데이터 수집 및 전처리 문서

LangChain 및 **RAG** 활용 의료 **LLM** 개발

(MediChain)

3차 프로젝트 **5**팀**(**권성호**,** 남의헌**,** 손현성**,** 이준배**,** 이준석**)**

프로젝트명 **:** 윈도우즈
작성일 **:** 2025-06-02
작성자 **:** 손현성

텍스트 데이터 전처리 문서 목차

1. 🔎 문서 개요

이 문서는 [프로젝트명]에 사용된 데이터의 수집 및 전처리 과정을 정리한 문서입니다. 분석의 신뢰성과 재현성을 확보하고자 작성되었습니다.

2. 📂 데이터 수집

2.1 데이터 출처

- 출처 : AI-Hub
- 링크/주소 :[필수의료 의학지식 데이터]
  (https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=data&dataSetSn=71875
)

  [전문 의학지식 데이터]
  (https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=data&dataSetSn=71874
)

  [초거대 AI 헬스케어 질의응답 데이터]

  (https://aihub.or.kr/aihubdata/data/view.do?pageIndex=1&currMenu=115&topMenu=100&srchOptnCnd=OPTNCND001&searchKeyword=%EC%B4%88%EA%B1%B0%EB%8C%80&srchDetailCnd=DETAILCND001&srchOrder=ORDER001&srchPagePer=20&srchDataRealmCode=&aihubDataSe=data&dataSetSn=71762)

2.2 수집 방법

- 사용 도구 : Python, 7zip
- 요약 코드

```
# 압축 해제할 첫 번째 파일 경로와 출력 디렉토리 설정
first_file = os.path.join(temp_dir, new_filenames[0])
output_dir = os.path.join(output_dir_base, base_name)
os.makedirs(output_dir, exist_ok=True)

# 7zip 실행하여 압축 해제
subprocess.run([
    seven_zip_path,  # 7zip 실행 파일 경로
    "x",             # extract 명령
    first_file,      # 압축 해제할 대상 파일
    f"-o{output_dir}"  # 출력 디렉토리
])
```

2.3 원시 데이터 정보

- 필수의료 의학지식 데이터
  1. 데이터 구축 규모
  - 원천데이터**: 101,400,003** 토큰
  - 라벨링데이터**: 19,201** 쌍

  2. 데이터 분포 **(**단위**:** 쌍**)**
  1). domain
  - 산부인과**: 2518**
  - 소아청소년과**: 3087**
  - 응급의학과**: 815**
  - 내과**: 12781**

  2). q_type
  - 객관식**: 15600**
  - 단답형**: 1814**
  - 서술형**: 1787**

  3. 원천데이터

대분류	소분류	포맷	수량	단위
국문	학술 논문 및 저널	JSON	15928056	토큰
	온라인 의료 정보 제공 사이트	JSON	515531	토큰
	정부기관 가이드라인	JSON	0	토큰
	학회 가이드라인	JSON	7709412	토큰
	국제기관 가이드라인	JSON	0	토큰
	의학 교과서	JSON	647538	토큰
	기타 (수술/검사/기타 동의서)	JSON	39799317	토큰
영문	학술 논문 및 저널	JSON	2355433	토큰
	온라인 의료 정보 제공 사이트	JSON	12030958	토큰
	정부기관 가이드라인	JSON	0	토큰
	학회 가이드라인	JSON	0	토큰
	국제기관 가이드라인	JSON	22413758	토큰
	의학 교과서	JSON	0	토큰
	기타 (수술/검사/기타 동의서)	JSON	0	토큰

  4. 라벨링 데이터

대분류	포맷	수량	단위
산부인과	JSON	2518	쌍
소아청소년과	JSON	3087	쌍
응급의학과	JSON	815	쌍
내과	JSON	12781	쌍

- 전문 의학 지식 데이터

1. 데이터 구축 규모
- 원천데이터**: 120,673,006** 토큰
- 라벨링데이터**: 15,286** 쌍

2. 데이터 분포 **(**단위**:** 쌍**)**
(1) domain
- 외과**: 3731**
- 예방의학**: 754**
- 정신건강의학과**: 1986**
- 신경과**/**신경외과**: 2161**
- 피부과**: 744**
- 안과**: 769**
- 이비인후과**: 563**
- 비뇨의학과**: 984**
- 방사선종양학과**: 108**
- 병리과**: 155**
- 마치통증의학과**: 944**
- 의료법규**: 129**
- 기타**: 2258**

