터 

**SK네트웍스 Family AI 과정 15기  
 모델링 및 평가 수집된 데이터 및 전처리 문서**



| **산출물 단계** | 모델링 및 평가 |
| --- | --- |
| **평가 산출물** | 수집된 데이터 및 전처리 문서 |
| **제출 일자** |  |
| **깃허브 경로** | 깃허브 주소 기재 |
| **작성 팀원** |  |

1. **데이터 수집 개요**

데이터 수집 기간:

수집 방식: (크롤링 / API / 수기입력 / 기타)

1.1. 수집 목적

수집 목적: 예) 사용자 챗봇 로그 데이터를 기반으로 상담 태그 분류 모델을 개발하기 위함.

1.2 데이터 출처

| **데이터명** | **출처** | **수집 방식** | **형식** |
| --- | --- | --- | --- |
| 상담 대화 로그 | 내부 DB | API 수집 (JSON) | JSON |
| 공개 문서 FAQ | 공공데이터포털 | 웹 크롤링 | CSV |

1.3 수집 대상 및 범위

* 대상: 예) 사용자 질의 및 응답 로그 (2023.01 ~ 2023.12)
* 총 수집 건수: 약 12,000건
* 개인정보 포함 여부: 예/아니오
* 민감 정보 여부 및 조치 사항: 예) 이름, 전화번호 → 수집 전 익명화 처리

1.4 법적·윤리적 고려 사항

* 개인정보보호법 준수 여부: O
* 데이터 수집 사전 동의 유무: O (서비스 약관에 명시됨)
* 자동 크롤링 시 robots.txt 확인 여부: O
* 수집 데이터의 활용범위 명시 여부: O

1. 데이터 저장 및 검증

**2.저장 방식**

* 저장 경로: 예) /data/raw\_data/consult\_log.json
* 저장 형식: JSON / CSV / SQLite 등
* 일관성 확보 방법: Key 유효성 검사, 필드 누락 검출

2.1 중복 및 정합성 검증

* 중복 제거 기준: Q&A 쌍, timestamp 기준 중복 제거
* 정합성 확보 방법:  
  + 날짜 필드 timestamp 형태 통일
  + 빈 질문/답변 제거
  + 텍스트 필드 이상값 필터링 (예: 문자 수 2 미만 제거)

**3. 데이터 전처리 절차**

3.1 이상치 탐지 및 처리

* 이상치 기준: 동일 질문에 상이한 태그 존재 / 부적절한 문장
* 처리 방법:  
  + 단어 수 2 이하 문장 제거
  + 욕설 필터링, 태그 누락 항목 제외
* 품질 기여 설명:  
   예) 태그 누락 및 중복 질문 제거로 학습 정확도 4.2%p 개선

3.2 결측치 처리

* 결측 필드: 질문, 응답, 태그 필드 일부 누락
* 처리 방법: 누락 응답 → “정보 없음” 대체 또는 해당 row 제거
* 적용 비율: 전체 데이터 중 약 3.1% 제거

3.3 데이터 변환 및 재현 가능성

* 텍스트 정제:  
  + 소문자화, 특수문자 제거, 이모지 필터링
* 토큰화:  
  + 예) KoNLPy 형태소 분석기 사용 (Twitter)
* 변환 코드 예시:  
    
   (예시)  
   import re  
   def clean\_text(text):  
   text = text.lower()  
   text = re.sub(r'[^가-힣a-zA-Z0-9\s]', '', text)  
   return text
* GitHub 또는 Colab 링크: [전처리 코드 링크 첨부]

3.4 전처리 절차 요약 (프로세스 플로우)

(도식 삽입 또는 순서대로 작성)

1. 데이터 수집
2. 중복 및 결측 제거
3. 텍스트 정제
4. 이상치 제거
5. 토큰화
6. 저장 (train.csv / test.csv)
7. 데이터셋 분리

* 학습/검증/테스트 비율:

| 구분 | 건수 | 비율 |
| --- | --- | --- |
| Train | 8,400 | 70% |
| Validation | 1,800 | 15% |
| Test | 1,800 | 15% |

* 분리 기준: 무작위(Random State=42), 클래스 비율 균형 유지 (Stratified Split)

**4. 부록**

* 수집 데이터 샘플 3건 (표 또는 JSON)
* 전처리 전/후 비교 예시
* 전처리 코드 전체 링크 또는 주요 코드 첨부
* 참조 문헌 및 API 문서 (공공 데이터, 크롤링 페이지 등)