**모델 정의서**

사랑의 홈쇼핑 창성핑 팀

이유송(팀장), 최창성, 송우선, 장은별, 김수현, 정수빈

**<목차>**

1. **개요**
   1. 문서 정보
   2. 프로젝트 개요
2. **데이터**

2.1 데이터 출처

2.2 데이터 테이블

2.3 데이터 전처리

1. **FastText**
   1. 모델 설명
   2. 하이퍼 파라미터
   3. 입출력 값
2. **KMeans**
   1. 모델 설명
   2. 하이퍼 파라미터
   3. 입출력 값
3. **CDAE**
   1. 모델 설명
   2. 하이퍼 파라미터
   3. 입출력 값
4. **개요**
   1. **문서 정보**

* 팀명: 사랑의 홈쇼핑 창성핑
* 팀원: 이유송(팀장), 장은별, 최창성, 정수빈, 송우선, 김수현
* 프로젝트 명: 트렌드 기반 홈쇼핑 상품 추천 서비스
* 작성일: 2025-01-21
  1. **프로젝트 개요**

본 프로젝트는 **네이버 트렌드** 및 **사용자 행동 데이터**를 기반으로 개인화된 상품 추천 서비스를 구현하여 사용자 경험을 개선하고 전자상거래 플랫폼의 매출을 증대하는 것을 목표로 합니다.

사용자의 클릭, 검색, 즐겨찾기 등 다양한 행동 데이터를 수집 및 분석하고, 트렌드에 민감한 젊은 세대를 타겟으로 하여 트렌드 워드를 반영하여 개별 사용자에게 최적화된 상품을 추천합니다. 이를 통해 구매 전환율을 높이고, 고객 만족도를 증진하며, 경쟁력을 강화합니다.

**프로젝트 목표**:

1. **정확도 높은 추천 제공**: 트렌드와 사용자 데이터를 결합하여 개별 사용자에게 최적화된 트렌드 상품 추천.
2. **구매 전환율 증대**: 사용자 행동 데이터를 활용한 효과적인 추천으로 구매 가능성 극대화.
3. **고객 만족도 향상**: 개인화된 추천을 통해 사용자 경험 개선.

이 프로젝트는 데이터 분석과 추천 알고리즘을 접목하여 전자상거래 플랫폼에서 사용자의 관심과 참여를 높이는 데 중점을 둡니다.

1. **데이터**
   1. **데이터 출처**

본 프로젝트에서는 총 다섯개의 데이터셋을 사용하였으며, 이는 다음과 같습니다.

* **홈쇼핑 데이터**: GS홈쇼핑, 공영 홈쇼핑, 신세계 홈쇼핑 외 총 17개의 홈쇼핑 사이트 편성표에 나와있는 제품을 크롤링 하여 약 18000천 여개의 데이터를 수집.
* **유저 데이터**: 사랑의 홈쇼핑 창성핑 사이트에 가입한 회원의 개인 정보, 회원 정보의 데이터. (가상 데이터 사용)
* **로그 데이터**: 사랑의 홈쇼핑 창성핑 사이트 회원의 로그 기록. 클릭, 검색 등의 이벤트를 실시간. (가상 데이터 사용)
* **트렌드 데이터**: 네이버 실시간 검색어 순위를 크롤링. 총 10개의 카테고리에 나와있는 10개의 인기 검색어를 일간, 주간, 월간으로 크롤링.
* **위키피디아 데이터**: 위키피디아에서 제공하는 한국어 데이터셋 API를 사용.
  1. **데이터 테이블**
* 홈쇼핑 데이터

컬럼명: \_id, title, datetime(방송 날짜), broadcasttime(방송 시작 시간), price, category, keyword1, keyword2, keyword3, url, image\_url

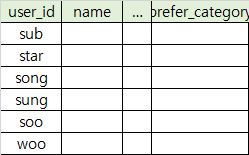
총 11개로 구성



* 유저 데이터

유저의 나이, 성별, 이름, 선호 카테고리 총 4가지의 데이터

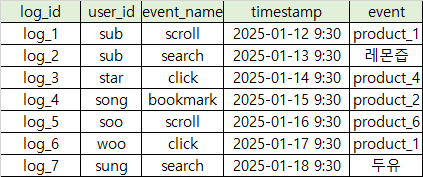
선호 카테고리는 총 10개의 카테고리 중 1 개 이상을 선택.



