**SK네트웍스 Family AI과정 9기 4팀 PetMind 1**

**김도연, 김정훈, 김하늘, 이다인, 임수연**

**수집된 데이터 전처리 문서**

* 개요
  + 제출일자: 2025-05-12
  + 작성자: 김도연
  + 내용 : 수집한 데이터의 전처리 과정과 이유를 서술함

1. RAG 콘텐츠 데이터

| 데이터 수집 개요 | * 수집 목적 : RAG용 * 수집 항목 : 반려견 관련 정보 콘텐츠 * 수집 기간 : 2025-04-23 ~ 2025-05-08 * 수집 출처 : 핏펫 블로그, 하이닥, 최영민동물 의료센터, 비마이펫 플랫폼 * 수집 방법 : 웹 크롤링(BeautifulSoup, Requests 라이브러리 사용) * 데이터 크기: 7.7MB * 데이터 수 : 1879개 * 데이터 컬럼 : title(제목), content(본문 내용), URL(링크), image(대표이미지) |
| --- | --- |
| 데이터 저장 및 관리 | * 데이터 파일명: merged\_rag\_data.csv * 저장 형식: CSV (UTF-8 인코딩) * 저장 경로: 구글 드라이브 내 공유 드라이브 폴더 * 관리 방법: 데이터 버전 관리(Google Drive Revision History 사용) |
| 데이터 전처리 과정 | * **문장 분리 및 줄바꿈 정리**  문단 내 텍스트를 마침표(.), 물음표(?), 느낌표(!), 물결표(~) 등의 문장 종결 기호를 기준으로 문장 단위로 분리하고, 다시 하나의 텍스트로 연결함. 문장 간 공백은 1칸으로 통일하고, 중복 공백과 줄바꿈 문자(\r, \n)를 제거하여 하나의 연속된 문장으로 만듦. * **특정 키워드 포함 문장 제거**  "오시는 길", "이 블로그", "NAVER Corp", "서울특별시", "핏펫", "다음 글", "TV동물농장", "최영민", "동물병원", "진료문의" 등 광고성·홍보성 키워드가 포함된 문장을 삭제하여 데이터 순수성을 확보함. * **불필요한 서두 및 특정 문구 제거**  기사 서두에 등장하는 작성자 및 직함 정보(예: ❚ 이수형 샤인동물메디컬센터 마취과장)나 "강남 동물병원 추천", "강남 동물병원" 같은 특정 키워드 구문을 삭제함. 키워드 사이 공백 수는 유동적으로 허용함. * **특수문자 및 불필요 문자 정리**   + '▶'는 쉼표(,)로 변환하여 자연스러운 문장 흐름 유지.   + \xa0(비표준 공백)는 일반 공백으로 변환.   + 기타 특수문자(★, ❚, △, ㅣ, …, “”, ‘’, !, ", ', - 등)는 삭제.   + (타이틀 처리 시 ., ?, !도 추가 삭제함)   + 허용하는 문자: 한글, 영문, 숫자, 공백, 마침표(.), 물음표(?), 느낌표(!), 퍼센트 기호(%). * **타이틀(title) 정리**  제목 끝에 붙는 "- 핏펫", "- 핏펫 블로그" 문구를 삭제하고, "수의사 칼럼 '...'" 또는 "수의사 칼럼 "..." 형식의 경우 따옴표 안 핵심 문장만 추출하여 제목을 간결화함. 또한 타이틀에서도 특정 키워드(예: "강남 동물병원 추천")를 삭제하고 중복 공백을 정리함. * **해시태그 제거**  텍스트 내 #으로 시작하는 모든 해시태그(예: #노령견안과질환, #노령견백내장치료)를 삭제하여 메타데이터 정보를 제거함. * **소문자 변환**  모든 텍스트와 타이틀을 소문자로 변환하여 대소문자에 의한 데이터 불일치를 방지하고 일관성을 유지함. * **괄호 및 대괄호 안 내용 제거**  소괄호 () 및 대괄호 [] 안에 포함된 모든 내용을 삭제하여 부가 설명이나 카테고리 정보를 제거하고, 본문 내용만 남김. * **특정 줄 제거** (content에만 적용)  텍스트를 줄 단위로 분리한 뒤, '글 ='이 포함된 줄을 찾아 삭제하여 메타 정보 또는 작성자 정보 제거. * **최종 공백 정리**  모든 전처리 후 여러 개의 연속된 공백은 하나로 통일하고, 텍스트 양쪽의 불필요한 공백을 제거하여 데이터 형태를 깔끔하게 마무리함. |
| 데이터 전처리 결과 | * **원시 데이터**  | index | contents | | --- | --- | | 3 | 이개혈종의 치료는 피고름을 빼내는 치료를 하고  1주 이상 붕대로 고정해야 합니다.  문제는 관리가 필요한데 붕대를 자꾸 강아지가  풀어버리고 귀를 다시 긁으면서 재발하는 경우가  많습니다.  #강아지귓병치료  따라서 강아지 이개혈종 치료는 치료 뿐 아니라  관리가 매우 중요합니다.  자주 붕대를 갈아주고 병원에 주기적으로 내원하여  신경을 써야 하므로 보호자의 관심이 필요한 질환이  강아지 이개혈종 귀질환 입니다.  #강아지귀치료동물병원  TV동물농장 자문위원이신 최영민 원장님께 강아지를  위한 효과적인 치료를 받도록 해주세요~  #강남강아지동물병원  모든 동물을 특별하게 케어하는  최영민동물병원  진료문의는 전화로 안내해 드립니다.  최영민동물병원 오시는 길  50m  © NAVER Corp.  최영민동물의료센터  서울특별시 강남구 논현로132길 24 양현빌딩2층 최영민동물의료센터  이 블로그의 체크인  이 장소의 다른 글 |  * **클리닝 데이터**  | index | contents | | --- | --- | | 3 | 이개혈종의 치료는 피고름을 빼내는 치료를 하고 1주 이상 붕대로 고정해야 합니다. 문제는 관리가 필요한데 붕대를 자꾸 강아지가 풀어버리고 귀를 다시 긁으면서 재발하는 경우가 많습니다. 따라서 강아지 이개혈종 치료는 치료 뿐 아니라 관리가 매우 중요합니다. 자주 붕대를 갈아주고 병원에 주기적으로 내원하여 신경을 써야 하므로 보호자의 관심이 필요한 질환이 강아지 이개혈종 귀질환 입니다. | |

