# 수집된 데이터 및 데이터 전처리 문서

## 1. 데이터 설명 및 구성

### 1.1. 데이터 및 필드 설명

| 구분 | 필드명 | 데이터 타입 | 설명 | 예시 |
|---|---|---|---|---|
| DATA-001 | perfume_id | int | 향수 ID (식별자) | 8092 |
| | perfume | string | 향수명 | Hypnotic Poision |
| | brand | string | 브랜드명 | Dior |
| | release_year | string | 출시년도 | 1998 |
| | concentration | string | 향수 농도 | Absolu de Parfum |
| | perfumer | list | 조향사 | [Jérôme Epinette, Robertet] |
| | accord | dict | accord 투표 정보 | {'Spicy': '4', 'Sweet': '1', 'Smoky': '3', 'Oriental': '1', 'Woody': '5', 'Resinous': '4', 'Fruity': '1', 'Floral': '2', 'Animal': '1'} |
| | audience | dict | style 투표 정보 | {'Masculine': '5', 'Feminine': '1', 'Classic': '1', 'Modern': '3'} |
| | season | dict | season 투표 정보 | {'Winter': '5', 'Fall': '5'} |
| | occasion | dict | occasion 투표 정보 | {'Leisure': '4', 'Night Out': '5', 'Evening': '5'} |
| | top_note | list | top note의 요소들 | [Saffron, Eucalyptus, Papyrus] |
| | middle_note | list | middle note의 요소들 | [Tuberose absolute, Oud, Jasmine absolute] |
| | base_note | list | base note의 요소들 | [Labdanum, Patchouli, Mineral amber, Sandalwood] |
| | img | string | 향수 이미지 링크 | https://media.parfumo.com/perfume_social/7f/7fa924_bois-obscur-byredo_1200.jpg?format=jpg&quality=90 |
| DATA-002 | review_id | int | 리뷰 ID (식별자) | 459 |
| | perfume_id | int | 향수 ID (왜래키) | 8092 |
| | content | string | 리뷰 내용 | Just my luck, try this on a whim and turns out I absolutely love it. Ugh! I'm a sucker for... |
| DATA-003 | perfume_id | int | 향수 ID (식별자) | 8281 |
| | perfume | string | 향수명 | Chelsea Flowers |
| | brand | string | 브랜드명 | Bond No. 9 |
| | fragrantica_link | string | 향수 링크 (fragrantica) | https://www.fragrantica.com/perfume/Bond-No-9/Chelsea-Flowers-2874.html |
| | review_count | int | 리뷰 갯수 | 32 |
| | description | string | 향수 설명 | Chelsea Flowers by Bond No 9 is a Floral fragrance for women. Chelsea Flowers was launched in 2003. The nose behind this fragrance is Laurent Le Guernec. Chelsea Flowers is a female perfume introduced in 2003 in a |

| 구분 | 필드명 | 데이터 타입 | 설명 | 예시 |
|---|---|---|---|---|
| | | | | translucent bottle, which was created by the designers of Bond No 9 house. The perfumer nose is Laurent Le Guernec. This perfume contains notes of peony, tulip, hyacinth, magnolia, rose, musk, sandalwood, vetiver and moss. It got its name from New York's Chelsea Flower Market, the large flower sale spot. The holiday 2008 edition was introduced in Swarovski All Stars collection in 50 ml bottle, covered with the finest tiny Swarovski crystals, emerald green colored, reflecting light the best. Apart from this one, the collection includes two more perfumes: Eau de New York and Chinatown. The price of this edition is 650 dollars. |
| | all_reviews | string | 향수에 대한 리뷰 | "Chelsea Flowers is a very generic verdant, floral fragrance with nothing special or distinctive. It reminded me of many, many perfumes I have smelled over the years, so unremarkable that I cannot even name one! I had hoped for something really special that reflected its name but all I got was a bouquet from the grocery store. |
| | collected_count | int | 수집된 리뷰 갯수 | 28 |
| | crawl_date | int | 수집된 날짜 | 2026-01-06 |

- 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(author)는 제거하였습니다.

---

## 3. 수집 자동화

- 수집 항목 및 품질 기준: 수집 데이터 필드와 정상 상태(기준) 작성
- 수집 방식 및 도구: selenium
- 데이터 유효성을 검증하는 방법: 검증 과정에 대한 설명(오류 발생 시 예외 처리 전략)

1. parfumo 요구 데이터 충족 여부 검사 (DATA-001):
    - 1차 : 가장 먼저 로딩되는 노트와 이미지 데이터의 여부를 검사
        - HTML에서 정적으로 로딩되는 노트 데이터의 경우 존재하지 않으면 사용하지 않기로 합의되었기 때문에 없으면 다음 링크로 이동
        - 예시: {"t": [], "m": [], "b": [], "n": []} 의 형태로 판단되는 경우 수집 안함

- 2차 : 어코드, 스타일, 계절, 상황데이터 여부 검사
    - 필수요소로 결정한 어코드, 스타일, 계절, 상황의 투표정보가 한 종류라도 부재하면 수집안함

2. fragrantica 요구 데이터 충족 여부 검사 (DATA-003):
    - 리뷰  90%이상 수집했는지 여부 검사
        - 제품별로 리뷰 개수를 저장해놓고 이를 목표치로 정하고 90%이상 수집한 경우에 다음 링크로 이동

- 수집 데이터 건수:  향수정보 3893건, 리뷰 180591건 수집
- 자동화 여부 및 주기: 자동화 예정

---

## 4. 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 | 어코드 등의 차트 데이터, 노트정보 결측행에 대해 제거 | pandas |
| 이상치 처리 | 비정상 데이터 제거 | 리뷰 데이터 중에 영문이외의 언어로 작성된 리뷰 제거 | re, langdetect |
| 정규화 | 텍스트 전처리 | 브랜드명의 표기 방법을 통일,<br>각 향수에 ID를 부여하고 리뷰와 ID 왜래키로 연결<br>D를 왜래키로 note, accord, audience, season, occasion의 테이블을 분리하고 원자화 진행,<br>노트 구분이 없는 경우 top, middle, base에 각각 복제하여 저장 | python, pandas |

- 도식화: 이미지 첨부