LLM 활용 소프트웨어

1. 개요

1. **목적** : 상품 추천시스템에서 활용될 상품 데이터의 데이터 전처리 과정에서 title과 description을 요약을 추가적으로 진행하여 데이터 전처리의 효율성을 높임
2. 상품 제목과 설명을 요약하고 추출하는데, gpt-4o-mini 모델을 활용하여 데이터 전처리 수행

2. 데이터 전처리

- **입력 데이터** : Amazon Reviews'23_All_Beauty (review, meta)
- **출처** : Amazon Reviews'23 https://amazon-reviews-2023.github.io/

요약 프로세스

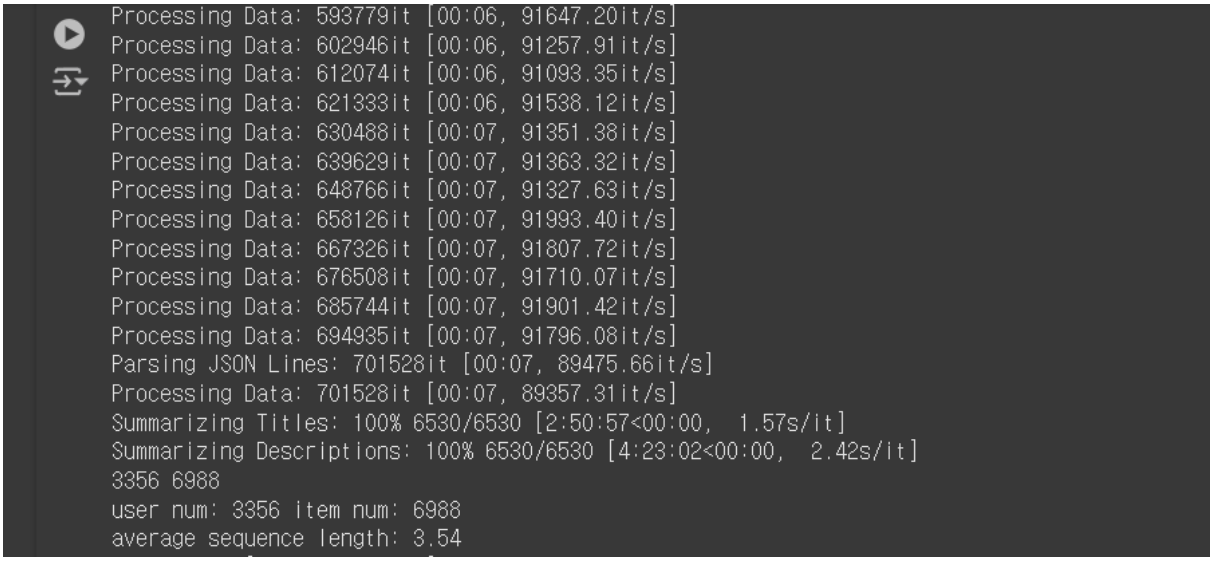
- **데이터 전처리** : 기존 Amazon Reviews'23 데이터 description 부분에 details 및 features 및 title 추가 하여 데이터를 만들어주고 전처리 코드에서 데이터를 불러와 파싱을 거치고 시간 필터링과 상호작용값을 카운트한 threshold 값으로 데이터를 전처리 후 title과 description을 매핑 및 요약을 진행함
- **GPT 활용 방식** : gpt-4o-mini 모델을 호출하여 상품 제목과 설명부분을 요약, 긴 제목에서 불필요한 형용사, 문자 제거 설명부분에서 불필요한 특수문자와 내용 요약
- **title 요약**(프롬프트 부분)

```python
def summarize_title(title, model="gpt-4o-mini"):
    try:
        chat_completion = client.chat.completions.create(
            model=model,
            messages=[
                {
                    "role": "system",
                    "content": (
                        "You are an extremely efficient title summarizer. "
                        "Your task is to summarize the provided title by identifying "
                        "and focusing on the important product name. Avoid unnecessary "
                        "details and focus on the key product name."
                    )
                },
                {
                    "role": "user",
                    "content": (
                        f"Here is the title: {title}. Summarize it by focusing on the "
                        "key product name and essential features, avoiding unnecessary details."
                    )
                }
            ],
            max_tokens=30,
            temperature=0.7,
        )
        time.sleep(1)
        summary = chat_completion.choices[0].message.content.strip()
        return summary

    except Exception as e:
        print(f"Error during GPT API call: {e}")
        return title
```

- **description 요약**(프롬프트 부분)

```python
def summarize_description(description, model="gpt-4o-mini"):
    try:
        chat_completion = client.chat.completions.create(
            model=model,
            messages=[
                {
                    "role": "system",
                    "content": (
                        "You are an efficient description summarizer. Your task is to "
                        "summarize the product description by focusing on the most important "
                        "features and aspects. Remove any unnecessary details and provide "
                        "a concise version of the description."
                    )
                },
                {
                    "role": "user",
                    "content": (
                        f"Here is the description: {description}. We condense the given "
                        "Amazon product metadata into a concise, relevant description that "
                        "focuses on the information that matters most to the customers and "
                        "eliminates unnecessary details. Given the title information and other "
                        "descriptions on the product, extract and summarize your product's key "
                        "features and benefits."
                    )
                }
            ],
            max_tokens=50,
            temperature=0.7,
        )
        time.sleep(1)
        summary = chat_completion.choices[0].message.content.strip()
        return summary

    except Exception as e:
        print(f"Error during GPT API call: {e}")
        return description
```

