모델 정의서

사랑의 홈쇼핑 창성핑 팀

이유송(팀장), 최창성, 송우선, 장은별, 김수현, 정수빈

<목차>

1. 개요

1.1 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

1.2 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.

---

2. 데이터

2.1 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.

2.2 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

product_id	title	price	category	...	url
product_1	title_1				
product_2	title_2				
product_3	title_3				
product_4	title_4				
product_5	title_5				
product_6	title_6				
product_7	title_7				

- 유저 데이터

  유저의 나이, 성별, 이름, 선호 카테고리 총 4가지의 데이터

  선호 카테고리는 총 10개의 카테고리 중 1 개 이상을 선택.

user_id	name	...	prefer_category
sub			
star			
song			
sung			
soo			
woo			

- 로그 데이터

  이벤트의 종류는 검색, 클릭, 즐겨찾기 총 3가지만 추출하여 구성

  event 컬럼은 이벤트가 발생한 키워드나 상품 아이디를 나타냄

log_id	user_id	event_name	timestamp	event
log_1	sub	scroll	2025-01-12 9:30	product_1
log_2	sub	search	2025-01-13 9:30	레몬즙
log_3	star	click	2025-01-14 9:30	product_4
log_4	song	bookmark	2025-01-15 9:30	product_2
log_5	soo	scroll	2025-01-16 9:30	product_6
log_6	woo	click	2025-01-17 9:30	product_1
log_7	sung	search	2025-01-18 9:30	두유

- 트렌드 데이터

datetime	categories	keyword	rank	period
2024.12.10.(화)	패션의류	패딩	1	일간
	패션잡화	귀걸이	3	주간
	도서	보험	4	월간
2024.12.10.(화)	식품	감자	10	일간

- 즐겨찾기 데이터

user_id	product_id	title
sub	product_1	
sub	product_2	
star	product_2	
star	product_3	
song	product_1	
sung	product_4	
sung	product_5	
sung	product_6	
woo	product_2	
woo	product_5	
soo	product_6	
soo	product_7	

2.3 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.

---

3. FastText

3.1 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.

3.2 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

.

3.3 입출력 값

입력값: 한글 문장 혹은 단어

출력값: 4차원 임베딩 값

```
로봇 청소기의 벡터: [-1.7788179e-03 -5.8815837e-01 -6.5078789e-01  5.8795279e-01
 -1.2720738e-01  3.4132397e-01  2.6876080e-01  5.0230384e-01
 -8.7457903e-02  1.2206471e-01 -2.4451119e-01 -9.9262305e-02
 -1.4013281e-01  3.2044092e-01  1.2392742e-01  3.8741477e-02
  6.9723167e-02  8.7537795e-02  5.7140589e-02 -3.0962139e-01
  2.5835150e-01  1.9304207e-01 -2.0732205e-01 -1.2482354e-01
 -2.8825653e-01 -1.5835159e-01 -2.3323356e-01  6.0323226e-01
 -2.6463017e-01  2.0227283e-01  4.2175001e-01 -1.9377600e-01
 -4.9271379e-02  8.8934191e-02 -3.5911098e-02  4.7801480e-02
 -6.8314996e-04 -3.7002051e-01  5.7036924e-01 -2.2707629e-01
 -1.9742228e-02 -1.1327995e-01 -1.1987417e-01  6.1944443e-01
  8.4298544e-02 -1.5493074e-01 -3.1978089e-01 -6.5937892e-02
 -2.3257338e-01 -1.8809098e-01  3.9755157e-01 -2.4740325e-01
 -1.6305858e-02 -2.6698536e-01 -7.1768904e-01  1.6775033e-01
  2.2197077e-01 -1.8888566e-01 -2.1026632e-01  5.1284939e-01
  1.4414358e-01 -2.2165994e-01 -2.9663080e-01  2.6418006e-01
  3.0032873e-01 -7.7541977e-02 -8.8535652e-02 -1.0783867e-01
  6.3118100e-02  5.8998865e-01  2.2170670e-01 -9.0612501e-02
 -1.2122989e-01  2.4942218e-01 -5.9161651e-01 -6.1919188e-01
 -4.3688316e-02 -5.3861237e-01 -1.0153527e-01 -1.0704379e-02
 -1.9667362e-01 -4.5605713e-01  4.6287033e-01 -9.6001714e-02
 -1.8257284e-01  1.9484939e-01 -3.0232590e-01 -4.9741855e-01
 -3.8108432e-01  1.5226430e-01  1.6154052e-01  6.5589869e-01
 -9.8703690e-02  1.2178837e-01  4.8012231e-02  1.2605570e-01
 -4.1578057e-01  1.2059755e-02 -1.0750228e-01 -6.4227283e-01]
```

4. KMeans

4.1 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

4.2 하이퍼 파라미터

파라미터	값	설명
n_cluster	8	생성할 클러스터 개수
n_init	10	반복할 클러스터링 수
max_iter	300	클러스터링 최대 반복 횟수

