

**SK네트웍스 Family AI과정 14기  
 모델링 및 평가 수집된 데이터 및 전처리 문서**



**□ 개요**

* 산출물 단계 : 모델링 및 평가
* 평가 산출물 : 수집된 데이터 및 전처리 문서
* 제출 일자 : 2025.09.12
* 깃허브 경로 : https://github.com/skn-ai14-250409/SKN14-Final-2Team
* 작성 팀원 : 한성규

## 1. 개요

### 1.1 데이터 설명

**본 문서는 ScentPick 프로젝트에서 사용자 자연어 설명으로부터 메인 어코드를 추출하기 위한 다중 라벨 분류 모델 학습에 필요한 데이터 수집 및 전처리 과정을 정리한 것이다. 데이터 출처, 데이터 유형, 전처리 단계, 라벨 분포 분석, 학습/평가 데이터 분할 및 산출물에 대해 기술한다.**

### 1.2 데이터 수집 목적

**ScentPick은 사용자의 향수 취향을 분석하여 맞춤형 추천을 제공하는 서비스다. 자연어 입력을 통해 추출된 설명에서 메인 어코드(향의 주요 특징)를 분류하기 위해, 다국어 입력을 처리할 수 있는 멀티라벨 학습 데이터셋이 필요하다.**

## 

## 

## 

## 2. 데이터 수집

### 2.1 데이터 출처

* **- 내부 데이터: perfume\_final.csv (향수 설명 텍스트 1,394건)**
* **- 외부 데이터: HuggingFace doevent/perfume (향수 description 및 fragrance 라벨, 26320건)**

### 2.2 데이터 유형

* **텍스트 데이터 (설명 description, 라벨 fragrances)구조**

## 3. 데이터 전처리

### 3.1 전처리 개요

* **1) 불필요 공백/특수문자 제거**
* **2) 브랜드명 및 불용어 제거**
* **3) 희소 라벨 제거 (RARE\_MIN\_COUNT = 7~10)**
* **4) 불필요 라벨 제거 ('of', 'the')**
* **5) JSON 라벨을 학습용 멀티라벨 벡터로 변환**

### 3.2 데이터 정제

* **불필요한 공백, 특수문자 제거**
* **브랜드명 제거, 불용어 제거 (효과 제한적) → 최종적으로 사용 X**
* **희소 라벨 제거: RARE\_MIN\_COUNT = 7~10**
* **불필요 라벨 제거: 'of', 'the'**

### 3.3 라벨 분포 및 불균형 처리

**라벨 분포가 심하게 불균형하여 Macro-F1이 낮게 나타남. 이를 개선하기 위해 class\_weight='balanced' 적용 및 per-label threshold 최적화를 수행.희소 라벨은 제거하고 라벨 분포 상위 그룹을 중심으로 학습 데이터셋을 구성.**

**3.4 학습/평가 데이터 분할**

**전체 약 26,000건 데이터 중 80% 학습(train), 20% 검증(valid)으로 분할.**

**- Train: 21,001건**

**- Valid: 5,251건**

## 5. 향후 계획

* **전처리 파이프라인을 코드화하여 재현성 확보**
* **주기적 라벨 분포 모니터링**
* **신규 수집 데이터(description)와의 병합 및 증강**
* **전처리 데이터셋 기반으로 모델 학습/평가 단계 진행**