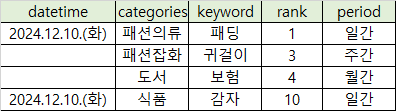
* 로그 데이터

이벤트의 종류는 검색, 클릭, 즐겨찾기 총 3가지만 추출하여 구성

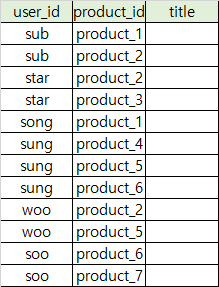
event 컬럼은 이벤트가 발생한 키워드나 상품 아이디를 나타냄



* 트렌드 데이터



* 즐겨찾기 데이터



* 1. **데이터 전처리**
* 홈쇼핑 데이터
  1. 추출 및 카테고리 선정: **Open AI**를 이용하여 해당 방송의 키워드를 추출, 1네이버 트렌드 카테고리에 맞게 총 10개의 카테고리 중 하나로 분류.
  2. 불용어 제거: 방송 제목 데이터가 들어간 **title** 컬럼에서 불용어를 제거. 영어나, 특수문자, 숫자뿐만 아니라 ‘특가’, ‘할인’, ‘묶음’ 등 임베딩을 진행할 때 혼동을 줄 수 있는 단어 제거.
  3. 가격 전처리: 보험 상품 혹은 렌탈 상품의 경우 ‘가격 없음’ 혹은 ‘상담사와 상담’ 등으로 가격이 되어있는 경우가 있다. 이런 경우 가격을 **nan** 값으로 통일하여 전처리.
* 로그 데이터

로그 데이터의 이벤트 활동에서 ‘클릭’, ‘검색’, ‘즐겨찾기’의 선호도를 나타낼 수 있는 값만 필터링.

* 임베딩 데이터

줄의 앞뒤 불필요한 공백 제거, 내용이 없는 빈 줄 제거 등의 간단한 전처리.

1. **FastText**
   1. **모델 설명**

이 모델은 한국어 Wikipedia 데이터를 기반으로 훈련된 FastText 임베딩 모델. FastText는 단어 수준 벡터를 학습하는 Word2Vec과 달리, 서브워드(subword) 정보를 포함하여 단어 벡터를 학습합니다. 이를 통해 희귀 단어 또는 \*\*미등록어(OOV)\*\*에 대해 더욱 강력한 일반화 성능을 발휘할 수 있다.

홈쇼핑 상품명이나 설명에는 자주 등장하지 않는 희귀 단어 또는 신조어, 외래어, 브랜드명이 포함될 가능성이 높습니다.

* **FastText는 서브워드(부분 문자열)를 활용해 OOV(Out-Of-Vocabulary, 미등록어) 단어도 벡터화할 수 있습니다.**
  + 예: 브랜드명 "아디다스오리지널"처럼 복합적이거나 새롭게 만들어진 단어도 "아디다스"와 "오리지널" 서브워드를 통해 의미를 학습 가능.
  + 한글의 음절 구조와 조합 특성상, 서브워드 기반 학습은 한국어 데이터에 특히 유용합니다.
* **한국어 Wikipedia 데이터는 광범위한 일반 지식을 포함하고 있어, 다양한 도메인에서 기본적인 어휘 관계를 잘 학습하고 있습니다.**
  + 홈쇼핑 상품 설명에도 일반적인 한국어 표현과 어휘가 자주 사용되기 때문에, 도메인에 맞는 추가 학습 없이도 높은 성능을 기대할 수 있습니다.
* **FastText는 학습 및 추론 속도가 빠르며, 실시간 애플리케이션에도 적합합니다.**
  + 홈쇼핑 상품은 지속적으로 업데이트되며, 새로운 상품에 대해 빠르게 임베딩을 생성해야 할 수도 있으므로 이점이 큽니다.
  1. **하이퍼 파라미터**