1. fine-tuning 학습용 QnA 데이터

| 데이터 수집 개요 | * 수집 목적 : 모델 fine-tuning 학습용 * 수집 항목 : 반려견 관련 질의응답 데이터 * 수집 기간 : 2025-04-23 ~ 2025-05-01 * 수집 출처 : <https://wayopet.com/trainer/qna> * 수집 방법 : 웹 크롤링(BeautifulSoup, Requests 라이브러리 사용) * 데이터 크기: 9,364KB * 데이터 수 : 2378개 * 데이터 컬럼 : url, 질문 제목, 증상과 행동, 시작된 시점, 보호자님 반응, 원인 분석, 솔루션 제안 |
| --- | --- |
| 데이터 저장 및 관리 | * 데이터 파일명: 6\_wayo\_qna\_data.json * 저장 형식: json (UTF-8 인코딩) * 저장 경로: 구글 드라이브 내 공유 드라이브 폴더 * 관리 방법: 데이터 버전 관리(Google Drive Revision History 사용) |
| 데이터 전처리 과정 | * **서두 이모티콘 기반 텍스트 제거**   + 텍스트의 시작 부분에 등장하는 :-) 또는 :) 기호(옵션으로 - 포함)까지의 모든 내용을 삭제함.   + 이모티콘 이전의 불필요한 인삿말, 잡담 등을 제거하여 본문 핵심 내용만 남겨 텍스트를 간결하고 직관적으로 만듦. * **텍스트 내 특수문자 제거**   + :-) , ^^ , ^\_^ , --;; 등 문자 기반 이모티콘을 삭제하여 가독성 및 정제도를 높임.   + 이모티콘 제거 후, 특수문자를 정리함. 단, 마침표(.), 쉼표(,)는 남기고 나머지 특수문자는 모두 삭제하여 순수한 텍스트 데이터만 남김. * **텍스트 내 모든 이모지 제거**   + 표준 유니코드 범위에 해당하는 모든 이모지(😂, 🐶, 🚗, ❤️ 등)를 정규표현식으로 탐지하여 삭제함.   + 다양한 범위 (이모티콘, 심볼, 플래그, 지도 아이콘 등) 를 포함하여 이모지를 완전히 제거함.   + 특수 유니코드 기호(⚡, ☀️ 등) 및 zero-width joiner(이모지 조합 기호)도 함께 제거하여 텍스트를 깔끔하게 만듦. |
| 데이터 전처리 결과 | * **원시 데이터**  | index | content | | --- | --- | | 0 | "url": "https://wayopet.com/trainer/qna/epilxkb2kmxl7ey5",  "질문 제목": "사료는 안먹는데 똥은 먹어요",  "증상과 행동": "사료를 배식해주면 몇시간을 안먹고 공복토도 했는데 그때만 먹고 다음 끼니는 또 안먹네요\n\n그런데 똥은 싸고 꼭 먹습니다\n\n왜 식사 훈련이나 식분증 훈련을 어떻게 해야할까요\n똥은 그 자리에서 먹는게 아니라 물고 나와서 방석이나 노는 공간에서 나눠 먹고 다 먹지도 않습니다",  "시작된 시점": "사료 안먹는건 한달 전\n식분은 2개월 전",  "보호자님 반응": "아침 사료는 노즈워크로 전부 세팅 (퇴근하고 오면 다 먹어 놓았습니더)\n똥은 그냥 치우기만 하고 같이 있을 때 똥 먹으려 하면 안돼! 하면서 크게 얘기해서 거부시그널 줬습니다",  "원인 분석": "안녕하세요 아리 보호자님 :-) 벌써 6개월이 지난 아리가 되었군요! 