

**SK네트웍스 Family AI 과정 10기  
 데이터 전처리 인공지능 데이터 전처리 결과서**



| **산출물 단계** | 데이터 전처리 |
| --- | --- |
| **평가 산출물** | 인공지능 데이터 전처리 결과서 |
| **제출 일자** | 2025.08.29 |
| **깃허브 경로** | https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN14-FINAL-6Team |
| **작성 팀원** | 서은선 |

1. **문서 개요**

* 프로젝트명: 신선식품 플랫폼 고객 응대 챗봇
* 전처리 목적: RAG 학습용 데이터 정제 및 벡터DB 구축
* 문제 정의:  
  마켓컬리 FAQ 및 이용약관 데이터를 구조화하여, 고객 문의에 대한 정확한 검색·응답이 가능한 데이터셋 구축

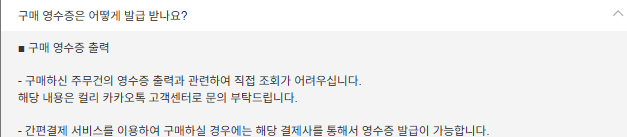
1. **데이터셋 개요**

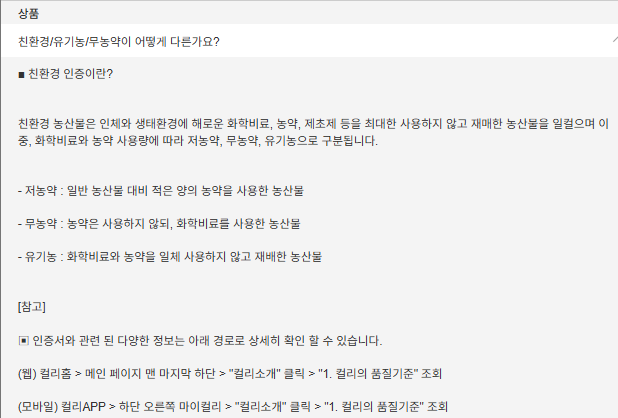
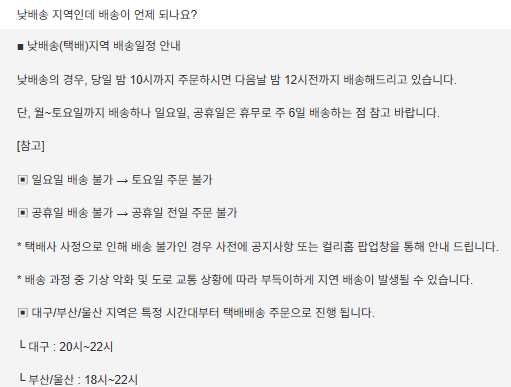
* 데이터 출처 및 수집 방법:  
  출처: 마켓컬리 공식 사이트 FAQ

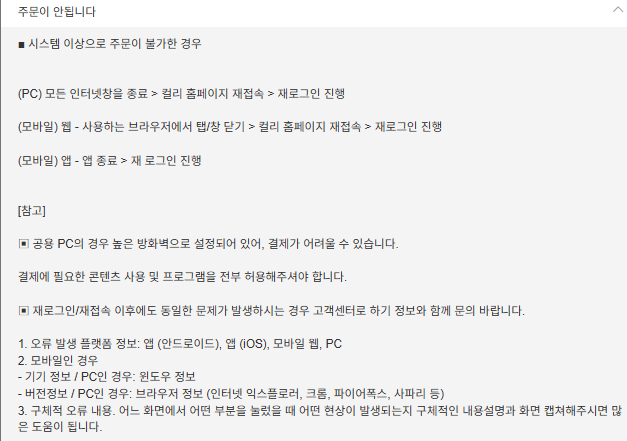
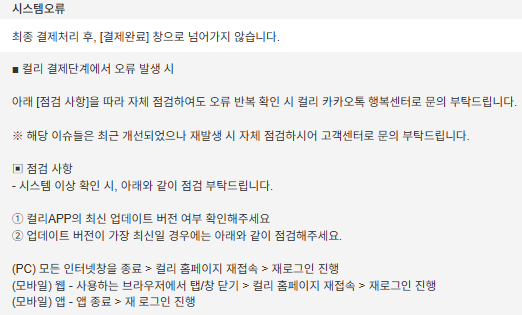
수집 방법: 공식 페이지에서 직접 발췌 후 JSON 파일로 정리