**FastText 모델 설정**:

vector\_size=100: 임베딩 차원 수를 100으로 설정.

window=5: 학습 시 고려할 컨텍스트 윈도우 크기.

min\_count=5: 최소 등장 빈도가 5 이상인 단어만 학습.

workers=4: 4개의 워커 스레드를 사용하여 병렬 학습.

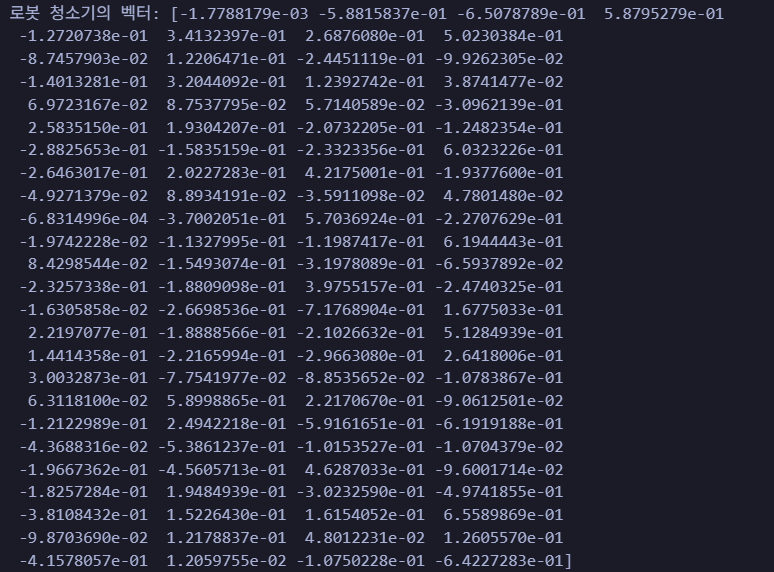
sg=1: Skip-gram 방식을 사용하여 단어 벡터 학습

.

* 1. **입출력 값**

입력값: 한글 문장 혹은 단어

출력값: 4차원 임베딩 값



1. **KMeans**
   1. **모델 설명**

**모델 이름:** K-Means

**알고리즘 유형:** 클러스터링 기반의 추천 시스템

**모델 선택 이유**:

* 데이터의 규모가 커질수록 KMeans는 상대적으로 빠른 학습과 예측을 제공함.
* 상품의 카테고리나 특성에 따라 유연하게 클러스터 수를 조정할 수 있어, 다양한 상품군에 맞는 추천 가능
* 클러스터 내의 상품을 추천할 때, 사용자가 관심있을 만한 다른 상품을 추천함으로써 사용자에게 의미있는 추천이 가능

**모델 파이프라인**

1. **데이터 전처리**
   * 홈쇼핑 데이터 전처리: 결측값 및 특수문자 전처리
2. **임베딩 생성**
   * 카테고리 및 키워드 컬럼을 결합하여 하나의 컬럼으로 생성
   * 생성한 컬럼을 기반으로 텍스트 벡터 생성
   * FastText를 통해 단어를 벡터로 변환 후 평균화
3. **KMeans 클러스터링**
   * 실루엣 계수를 구해 최적의 k값을 선정 후 클러스터링 진행
4. **상품 추천**
   * 사용자가 입력한 키워드를 기반으로 클러스터 내에서 관련성이 높은 상품을 출력
   * 각 상품의 임베딩을 벡터로 비교하여 유사도를 계산하고, 가장 유사한 상품을 추천
   1. **하이퍼 파라미터**

| 파라미터 | 값 | 설명 |
| --- | --- | --- |
| n\_cluster | 8 | 생성할 클러스터 개수 |
| n\_init | 10 | 반복할 클러스터링 수 |
| max\_iter | 300 | 클러스터링 최대 반복 횟수 |