- 데이터 처리된 파일을 모델에 들어가기 위한 파일들
- 데이터 처리된 파일의 title과 description의 매칭을 확인하기 위한 파일들
- 데이터 처리 진행 중



- **제목 요약 처리**
  - **원본 제목**: NIRA Skincare Laser & Serum Bundle - Includes Anti-Aging Laser & Hyaluronic Acid Serum - Reduces Appearance of Fine Lines & Wrinkles - FDA Cleared
  - **요약된 제목**: NIRA Skincare Laser & Serum Bundle

  - **원본 제목**: Philips Sonicare Essence+ Gum Health & Airfloss Rechargeable Electric Flosser, Bundle Value Pack, HX8218/02
  - **요약된 제목**: Philips Sonicare Essence+ Electric Flosser Bundle

  - **원본 제목**: APIVITA Queen Bee Holistic Age Defense Night Cream 1.69 fl.oz. | Intensive Night Treatment That Speeds Skin Regeneration, Smooths Wrinkle & Increases Skin Elasticity with Royal Jelly & Hyaluronic Acid
  - **요약된 제목**: APIVITA Queen Bee Night Cream
- **설명 요약 처리**
  - **원본 설명**:"NIRA Skincare Laser & Serum Bundle - Includes Anti-Aging Laser & Hyaluronic Acid Serum - Reduces Appearance of Fine Lines & Wrinkles - FDA Cleared. POWERFUL ANTI-AGING DUO - This powerful anti-aging duo includes our Skincare Laser & Hyaluronic Acid Serum Bundle. Reduce fine lines & wrinkles, stimulate natural collagen production & hydrate skin for glowing, radiant results. AT-HOME LASER

TREATMENT - The NIRA Anti-Aging Laser is the first & only painless at-home laser for wrinkle reduction. Our non-fractional, patented technology is FDA cleared to reduce fine lines for visible results in just 2 minutes a day. HYALURONIC ACID SERUM - Our Hyaluronic Acid Serum contains four all-natural ingredients to lock in moisture, diminish the appearance of fine lines & wrinkles, and drive antioxidant activity, resulting in skin that appears softer, smoother & firmer. SAME GREAT RESULTS - The treatment is simple, fast & effective—and can be used in the comfort of your own home. NIRA's technology is the same used by professional dermatologists— but at a fraction of the price & without irritation. HOW TO USE - Wash & dry face completely. Use NIRA Anti-Aging Skincare Laser for 2 minutes to stimulate collagen & reduce wrinkles. Finish with NIRA Hyaluronic Acid Serum for dramatic, faster & longer-lasting results."

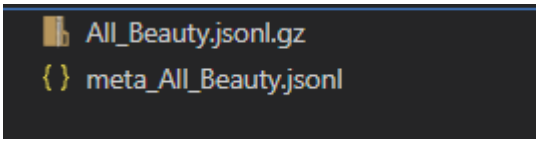
- **요약된 설명**:"**NIRA Skincare Laser & Serum Bundle**: This bundle features a powerful anti-aging duo, including an FDA-cleared at-home laser and a hydrating Hyaluronic Acid serum. The painless NIRA Anti-Aging Laser effectively reduces fine lines and stimulates collagen. The Hyaluronic Acid Serum locks in moisture and enhances skin firmness, offering a spa-like treatment at home with results similar to those from professional dermatologists."

- **원본 설명**:"Caroline Keller Keratin Shampoo for dry and damaged hair and scalp. With Argan oil, Avocado oil, Keratin, and Vitamins. Specially formulated for Post Keratin Treatments. Salt Free. 16.9 fl.Oz. This shampoo is designed to revitalize dry, over-processed hair, and improve hair strength, elasticity, and hydration. Argan oil and Avocado oil deeply nourish the hair, while Keratin repairs damaged hair structures, making it softer, smoother, and more manageable. Suitable for all types of chemically-treated or color-treated hair."
- **요약된 설명**:"**Caroline Keller Keratin Shampoo**: Specially formulated for dry, damaged, and chemically-treated hair, making it ideal for post-keratin treatment care. Infused with Argan oil, Avocado oil, and vitamins, this salt-free shampoo nourishes, strengthens, and hydrates, leaving hair soft and manageable."