* 데이터 구성:

| **항목명** | **설명** | **예시** |
| --- | --- | --- |
| category | 질문 종류 | 취소/교환/환불 |
| question | 질문 | “서비스 이용 기준이 있나요?” |
| answer | 답변 | “실결제금액이 15,000원 이상일 경우, 자동으로 무료배송 혜택이 적용됩니다.” |

* 원본 데이터 샘플(5~10건 첨부):  
   

1. **전처리 프로세스 개요**

* 전체 흐름도:  
    
   ① 수집 → ② 정규화(텍스트 전처리) → ③ 정규화(고유명사 치환)
* 전처리 파이프라인 요약:

| **단계** | **목적** | **수행 작업** | **사용 도구/라이브러리** |
| --- | --- | --- | --- |
| 정규화 | 텍스트 전처리 | 특수문자 치환 (■▣▶※◆◇● → ■) | Python-re |
| 정규화 | 고유명사 치환 | “마켓컬리” -> “Qook” | Python-re |

1. **세부 전처리 단계**

4.1 정규화

* 텍스트 정규화::  
  + 특수문자 치환: “▣ Qook패스” → “■ Qook패스”
  + 고유명사 치환: “마켓컬리” -> “Qook”
* 사용 라이브러리: re
* 코드 예시 (Python):
* text = re.sub(r"[■▣▶※◆◇●]", "■", text)
* return re.sub(r"(마켓컬리|컬리)", "Qook", text)

5. **전처리 결과 요약 및 평가**

* 품질 향상 지표:  
  + 특수문자 치환: 137건 정제
  + 고유명사 치환: 110건 정제
* 향후 활용 방안:  
  + RAG 기반 정보 검색 및 응답 생성

1. **변경 이력**

| 변경일 | 변경자 | 변경 내용 | 비고 |
| --- | --- | --- | --- |
| 2025-08-28 | 서은선 | 마켓컬리/컬리 글자 치환 | 플랫폼 명칭 통일 |

1. **문서 개요**

* 프로젝트명: 신선식품 플랫폼 고객 응대 챗봇
* 전처리 목적: RAG 학습용 데이터 정제 및 벡터 DB 구축
* 문제 정의:  
  마켓컬리 FAQ 및 이용약관 데이터를 구조화하여, 고객 문의에 대한 정확한 검색·응답이 가능한 데이터셋 구축

1. **데이터셋 개요**

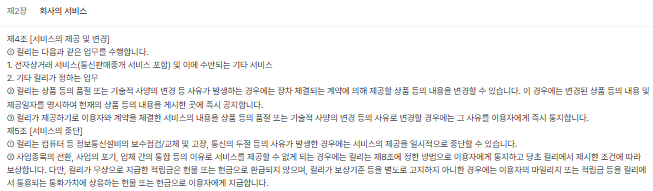
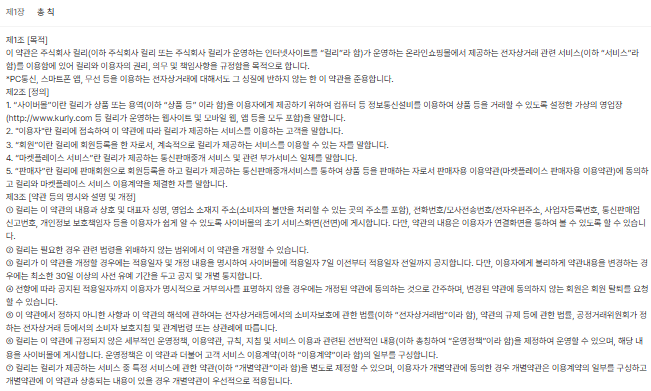
* 데이터 출처 및 수집 방법:  
  출처: 마켓컬리 공식 사이트 이용약관 페이지

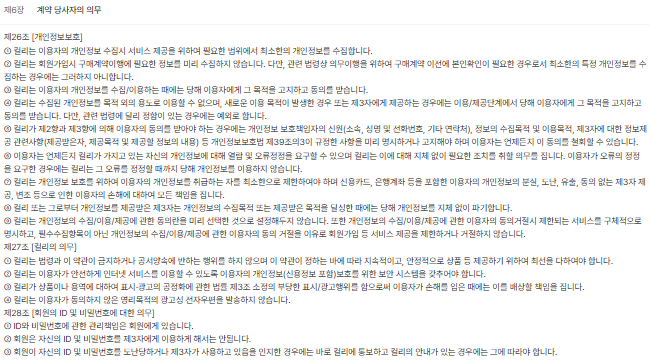
수집 방법: 공식 페이지에서 직접 발췌 후 JSON 파일로 정리

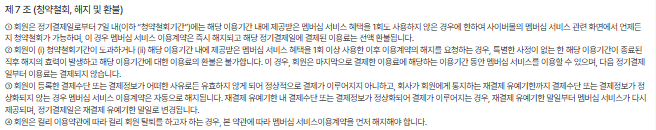
* 데이터 구성:

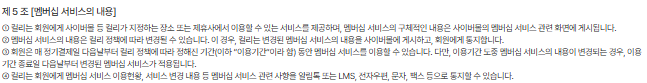
| **항목명** | **설명** | **예시** |
| --- | --- | --- |
| article\_number | 조항 번호 | 마켓컬리 이용약관 제1조 |
| title | 조항 제목 | 서비스의 제공 및 변경 |
| content | 조항 내용 | 마켓컬리가 제공하기로 이용자와 계약을 체결한 서비스의 내용을 상품 등의 품절 또는 기술적 사양의 변경 등의 사유로 변경할 경우에는  그 사유를 이용자에게 즉시 통지합니다. |

* 원본 데이터 샘플(5~10건 첨부):









1. **전처리 프로세스 개요**

* 전체 흐름도:  
    
   ① 수집 → ② 구조화 → ③ 정규화
* 전처리 파이프라인 요약:

| **단계** | **목적** | **수행 작업** | **사용 도구/라이브러리** |
| --- | --- | --- | --- |
| 구조화 | 약관 조항 세부 구조 구분 | 최상위 조항, 숫자 항목,  한글 소분류 구분 | Python-re |
| 정규화 | 고유명사 치환 | “마켓컬리” -> “Qook” | Python-re |

1. **세부 전처리 단계**

4.1 구조화

* 구조화 방식:
* 최상위 조항(①,②,…) → 새로운 Document 시작
* 숫자 항목(1,2,3…) → / 기호로 구분
* 한글 소분류(가,나,다…) → // 기호로 구분
* 코드 예시 (Python):

# ①,②,… 최상위 항목

if re.match(r'^[①-⑳]', line):

if buffer and current\_article:

docs.append(Document(

page\_content=f"{current\_article} " + " ".join(buffer),

metadata={

"article": item["article\_number"],

"title": item["title"],

"type": "terms"

} ))

elif re.match(r'^\d+', line):

buffer.append("/" + line)

elif re.match(r'^[가-힣]', line):

if buffer:

buffer[-1] += " //" + line

else:

buffer.append("//" + line)

4.2 정규화

* 텍스트 정규화:
  + 고유명사 치환: “마켓컬리” -> “Qook”

* 사용 라이브러리: re
* 코드 예시 (Python):

content = re.sub(r"(마켓컬리|컬리)", "Qook", content)

1. **전처리 결과 요약 및 평가**

* 품질 향상 지표:  
  + 계층 구조 구분하여 정제
  + 고유명사 치환: 296건 정제
* 향후 활용 방안:  
  + RAG 기반 정보 검색 및 응답 생성

1. **변경 이력**

| 변경일 | 변경자 | 변경 내용 | 비고 |
| --- | --- | --- | --- |
| 2025-08-28 | 서은선 | 숫자 항목 분류 앞에  ‘/’ 기호 추가 | 계층 구조를 통한 검색 안정성 확보 |

1. **문서 개요**

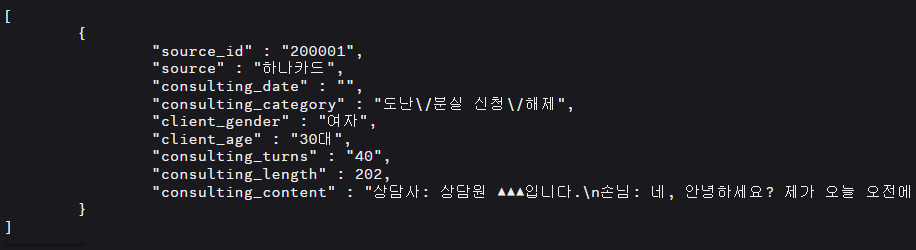
* 프로젝트명: 신선식품 플랫폼 고객 응대 챗봇
* 전처리 목적: sLLM 상담사의 말투와 대화 패턴을 학습 데이터셋 구축
* 문제 정의: 원본 상담 데이터(금융사, 여행사, 통신사 등)에는 개인정보, 불필요한 반복, 비표준화된 표현이 포함되어 정제·표준화 과정을 통해 고객–상담사 멀티턴 대화 데이터셋을 구축