* 1. **입출력 값**

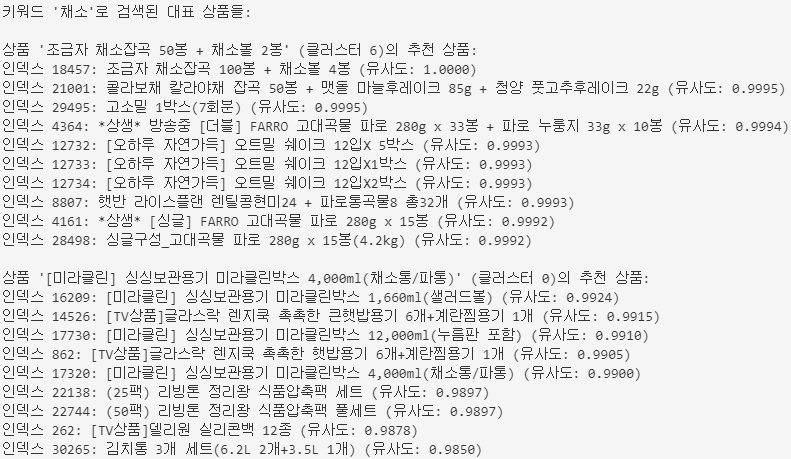
**입력 데이터**

* **상품 데이터:** 상품 이름, 가격, 카테고리, 임베딩 값

**출력 데이터**

* 사용자가 입력한 키워드와 상품 데이터의 유사도를 계산하여 유사도가 높은 10개의 상품 추천

**출력 형식**



1. **CDAE**
   1. **모델 설명**

**모델 이름**: Contractive Denoising Autoencoder (CDAE).

**알고리즘 유형**: 딥러닝 기반의 추천 시스템.

**모델 선택 이유**:

사용자 행동 데이터와 상품 데이터를 결합하여 개인화된 추천을 제공하기 적합.

Gaussian Noise와 Dropout을 사용해 일반화를 강화하고 과적합 방지.

Denoising이라는 이름처럼, 원래 데이터를 약간 손상(잡음 추가)시키고 복원하는 과정을 통해 모델이 더 강인해지도록 하는 특징을 가짐.

**모델 아키텍쳐**

* **입력 레이어**:

상품 데이터 벡터 (item\_input)와 사용자 ID (user\_input)를 입력.

* **Encoder**:

상품 데이터를 저차원 벡터로 변환.

Gaussian Noise와 Dropout으로 일반화.

* **사용자 임베딩**:

사용자 ID를 벡터로 변환해 상품 데이터와 결합.

* **Decoder**:

결합된 정보를 디코딩하여 각 상품에 대한 추천 점수를 예측.

#### 모델 파이프라인

1. **데이터 전처리**:
   * 행동 로그 정리 및 결측값 처리.
   * 상품 및 사용자 데이터 정규화.
2. **가중치 계산**:
   * 이벤트(클릭, 검색, 즐겨찾기), 시간, 카테고리, 트렌드 가중치를 계산.
3. **모델 학습**:
   * 사용자 ID와 상품 가중치를 입력으로 학습.
4. **추천 생성**:
   * 상위 10개의 추천 상품과 점수 반환.

