인공지능 데이터 전처리 결과서

1. 데이터 개요

1.1. 전처리 목적

- 노트 - 탑/미들/베이스로 분류되지 않은경우 해당 노트들을 탑/미들/베이스 노트에 복제하여 채워넣음
- 전체 15029개의 데이터중 필수 요소인 노트정보, 이미지 정보, 어코드,계절, 성별, 상황정보가 포함되지 않은 행을 제거
- 리뷰 - ...등의

1.2. 문제 정의

구분	데이터명	데이터 수집 목적	데이터 형태	데이터 출처	데이터 크기	처리 상태	비고
DATA-001	perfume	향수명	str	parfumo	3893 건		
	brand	브랜드명	str	parfumo	3893 건		
	release_year	출시년도	str	parfumo	3630 건		
	concentration	향수 농도	str	parfumo	1128 건		
	perfumer	조향사	list	parfumo	2584 건		
	accord	accord 정보를 통해 향수 분위기 매칭	dict	parfumo	3893 건		
	audience	남/여 - 클래식/모던 향수 구분하여 추천에 이용	dict	parfumo	3893 건		
	season	어울리는 계절을 추천에 이용	dict	parfumo	3893 건		
	occasion	어울리는 상황을 매치하여 추천에 이용	dict	parfumo	3893 건		
	top_note	노트 정보를 이용하여 향에 대한 정보를 추천에 이용	list	parfumo , BaseNote	3893 건		
	middle_note	노트 정보를 이용하여 향에 대한 정보를 추천에 이용	list	parfumo, BaseNote	3893 건		
	base_note	노트 정보를 이용하여 향에 대한 정보를 추천에 이용	list	parfumo, BaseNote	3893 건		
	img	이미지 링크를 통해 향수 이미지 제시	str	parfumo	3893 건		
DATA-002	parfumo-reviews	parfumo사이트에서 크롤링한 리뷰 데이터들	str	parfumo	16731 건		
DATA-003	fragrantica-reviews	리뷰 정보를 바탕으로 사용자에게 상세한 추천을 제공	str	fragrantica	163860 건		

- 데이터 수집 기간:2025/12/31 ~ 2026/01/08
- 전체 수집 데이터 건수: 향수정보 3893건, 리뷰 180591건 수집

---

2. 데이터 수집 및 활용의 적법성 검토

구분	상업 이용 가능	학습 사용 허용	크롤링 이용약관 준수	개인정보 보호	비고
DATA-001	X	X	O	O	
DATA-002	X	X	O	O	
DATA-003	X	O	O	O	

- 저작권 준수: 해당 저작권 규정을 준수하기 위해 개인적이고 비상업적인 용도로만 허용됩니다
- 크롤링 이용약관 준수: Robots.txt 규정에 따라 서버에 부하를 주지 않는 방식으로 크롤링을 수행하였습니다.
- 개인정보 보호: 수집 과정에서 개인 식별 정보(id)는 제거하였습니다.

---

3. 데이터 수집 방법

[참고] 수집 방법이 동일한 데이터는 하나의 방식으로 통합하여 기재해주세요.

1. DATA-001

- 수집 방식 및 도구: 크롤링, selenium, DrissionPage,BeautifulSoup

2. DATA-002

- 수집 방식 및 도구: 크롤링, selenium, DrissionPage

3. DATA-003

- 수집 방식 및 도구: 크롤링, selenium

---

4. 데이터 저장 및 관리

1. DATA-001

- 저장 형식: tsv
- 저장 환경: Postgre DB, pgvector DB
- 데이터 구조: 해당 컬럼들을 그대로 단일 table로 구성하고 벡터 DB에는 "Brand의 Name은 Accord한 느낌을 가진 향수로 Season에 잘 어울리며
  Style한 향수입니다. Occasion한 상황에 잘 어울립니다."의 형태로 임베딩후 metadata에 컬럼의 내용을 dict형태로 저장
- 데이터 정제 및 전처리: 투표로 얻어지는 항목들에 대해서는 결측치의 경우 모두 제거하였고 top, middle, base의 구분이 명확하지 않은 경우 모든 note를 동일하게 채워 넣음

2. DATA-002

- 저장 형식: tsv
- 저장 환경: pgvector DB
- 데이터 구조: 해당 리뷰들을 그대로 vector 임베딩후 metadata에 향수 ID 저장

