족도를 극대화할 수 있습니다.</p> <p>향후 발전 방향</p> <ul style="list-style-type: none"> - 언어지원확대: 다국어 지원 기능을 통해 글로벌 사용자를 위한 맞춤형 서비스 제공 - 멀티턴 대화 기능 강화 : 사용자의 의도를 더 깊이 이해하기 위한 컨텍스트 유지 기술 적용 - 긴급 상황 탐지 및 대응 기능 추가 : 응급 상황의 가능성을 탐지하여 사용자에게 적절한 경고 및 빠른 대응 방안을 제시 - 실시간 예약 시스템 연동 : 사용자의 질병과 관련된 진료소와 실시간 예약 기능 구 - 의료 데이터 추가 및 정기 업데이트 : 진료 후기, 의료진 정보, 전문 분야별 상세 정보 등 더욱 풍부한 데이터 확보 / 보험 적용 여부, 비용 정보 등 경제적 요소 관련 데이터 추가 - 추천 알고리즘 고도화 : 사용자의 검색이력, 진료 기록, 건강 상태 등을 분석하여 알고리즘에 적용 - 사용자 인터페이스 및 경험 개선 : 지도기반검색, 특정 질환별 검색 등 다양한 검색 옵션 제공하여 편의성 증가 - 파트너십 확장 : 다양한 의료기관, 건강 관련 기업들과 파트너십을 통해 서비스의 질 향상 및 더 많은 혜택 제공, 전문병원, 종합병원과의 연계를 통한 진료 의뢰 시스템 구축