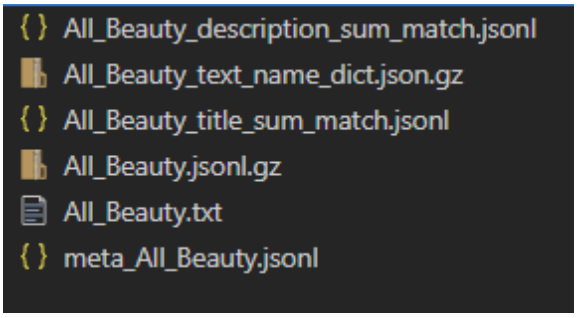
- **원본 설명**:"Philips Sonicare Essence+ Gum Health & Airfloss Rechargeable Electric Flosser, Bundle Value Pack, HX8218/02. Improve your oral health with the new Philips Sonicare Essence+ Gum Health & Airfloss Rechargeable Electric Flosser Value Bundle. This value bundle features a rechargeable toothbrush with unique and effective Sonicare technology. The Essence+ Gum Health toothbrush reduces the amount of plaque at the gum line versus a manual toothbrush, and is designed to help prevent gingivitis and gently lifts away stains for naturally whiter teeth. Included in your value pack is the Philips Sonicare Airfloss Rechargeable Electric Flosser, which offers an easy, fast routine to help prevent cavities and improve long term gum health in between teeth. It uses advanced technology to gently remove plaque between teeth with a rapid burst of air and micro-droplets in just 30 seconds. It is most ideal for inconsistent flossers."
- **요약된 설명**:"**Philips Sonicare Essence+ Electric Flosser Bundle**: Enhance your oral health with this comprehensive value bundle, which includes a rechargeable toothbrush and Airfloss electric flosser. The Sonicare Essence+ toothbrush reduces plaque at the gum line and helps prevent gingivitis, while the Airfloss offers quick and efficient plaque removal between teeth in 30 seconds."

- **데이터 전처리 파일**

전



후



3. 기대 효과

본 전처리 작업은 title과 description을 간결하게 요약함으로써, 데이터의 일관성과 용량 최적화를 통해 추천 시스템의 정확도를 개선함.

- **데이터 일관성 향상**: 긴 제목을 간결하게 정리함으로써, 모델 입력 데이터의 일관성 강화, 이를 통해, 모델이 다양한 데이터에서 일관된 패턴을 학습할 수 있게 되어 추천의 신뢰성이 높아짐.
- **모델 성능 최적화**: 데이터 요약을 통해 모델이 불필요한 정보가 아닌 핵심 정보에 집중할 수 있도록 지원함. 이는 모델이 보다 정확한 추천 결과를 산출하도록 돕고, 추천 시스템의 정확도 향상에 기여.
- **데이터 용량 감소**: 긴 제목과 설명에서 불필요한 부분을 제거함으로써 데이터 용량이 줄어들어 저장 및 처리 효율이 크게 향상. 이는 전반적인 데이터 처리 시간 단축과 더불어, 기존 로컬 환경에서 모델이 너무 커져 돌아가지 않던 상황에서 모델을 사용할 수 있게 해주며 모델이 더 빠르고 효율적으로 학습할 수 있도록 함.

4. 도구 및 환경

- **GPT 모델**: OpenAI의 GPT-4o-mini 모델을 사용하여 title과 description 요약 작업을 수행, 이 모델은 긴 텍스트에서 중요한 정보를 추출하고, 불필요한 부분을 제거하여 간결한 형태로 요약 최적화
- **API 연동 방식**: OpenAI API 호출을 통해 데이터 전처리 과정에서 요약 작업을 자동화, API를 활용한 요약 결과는 원본 데이터와 매핑되어 데이터셋에 반영, 이를 통해 전처리 단계에서 일관된 데이터 요약 가능
- **데이터셋**: Amazon Reviews'23 데이터의 All_Beauty 섹션에서 review와 meta 정보를 사용하여, 상품의 title, description, details, features 등 종합적 활용, 해당 데이터를 기준으로 요약된 텍스트를 구성하여 추천 모델에 최적화된 데이터 구축
- **전처리 환경**:
    - 데이터 전처리 과정은 Python과 관련 라이브러리(json등)를 수행하며, 데이터 로딩, 파싱, 시간 필터링, 상호작용값 기반의 threshold 처리 진행

결론

이번 전처리 과정은 llm을 활용하여 title과 description 요약을 통해 데이터의 일관성과 효율성을 높이고, 추천 모델의 성능을 개선하는데 초점을 둠

 요약된 데이터는 기존 대비 용량이 줄어들어 저장 및 처리 효율성을 크게 향상시켰고, 추천 시스템이 핵심 정보에 집중하도록 함으로써 추천 정확도를 높임

- **주요 성과**:
    - 데이터의 일관성 강화: 간결한 제목과 설명을 제공으로, 모델에 일관된 입력 데이터를 공급

    - 모델 효율성 향상: 요약된 데이터는 모델의 학습 및 예측 단계에서 처리 부담을 줄이고, 더 신속한 반응을 가능하게 하여 사용자 경험을 개선하는 데 긍정적인 영향
    - 데이터 용량 최적화: 요약 과정을 통해 불필요한 정보가 제거, 데이터 크기 감소로, 전처리와 저장의 효율성을 높임
- **향후 계획**:
    - 요약 성능을 지속적 모니터링 및 데이터 품질을 유지함으로써 개선할 수 있는 방안을 검토
    - 추천 모델의 성능을 주기적으로 평가하여, 요약된 데이터가 추천 결과에 대한 영향 분석