- 데이터 정제 및 전처리: ...으로 끝나는 등의 잘린 데이터와 중복데이터를 제거

3. DATA-003

- 저장 형식: tsv
- 저장 환경: pgvector DB
- 데이터 구조: 해당 리뷰들을 그대로 vector 임베딩후 metadata에 향수 ID 저장
- 데이터 정제 및 전처리: 배송이 빨라요 등, 향수와 관련 없는 내용은 제거

---

5. 데이터 전처리

5.1. 이상치 탐지 및 처리

1. 리뷰데이터 언어 필터링

- 이상치 기준: 리뷰 문장내에 알파벳 비율이 50% 이하이거나 Langdetect 라이브러리가 영어로 탐지한 확률이 80% 이하인 리뷰
- 처리 방법: 이상치로 판단후 제거
- 처리 결과: 16862건의 리뷰중 131건 제거 후 16731개의 리뷰 데이터 저장

5.2. 결측치 처리

1. 노트 정보 NULL

- 대상 결측 필드: 노트, 이미지, 어코드, 성별, 상황, 계절, 브랜드, 향수명
- 결측치 발생 건수(비율): 차트 데이터의 경우 부족할 경우 아예 수집에서 제외 노트 데이터는 전체 4000건중 1947건의 결측치 발생 (48.675 %)
- 처리 방법: 노트 정보가 부족한 경우 일단 Base Notes에서 향수 이름과 조향사, 브랜드를 기준으로 매핑후 노트 추가 크롤링 진행 이후 BaseNotes에도 없는 경우는 해당 행을 제거
- 처리 결과: BaseNote에 매칭된 건에 대해 추가 노트 수집 최종적으로 3893건의 향수에 대한 노트 정보 저장

2. 노트 정보 불균형

- 대상 결측 필드: 노트에 구분이 없는 경우 존재
- 결측치 발생 건수(비율): 노트 구분이 없는 경우는 1342건
- 처리 방법: 노트 구분이 없는 경우는 top middle base를 모두 해당 노트들로 복사하여 채워넣음
- 처리 결과: 각 3893건의 노트 정보 저장

5.2. 데이터 정제

[참고] 정규화/표준화를 위한 추가 전처리 내역 작성

(1) 필드명 표준화 적용

- 기준: 스네이크 케이스 규칙 적용
- 처리 과정:
- 적용 내용: 서로 다른 표기 형태의 필드명을 통합하여 관리
- 활용 방안: 표준화된 필드명을 RAG 인덱싱 및 파인튜닝 학습 데이터셋 구성에 활용하여, 데이터 처리 자동화와 검색 품질의 일관성을 확보

(2) 브랜드명 특수 케이스 처리

- 기준: 영문 알파벳 이외의 문자가 포함되거나 브랜드명이 바뀌어서 과거와 현재의 브랜드명이 교차되어 사용된 경우
- 처리 과정: 브랜드명을 수집하고 특수 케이스의 문자들을 확인후 알파벳으로 통합하기로 합의, 과거의 브랜드 명이나 Victoria's Secret같이 특수한 방식의 표기가 있는 경우 가장 대중적인 이름 하나로 통일하도록 합의
- 적용 내용: 영문 알파벳 이외의 문자는 알파벳으로 치환하여 적용하고 과거의 브랜드는 현재 사용중인 브랜드 명으로 치환 적용
- 활용 방안: 여러 사이트에서 크롤링한 브랜드 정보를 통합하여 사용 가능

(3) 향수명 특수 케이스 처리

- 기준: 향수명이 두 사이트에서 표기방식이나 특수 문자 처리등이 다른 경우 발생
- 처리 과정: 유사도 기반으로 처리하려고 시도했으나 전혀 다른 향수가 매칭되는 경우가 발생하여 최종적으로 향수에 대한 정보가 있는 Parfumo쪽의 데이터에 완전히 일치하는 향수만 사용하고 이름을 통일
- 적용 내용: Parfumo의 향수명으로 사용하도록 통일
- 활용 방안: 두 사이트의 데이터를 연결 가능

---

6. 전처리 프로세스 개요

- 전체 흐름도:

① 수집 → ② 결측치 처리 → ③ 이상치 탐지후 처리 → ④ 정규화

- 전처리 파이프라인 요약:

단계	목적	수행 작업	사용 도구/라이브러리
결측치 처리	누락값 제거	Null 행 제거, 특수값 대체	pandas
이상치 처리	비정상 데이터 제거	영어 리뷰만, 브랜드 매핑	pandas, langdetect
정규화	텍스트 전처리	소문자 변환	re