

**SK네트웍스 Family AI 과정 14기  
 데이터 전처리 인공지능 데이터 전처리 결과서**



| **산출물 단계** | 데이터 전처리 |
| --- | --- |
| **평가 산출물** | 인공지능 데이터 전처리 결과서 |
| **제출 일자** | 2025-09-25 |
| **깃허브 경로** | https://github.com/skn-ai14-250409/SKN14-Final-2Team |
| **작성 팀원** | 한성규 |

1. **문서 개요**

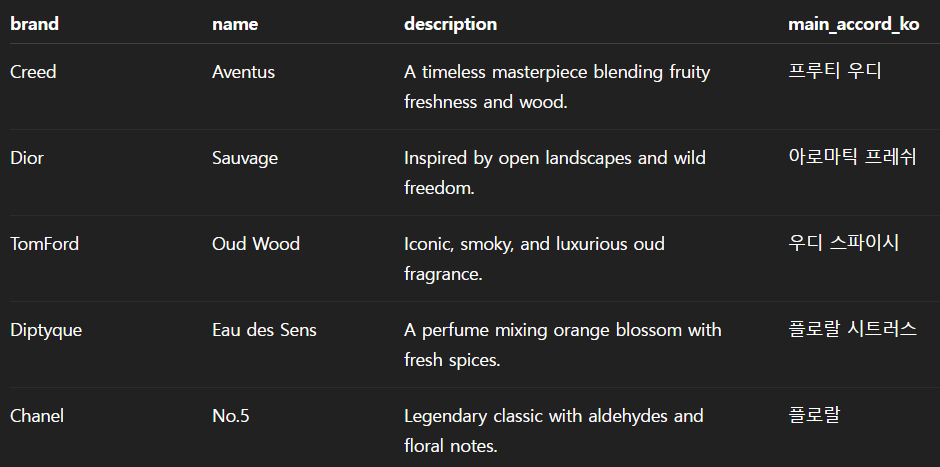
* 프로젝트명: ScentPick (대화형 향수 추천 챗봇)
* 전처리 목적: 향수 설명(description)과 라벨(main accord)을 기반으로 다중 라벨 분류 모델 학습을 위한 데이터 정제 및 불필요한 텍스트 제거
* 문제 정의:  
  + 원본 향수 설명 데이터에는 브랜드명, 광고성 표현, 불필요한 숫자, 일반적인 마케팅 단어들이 포함되어 있음
  + 이러한 노이즈 텍스트가 그대로 학습에 사용될 경우 모델 성능이 저하됨
  + 따라서 모델이 **향과 관련된 핵심 표현**만 학습하도록 데이터 전처리 필요

1. **데이터셋 개요**

* 데이터 출처 및 수집 방법: Fragrantica 크롤링 및 Huggingface 공개 데이터셋 활용

* 자체 CSV 병합 및 라벨 매핑 수행
* 데이터 구성:

| **항목명** | **설명** | **예시** |
| --- | --- | --- |
| brand | 향수 브랜드명 | Creed |
| name | 향수 이름 | Aventus |
| description | 영어 설명(원본) | “A timeless masterpiece with fresh citrus notes” |
| main\_accord\_ko | 주요 어코드 (라벨, 한국어) | 시트러스 우디 |

* 원본 데이터 샘플(5~10건 첨부):  
  

1. **전처리 프로세스 개요**

* 전체 흐름도:  
    
  ① 수집 → ② 결측치 처리 → ③ 불필요 단어 제거 → ④ 브랜드/향수명 제거 → ⑤ 숫자/특수문자 제거 → ⑥ 소문자 변환 → ⑦ 학습/검증 데이터 분리
* 전처리 파이프라인 요약:

| **단계** | **목적** | **수행 작업** | **사용 도구/라이브러리** | **적용** |
| --- | --- | --- | --- | --- |
| 결측치 처리 | description, label 공란 제거 | Null → 제거 | pandas | O |
| 불필요 단어 제거 | 광고성/마케팅 단어 제거 | “masterpiece”, “luxury”, “exclusive”, “perfume” 등 제거 | re | X |
| 브랜드 제거 | 브랜드명 제거 | Creed, Dior, TomFord 등 모든 브랜드명 삭제 | pandas | X |
| 향수명 제거 | 향수 개별 이름 제거 | Aventus, Sauvage 등 삭제 | pandas | X |
| 숫자 제거 | 숫자 제거 | 2021, 100ml → 제거 | re | X |
| 소문자 변환 | 일관성 확보 | 전체 문자열 소문자 변환 | pandas | X |
| 데이터 분리 | 학습/검증용 분리 | Multilabel Stratified Split (train 90% / valid 10%) | sklearn | O |