4.3 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

```
키워드 '채소'로 검색된 대표 상품들:

상품 '조금자 채소잡곡 50봉 + 채소볼 2봉' (클러스터 6)의 추천 상품:
인덱스 18457: 조금자 채소잡곡 100봉 + 채소볼 4봉 (유사도: 1.0000)
인덱스 21001: 콜라보채 칼라야채 잡곡 50봉 + 맷돌 마늘후레이크 85g + 청양 풋고추후레이크 22g (유사도: 0.9995)
인덱스 29495: 고소밀 1박스(7회분) (유사도: 0.9995)
인덱스 4364: *상생* 방송중 [더블] FARRO 고대곡물 파로 280g x 33봉 + 파로 누룽지 33g x 10봉 (유사도: 0.9994)
인덱스 12732: [오하루 자연가득] 오트밀 쉐이크 12입X 5박스 (유사도: 0.9993)
인덱스 12733: [오하루 자연가득] 오트밀 쉐이크 12입X1박스 (유사도: 0.9993)
인덱스 12734: [오하루 자연가득] 오트밀 쉐이크 12입X2박스 (유사도: 0.9993)
인덱스 8807: 햇반 라이스플랜 렌틸콩현미24 + 파로통곡물8 총32개 (유사도: 0.9993)
인덱스 4161: *상생* [싱글] FARRO 고대곡물 파로 280g x 15봉 (유사도: 0.9992)
인덱스 28498: 싱글구성_고대곡물 파로 280g x 15봉(4.2kg) (유사도: 0.9992)

상품 '[미라클린] 싱싱보관용기 미라클린박스 4,000ml(채소통/파통)' (클러스터 0)의 추천 상품:
인덱스 16209: [미라클린] 싱싱보관용기 미라클린박스 1,660ml(샐러드볼) (유사도: 0.9924)
인덱스 14526: [TV상품]글라스락 렌지쿡 촉촉한 큰햇밥용기 6개+계란찜용기 1개 (유사도: 0.9915)
인덱스 17730: [미라클린] 싱싱보관용기 미라클린박스 12,000ml(누름판 포함) (유사도: 0.9910)
인덱스 862: [TV상품]글라스락 렌지쿡 촉촉한 햇밥용기 6개+계란찜용기 1개 (유사도: 0.9905)
인덱스 17320: [미라클린] 싱싱보관용기 미라클린박스 4,000ml(채소통/파통) (유사도: 0.9900)
인덱스 22138: (25팩) 리빙톤 정리왕 식품압축팩 세트 (유사도: 0.9897)
인덱스 22744: (50팩) 리빙톤 정리왕 식품압축팩 풀세트 (유사도: 0.9897)
인덱스 262: [TV상품]델리원 실리콘백 12종 (유사도: 0.9878)
인덱스 30265: 김치통 3개 세트(6.2L 2개+3.5L 1개) (유사도: 0.9850)
```

---

5. CDAE

5.1 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

5.2 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

5.3 입출력 값

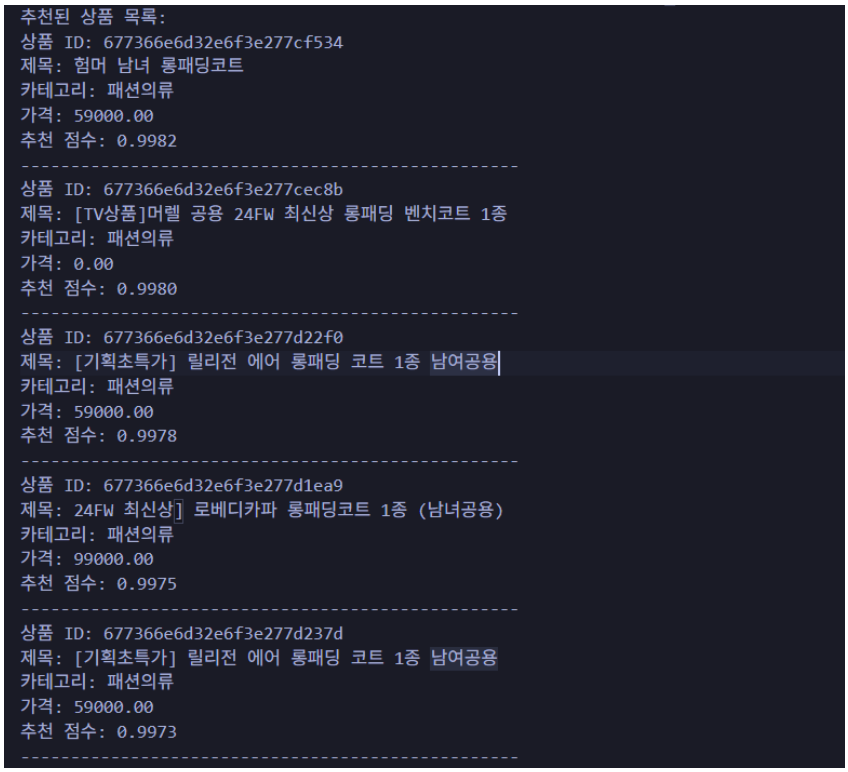
사용자 입력

- 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.