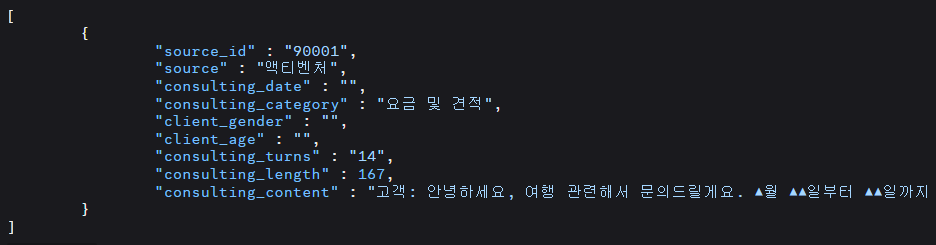
1. **데이터셋 개요**

데이터 구성:

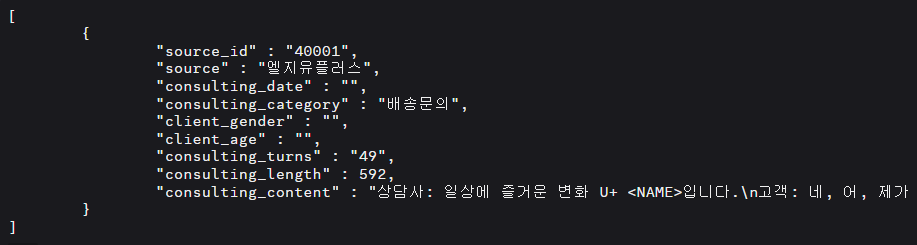
| **항목명** | **설명** | **예시** |
| --- | --- | --- |
| source\_id | 원본 데이터 고유 식별자 | 200001 |
| source | 출처(기관/기업명) | 하나카드 / 액티벤처 / 엘지유플러스 |
| consulting\_date | 상담 일자 | (공란 또는 YYYY-MM-DD) |
| client\_gender | 고객 성별 | 여자 / 공란 |
| client\_age | 고객 연령대 | 30대 / 공란 |
| consulting\_turns | 대화 턴 수 | 14 / 40 / 49 |
| consulting\_length | 대화 전체 길이(토큰) | 202 / 167 / 592 |
| consulting\_content | 원문 대화 | 다중 발화 포함 (상담사/고객 대화) |

### 금융사 (하나카드): 카드 도난/분실 해제 관련 문의



* 액티비티 여행사: 가족 여행 일정 및 숙소 예약 문의

### 통신사(엘지유플러스): 사은품/보험/요금제/정지 해제 등 복합 문의



1. **전처리 프로세스 개요**

* 전체 흐름도:  
    
   ① 수집 → ② 구조화
* 전처리 파이프라인 요약:

| **단계** | **목적** | **수행 작업** |
| --- | --- | --- |
| 구조화 | Q&A 구조화 | 상담사 및 손님 발화 분리 |

1. **세부 전처리 단계**

4.1 구조화

* 구조화 방법
  + “손님” 발화 -> “Text”
  + “상담사” 발화 -> “Completion”
  + 한 턴 단위로 JSON 포맷 생성
* 코드 예시 (Python):

for line in lines:

if line.startswith("손님:"):

# '손님:' 제거 후 발화 내용만 추출

current\_text = line.replace("손님:", "").strip()

elif line.startswith("상담사:") and current\_text:

# '상담사:' 제거 후 발화 내용만 추출

completion = line.replace("상담사:", "").strip()

qa\_pairs.append({

"Text": current\_text,

"Completion": completion

})

1. **학습/검증 데이터 분리**

* 분리 기준 및 방법:  
  + 기준: 별도 분리 없음
  + 사유: 데이터 출처(AI-Hub)에서 이미 학습/검증 세트로 분리된 상태로 배포 -> 추가 분리 불필요함

1. **전처리 결과 요약 및 평가**

* 품질 향상 지표:  
  + Q&A 구조화 완료 → 모델 학습 데이터셋으로 직접 활용 가능
* 향후 활용 방안:  
  + 민원 상담 맥락에서 손님/상담사 말투를 각각 학습시켜 모델의 자연스러운 응대를 위함

1. **변경 이력**

| 변경일 | 변경자 | 변경 내용 | 비고 |
| --- | --- | --- | --- |
| 2025-08-27 | 김광령 | 상담사/손님 발화 구분 | 손님·상담사 화법 구분을 통한 파인튜닝 데이터셋 구축 |