→ transformers 기반 모델로 주변 단어 삭제 시 문맥을 파악이 어려워져 성능이 떨어지는 문제 발생

1. **세부 전처리 단계**

**4.1 결측치 처리**

* 결측치 존재: 있음
* 처리 방법: description과 main\_accord\_ko 중 빈 값 → 제거

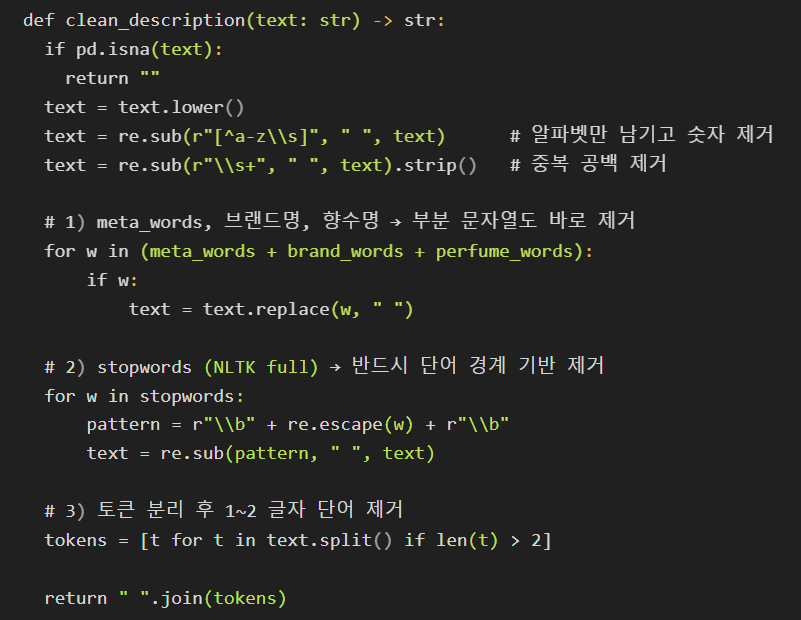
**4.2 불필요 단어/브랜드 제거**

* meta\_words 사전 기반 제거 (masterpiece, bestseller, iconic, luxury 등)
* 전체 브랜드명(약 1,000개 이상) 및 향수명 제거
* 숫자/특수문자 제거

**4.3 정규화**

* 영어 설명 → 소문자 변환
* 다중 공백 제거
* 길이가 1~2자인 불필요 단어 제거 (ex: “a”, “is”, “of” 등)

**코드 구현**



**4.5 데이터 분리**

* **방법:** MultilabelStratifiedShuffleSplit (train:test = 9:1)
* **결과:** train 23,442건 / valid 2,613건

**5. 전처리 결과 요약 및 평가**

* 원본: 26,319건
* 전처리 후: 26,055건
* **정리 결과:**
  + 불필요 단어,브랜드/향수명 제거 → 성능저하로 적용 X
  + 학습용 데이터 분포 균형 개선
* **최종 사용 데이터 : 전처리 과정을 거치지 않은 raw description**

→ 모델 무게와 처리 속도 이슈로 DL 대신 ML 모델을 적용했으며, 학습 결과 전처리하지 않은 description 사용 시 정확도가 더 높게 나타나 기존 전처리 데이터셋을 폐기하고 원본 데이터셋을 사용하기로 결정.

**6. 향후 활용 방안**

* description을 입력으로 다국어 임베딩 모델(XLM-RoBERTa) 학습
* main\_accord\_ko 다중 라벨 분류 태스크 수행
* ScentPick 챗봇의 사용자 자연어 입력 → 어코드 예측 및 향수 추천

1. **변경 이력**

| 변경일 | 변경자 | 변경 내용 | 비고 |
| --- | --- | --- | --- |
| 2025-08-28 | 한성규 | 브랜드/향수명/메타 단어 등 불필요 단어 제거 로직 추가 | main accord 추출을 위한 학습에 불필요한 Description의 단어 제거 |
| 2025-08-28 | 한성규 | stopwords 제거 로직 적용 & 변경 | Is, a 등 영어 불용어 제거, 로직 강화 |
| 2025-08-29 | 한성규 | 최종 전처리 데이터셋 확정 | 문맥 손실 방지 위해 단어 및 stopwords 제거 적용 X |
| 2025-09-04 | 박빛나 | 전처리 없이 description 그대로 사용 | 모델 무게와 처리 속도 이슈로 DL 대신 ML 모델을 적용했으며, 학습 결과 전처리하지 않은 description 사용 시 정확도가 더 높게 나타나 기존 전처리 데이터셋을 폐기하고 원본 데이터셋을 사용하기로 결정. |