데이터 수집 및 전처리 문서

SKN FAMILY AI CAMP 20기

3차 프로젝트_3조(김나현, 박찬, 안채연, 이경현, 이도경)

1. 데이터 수집

본 프로젝트에서는 AI Hub(https://www.aihub.or.kr)에서 제공하는 「반려견 성장 및 질병 관련 말뭉치 데이터」를 활용하였다. 데이터는 크게 두 가지 유형으로 구성된다.

1) 의학지식데이터

의학지식데이터는 수의학 서적 및 논문에서 발췌한 전문 설명문으로, 각 문서는 다음과 같은 JSON 구조를 가진다.

```
{
  "title": "소동물 주요 질환의 임상추론과 감별진단",
  "author": "현창백 내과아카데미 역",
  "publisher": "(주)범문에듀케이션",
  "department": "내과",
  "disease": "<질병·증례 설명 텍스트>"
}
```

---

2)질의응답(QA) 데이터

보호자와 수의사의 실제 상담 기록을 기반으로 대화형 데이터다. 해당 데이터셋은 실제 보호자들의 질의를 참고해 수의사가 직접 답변하는 형식으로 수집됐다.

(- 대한 수의사회에 소속된 각 병원에서 기보유하고 있는 EMR차트 내 상담내용을 토대로 수의사가 직접 답변을 작성
- 펫앤라이프에서 기보유한 쿠팡 로켓펫닥터 질문 내용을 토대로 수의사가 질문과 답변을 직접 작성)

```
{
"meta": {
"lifeCycle": "자견",
"department": "내과",
"disease": "기타"
},
"qa": {
"instruction": "너는 반려견 건강 전문가야....",
"input": "보호자의 실제 질문 내용...",
"output": "수의사 상담 형태의 답변 내용..."
```

```
}
}
```

해당 데이터는 실제 임상 환경과 수의학 지식을 모두 반영하도록 구성되어 있어, 반려견 질병 상담을 위한 RAG 시스템 구축에 적합하다. 의학지식데이터는 수의학적 전문 정보를 제공하는 반면, QA 데이터는 실제 사례 기반의 문맥을 제공하여, 두 데이터 유형이 상호 보완적으로 사용될 수 있다.

2. 데이터 전처리

데이터 활용 목적(RAG 기반 상담 모델 구축)에 맞추어, 모든 원천 데이터를 **Document 형태**로 변환하였다.

이 과정에서 데이터 유형을 구분하기 위해 `source_type` 메타필드를 추가하였다.

데이터 유형	추가 메타필드	설명
의학지식데이터	`"source_type": "medical_data"`	질병 해설·정의·시술 설명 등 긴 설명문으로 구성
QA 상담기록	`"source_type": "qa_data"`	실제 상담 기록(질의응답 형태)

원본 메타데이터는 모두 유지했으며, 불필요한 특수문자는 제거하였다.

3. 텍스트 분할(Chunking)

의학 데이터와 QA 데이터는 텍스트 성격이 다르기 때문에, 서로 다른 청킹 전략을 적용하였다.

1. 의학지식데이터(chunk_size=500)

- 긴 설명문 중심 → 과도한 분절 시 의미 손실 발생
- `RecursiveCharacterTextSplitter` 사용
- `chunk_size=500`, `chunk_overlap`는 최소화

2. QA 데이터(chunk_size=800)

- 질문과 답변은 하나의 맥락을 이루므로 더 큰 단위 유지 필요
- 동일한 Splitter 사용하되 `chunk_size=800`으로 길게 유지

3. 청킹 결과

- **총 문서(chunk) 수:** 약 **34,600개 생성**