(2) q_type
객관식**: 11556**
단답형**: 1796**
서술형**: 1934**

3. 원천데이터

대분류	분류	포맷	수량	단위
국문	학술 논문 및 저널	JSON	1390992	토큰
	온라인 의료 정보 제공 사이트	JSON	60208849	토큰
	정부기관 가이드라인	JSON	0	토큰
	학회 가이드라인313	JSON	935160	토큰
	국제기관 가이드라인	JSON	0	토큰
	의학 교과서	JSON	581216	토큰
	기타 (수술/검사/기타 동의서)	JSON	2364782	토큰
영문	학술 논문 및 저널	JSON	7330141	토큰
	온라인 의료 정보 제공 사이트	JSON	32835105	토큰
	정부기관 가이드라인	JSON	0	토큰
	학회 가이드라인	JSON	1080	토큰
	국제기관 가이드라인	JSON	15021648	토큰
	의학 교과서	JSON	2689	토큰
	기타 (수술/검사/기타 동의서)	JSON	1344	토큰

4. 라벨링 데이터

대분류	포맷	수량	단위
외과	JSON	3731	쌍
예방의학	JSON	754	쌍
정신건강의학과	JSON	1986	쌍
신경과/신경외과	JSON	2161	쌍
피부과	JSON	744	쌍
안과	JSON	769	쌍
이비인후과	JSON	563	쌍
비뇨의학과	JSON	984	쌍
방사선종양학과	JSON	108	쌍
병리과	JSON	155	쌍
마취통증의학과	JSON	944	쌍
의료법규	JSON	129	쌍
기타	JSON	2258	쌍

- 초거대 **AI** 헬스케어 질의응답 데이터

1. 메타데이터 구조

데이터 영역	한국어	**데이터 유형**	텍스트
데이터 형식	.txt	**데이터 출처**	자체 수집, 의료 문서(온라인 기사)
라벨링 유형	질의응답(자연어)	**라벨링 형식**	json
데이터 활용 서비스	의료용 챗봇 서비스	**데이터 구축년도/ 데이터 구축량**	2023년/라벨링 데이터 기준 242,870,922 어절(원천데이터 동일)

2. 최종 증강 데이터 구축량

질문 어절 수	
질환 분류	어절
호흡기질환	540,515
신장비뇨기질환	664,627
순환기질환	746,680
뇌신경정신질환	1,179,013
유방내분비질환	653,752
소아청소년질환	438,335
근골격질환	1,123,839
치과질환	237,203
응급질환	656,608
유전질환	194,578
귀코목질환	788,691
기타	371,874
소화기질환	946,099
여성질환	741,541
눈질환	502,573
감염성질환	959,116
피부질환	906,534
성형미용 및 재건	205,132
종양혈액질환	984,867
전체	12,841,577

의료진 답변 어절 수	
질환 분류	어절
호흡기질환	1,694,986
신장비뇨기질환	3,842,939
순환기질환	2,439,279
뇌신경정신질환	4,957,681
유방내분비질환	2,955,153
소아청소년질환	2,416,840
근골격질환	4,766,769
치과질환	932,301
응급질환	3,116,101
유전질환	1,659,768
귀코목질환	4,130,182
기타	1,392,876
소화기질환	5,040,068
여성질환	4,659,717
눈질환	2,633,944
감염성질환	4,278,142
피부질환	4,524,034
성형미용 및 재건	1,106,395
종양혈액질환	5,111,776
전체	61,658,951

의료진 답변 어절 수	
질환 분류	어절
호흡기질환	1,694,986
신장비뇨기질환	3,842,939
순환기질환	2,439,279
뇌신경정신질환	4,957,681
유방내분비질환	2,955,153
소아청소년질환	2,416,840
근골격질환	4,766,769
치과질환	932,301
응급질환	3,116,101
유전질환	1,659,768
귀코목질환	4,130,182
기타	1,392,876
소화기질환	5,040,068
여성질환	4,659,717
눈질환	2,633,944
감염성질환	4,278,142
피부질환	4,524,034
성형미용 및 재건	1,106,395
종양혈액질환	5,111,776
전체	61,658,951

3. 데이터 분포

- 카테고리는 질환 분류, 질환명, 질문 의도 등으로 구성
- 질환 분류: 감염성 질환, 귀코목 질환, 근골격 질환, 눈질환 등19개
- 질환명: 궤양성 대장염, 급성 심근경색증, 다운증후군 등 주요질환 500여개
- 진료과: 마취통증의학과, 피부과, 응급의학과, 이비인후과, 가정의학과, 일반외과, 내과, 순환기내과, 내분비내과,소화기내과, 혈액종양내과, 감염내과, 신장내과, 호흡기내과, 류마티스내과, 심장내과, 알레르기 내과 등
- 의도: 정의, 증상, 진단, 치료, 예방, 약물, 운동, 재활, 식이/생활 등 11개
- 카테고리별 목표 비율로 데이터 구축