이전 6월 1일에 답변 드렸던 식분증에 관한 문의글을 남겨주셨던 것으로 확인이 되는데요 당시 답변 드렸던 내용을 토대로 함께 진행해주시고 더불어 추가 답변 도와드리겠습니다.",  "솔루션 제안": "사료를 먹지 않는 다는 것은 다른 대체제는 먹고 있다는 것으로 예상이 되는데요, 평소 육류의 간식 및 노즈워크 제공을 해주기 위한 여러 간식을 제공하실 경우 사료를 먹지 않고자 할 수 있습니다.\n\n이전 말씀 드린것과 같이 산책과 퍼피 트레이닝(사회화 교육)은 충분히 하고 계실지 궁금한데요, 한창 호기심이 많은 시기에는 이러한 것들을 해소하기 위해 자신의 배변을 먹고자 할 수 있으니 이점 참고하시어 이전 글과 함께 적용해주시면 좋겠습니다!\n\n이제 6개월이 지난 시점에는 이전에는 없던 짖음, 분리 불안 등 다양한 문제점들이 보일 수 있습니다. 현재 크게 문제 행동이 발현 되지 않는다고 하여도 미리 예방을 해보시면 좋을 것 같습니다 :)\n\n아리와 건강한 반려생활을 위해 여러모로 노력해주시느라 고생이 많습니다! 그럼 추후 아리 소식 또 기다리겠습니다🥰" |  * **클리닝 데이터**  | index | content | | --- | --- | | 0 | "url":"https:\/\/wayopet.com\/trainer\/qna\/epilxkb2kmxl7ey5",  "질문 제목":"사료는 안먹는데 똥은 먹어요",  "증상과 행동":"사료를 배식해주면 몇시간을 안먹고 공복토도 했는데 그때만 먹고 다음 끼니는 또 안먹네요. 그런데 똥은 싸고 꼭 먹습니다. 왜 식사 훈련이나 식분증 훈련을 어떻게 해야할까요. 똥은 그 자리에서 먹는게 아니라 물고 나와서 방석이나 노는 공간에서 나눠 먹고 다 먹지도 않습니다",  "시작된 시점":"사료 안먹는건 한달 전\n식분은 2개월 전",  "보호자님 반응":"아침 사료는 노즈워크로 전부 세팅 (퇴근하고 오면 다 먹어 놓았습니더)똥은 그냥 치우기만 하고 같이 있을 때 똥 먹으려 하면 안돼! 하면서 크게 얘기해서 거부시그널 줬습니다",  "원인 분석":"벌써 6개월이 지난 아리가 되었군요 이전 6월 1일에 답변 드렸던 식분증에 관한 문의글을 남겨주셨던 것으로 확인이 되는데요 당시 답변 드렸던 내용을 토대로 함께 진행해주시고 더불어 추가 답변 도와드리겠습니다.",  "솔루션 제안":"사료를 먹지 않는 다는 것은 다른 대체제는 먹고 있다는 것으로 예상이 되는데요, 평소 육류의 간식 및 노즈워크 제공을 해주기 위한 여러 간식을 제공하실 경우 사료를 먹지 않고자 할 수 있습니다. 이전 말씀 드린것과 같이 산책과 퍼피 트레이닝사회화 교육은 충분히 하고 계실지 궁금한데요, 한창 호기심이 많은 시기에는 이러한 것들을 해소하기 위해 자신의 배변을 먹고자 할 수 있으니 이점 참고하시어 이전 글과 함께 적용해주시면 좋겠습니다. 이제 6개월이 지난 시점에는 이전에는 없던 짖음, 분리 불안 등 다양한 문제점들이 보일 수 있습니다. 현재 크게 문제 행동이 발현 되지 않는다고 하여도 미리 예방을 해보시면 좋을 것 같습니다. 아리와 건강한 반려생활을 위해 여러모로 노력해주시느라 고생이 많습니다 그럼 추후 아리 소식 또 기다리겠습니다" | |