수집 데이터

1. 데이터 개요

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		

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	노트 정보의 경우 부족분을 BaseNote 사이트에서 추가 크롤링 진행
DATA-002	X	X	O	O	

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

3. 데이터 저장 및 관리

1. DATA-001

- 저장 형식: tsv,jsonl
- 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

▼ 데이터 구조

```
# tsv
perfume_id perfume brand release_year concentration perfumer accord audience season occasion top_note middle_note base_note tags img link
P_08272 Enigmatic Saffiano Hugo Boss 2024.0 {'Spicy': '3', 'Leathery': '4', 'Woody': '2', 'Fresh': '2', 'Creamy': '1', 'Floral': '1', 'Aquatic': '4'} {'Masculine': '3', 'Feminine': '1', 'Modern': '1'} {'Winter': '1', 'Fall': '2', 'Summer': '2', 'Spring': '2'} {'Leisure': '3', 'Night Out': '2', 'Business': '2', 'Evening': '2'} Marine notes, Morocco leather Marine notes, Morocco leather Marine notes, Morocco leather https://media.parfumo.com/perfume_social/c0/c042b8-enigmatic-saffiano-hugo-boss_1200.jpg?format=jpg&quality=90 https://www.parfumo.com/Perfumes/Hugo_Boss/enigmatic-saffiano
```


```
{
"perfume_id":"P_08272",
"perfume":"Enigmatic Saffiano",
"brand":"Hugo Boss",
"release_year":2024.0,
"concentration":null,
"perfumer":null,
"accord":"{'Spicy': '3', 'Leathery': '4', 'Woody': '2', 'Fresh': '2', 'Creamy': '1', 'Floral': '1', 'Aquatic': '4'}",
"audience":"{'Masculine': '3', 'Feminine': '1', 'Modern': '1'}","season":"{'Winter': '1', 'Fall': '2', 'Summer': '2', 'Spring': '2'}",
"occasion":"{'Leisure': '3', 'Night Out': '2', 'Business': '2', 'Evening': '2'}",
"top_note":"Marine notes, Morocco leather",
"middle_note":"Marine notes, Morocco leather",
"base_note":"Marine notes, Morocco leather",
"tags":null,
"img":"https:\/\/media.parfumo.com\/perfume_social\/c0\/c042b8-enigmatic-saffiano-hugo-boss_1200.jpg?format=jpg&quality=90",
"link":"https:\/\/www.parfumo.com\/Perfumes\/Hugo_Boss\/enigmatic-saffiano"
}
```


2. DATA-002

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