4. 최종 학습용 데이터 구성**(**라벨링 데이터 **JSON**파일수 기준**)**

분류	데이터 구축 총량 (100%)	학습 데이터 (80%)	테스트 데이터 (10%)	검증 데이터 (10%)
질문데이터	1,431,243	1,143,131	146,820	141,292
답변데이터	2,461,960	1,966,902	249,654	245,404
합계	3,893,203 (242,870,922어절)	3,110,033 (194,061,198어절)	396,474 (24,625,829어절)	386,696 (24,183,895어절)

🧹 3. 데이터 전처리

3.1 전처리 목적

- **content** 추출 **:** JSON 구조에서 텍스트 본문(Content)만 추출하여 노이즈를 제거하고 분석 가능한 형태로 정리.
- 불용어 제거 **:** 의미 없는 단어를 제거하여 텍스트의 품질을 높이고, 모델 학습의 효율성을 향상시킴. 한국어 불용어 리스트 및 사용자 정의 리스트를 병행 사용
- 벡터**DB**화 **:** 정제된 텍스트를 jhgan/ko-sroberta-multitask(한국어 특화 모델)과 sentence-transformers/all-MiniLM-L6-v2 (다국어 대응 모델) 임베딩 모델로 벡터화한 후, 유사도 기반 검색을 위해 FAISS를 사용하여 벡터 DB에 저장

3.2 주요 전처리 작업

작업 항목	처리 내용	방법	예시
Content 추출	JSON 구조에서 content 필드만 추출	Python dict 접근 (item['content']) 또는 pandas apply + json.loads	{..., "content": "리뷰 내용"} → "리뷰 내용"
의미없는 문자 제거	특수문자 제거	정규표현식 (re.sub)	“구강암 예방을 위해 금연이 필수적인가요?” -> “구강암 예방을 위해 금연이 필수적인가요”
불용어 제거	분석에 필요 없는 조사, 접속사 등 제거	한국어 불용어 사전 + 사용자 정의 불용어 리스트 적용	"나는 오늘 갔다" → "오늘 갔다"
텍스트 벡터화	문장을 벡터로 변환하여 유사도 비교 가능하도록 함	jhgan/ko-sroberta-multitask 또는 sentence-transformers/all-MiniLM-L6-v2 모델 을 이용한 임베딩	"오늘 날씨 좋다" → [0.12, -0.03, ..., 0.45]
벡터 DB 저장	벡터 데이터를 검색 가능하도록 저장	FAISS를 사용하여 내적 기반 인덱스 생성 및 저장	벡터 → FAISS 인덱스 (IndexFlatIP 등)

📙 4. 벡터DB

4.1 벡터DB 정보

벡터**DB** 파일명 : 벡터DB 파일명
원천데이터 : 벡터 DB제작에 사용된 데이터
임베딩 모델 : 사용한 임베딩 모델명
불용어 제거 유무 : 의료 Q&A 챗봇 학습시 데이터의 불용어 제거를 권장하지 않아 전처리 유무를 나눔
청크 개수 : 벡터DB 청크 개수

벡터**DB** 파일명	원천데이터	임베딩 모델	불용어 제거 유무	청크 개수
QA_random_pair_part1_chunks1.txt	초거대 AI 헬스케어 질의응답 데이터	jhgan/ko-sroberta-multitask	O	8622
QA_random_pair_part2_chunks1.txt	초거대 AI 헬스케어 질의응답 데이터	jhgan/ko-sroberta-multitask	O	2286
index.faiss	전문 의학지식 데이터	sentence-transformers/all-MiniLM-L6-v2	O	299111
pilsu_pro_no_prepro_chunks1.txt	필수의료 의학지식 데이터, 전문 의학지식 데이터	jhgan/ko-sroberta-multitask	X	54813
texts.pkl	초거대 AI 헬스케어 질의응답 데이터, 필수의료 의학지식 데이터, 전문 의학지식 데이터	jhgan/ko-sroberta-multitask , sentence-transformers/all-MiniLM-L6-v2	X	1 (청크로 분류하지 않음)

💾 5. 메타데이터

5.1 메타데이터 생성

카테고리명	설명	값
category	사용자의 질문을 다음과 같은 카테고리로 분류	["medicine", "treatment", "assist_answer", "assist_question", "internal_answer", "internal_question"]
유사도	사용자의 질문에 대한 청크의 유사도	0 ~ 1(float)