#### 데이터 전처리

* 상품 데이터는 fasttext 모델을 이용하여 임베딩
* 해당 사용자 id와 일치하는 로그 데이터를 필터링. 이벤트 내용 중 상품에 대한 선호도를 알 수 있는 ‘검색’, ‘클릭’, ‘즐겨찾기’의 이벤트만 추출.
* 이벤트 내용의 상품 id를 통해 상품 데이터 베이스와 병합.
* 각각의 상품에 가중치를 부여
  + 사용자의 행동 패턴과 빈도를 통해 [즐겨찾기> 검색> 클릭] 순으로 해당 상품과 비슷한 상품에 대하여 가중치를 다르게 부여
  + 트렌드 데이터의 키워드를 통해 해당 키워드를 포함한 상품에 가중치를 부여
  + 시시각각 변화하는 사용자의 관심도를 반영하기 위해 로그 데이터를 최신순으로 20대 추출 후 시간의 순서에 따라 가중치를 다르게 부여

→ 최신의 로그일수록 가중치가 높아지게 설정

* 1. **하이퍼 파라미터**

| 파라미터 | 값 | 설명 |
| --- | --- | --- |
| encoding\_dim | 32 | 임베딩 벡터의 차원 크기 |
| dropout\_rate | 0.3 | Dropout 확률 |
| noise\_factor | 0.1 | Gaussian Noise 추가 비율 |
| epochs | 100 | 학습 반복 횟수 |
| batch\_size | 12 | 학습 시 배치 크기 |

* 1. **입출력 값**

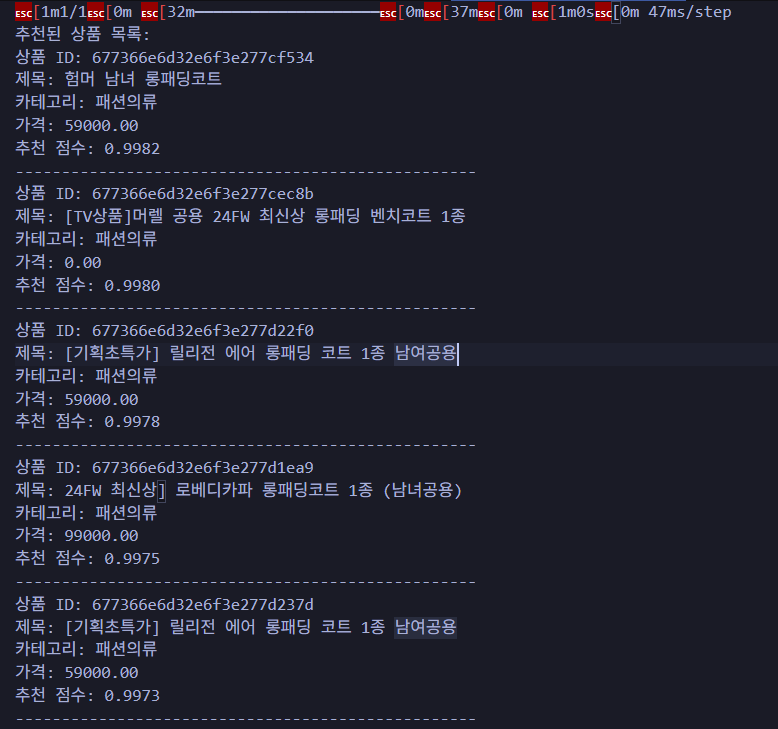
**사용자 입력**

* 사용자 데이터 : 사용자 ID, 선호 카테고리, 나이, 성별
* 로그 데이터 : 사용자 ID, 이벤트, 이벤트 내용, 이벤트 발생 시간
* 트렌드 데이터 : 네이버 트랜드 키워드, 일자, 카테고리
* 상품 데이터: 상품 이름, 가격, 카테고리, 임베딩 값

**모델 예측**

* 가중치와 유사도를 계산하여 추천 점수를 계산하여 상위 10개의 상품을 사용자에게 추천.

**출력 형식**: [예: 추천 항목 리스트]



가장 점수가 높은 상품들부터 차례로 10개가 출력되는 것을 알 수 있다.