모델정의서

Team P

백지연, 이경준, 이예림, 최동연

IFITV 프로젝트 모델 정의서 목차

1. 프로젝트 개요

- **프로젝트명:** IFITV
- **목적**
    - 사용자의 취향을 정밀하게 반영한 VOD 콘텐츠 및 실시간 방송 추천 제공.
    - 복합 피처(이미지·텍스트·메타데이터·행동 데이터)를 활용한 최적화된 맞춤 추천 시스템 개발.
    - 추천 품질 혁신(정확도·다양성) 및 대규모 자동화 서비스 지원.

2. 데이터 설명

2.1 공통 데이터 소스

- **메타데이터**
    - 작품명(title), 장르(genre), 서브장르(subgenre), 썸네일 이미지 URL(thumbnail), 줄거리(description), 출연진, 감독 등.
- **시청 로그/사용자 행동 데이터**
    - user_id, 콘텐츠 제목, 시청 이력, 시청 완료 비율, 찜(MyList), 사용 경로 등.

2.2 이미지 유사도 딥러닝 모델용 데이터

- **이미지 데이터**
    - TVING 플랫폼 내 모든 콘텐츠의 대표 썸네일 이미지를 수집, 사전 전처리(결측치·중복 제거) 후 사용.
    - 임베딩: ResNet50 기반 이미지 벡터(2048 차원).
- **텍스트 데이터**
    - 각 콘텐츠의 공식 줄거리(description).
    - 임베딩: KoBERT(768 차원), TF-IDF(명사 추출, 실험적 사용).
- **메타 정보**
    - 장르, 서브장르 정보(카테고리형), Jaccard 일치 기준 벡터화.

2.3 콘텐츠 기반 하이브리드 추천 모델용 데이터

- **원본 메타데이터**
    - title, synopsis, genre, subgenre, cast, director, thumbnail_url 등.
- **사용자 로그**
    - 시청 횟수, 시청 완료 비율, 찜, 시청 시간, 선호 장르/서브장르, 개인별 행동피처 기록 등.
- **임베딩 데이터**

    - TF-IDF 임베딩: 상위 1,000~2,000 개 단어 기준 줄거리 희소행렬(.npz)
    - KoBERT 임베딩: synopsis 문장 벡터(.npy, 768 차원)
- **실시간 편성표 정보**
    - 방송 시간, 채널, 실시간 인기 등.

3. 모델 설계

3.1 이미지 유사도 딥러닝 하이브리드 모델 (ResNet50 + KoBERT)

3.1.1 구조 개요

- 썸네일 이미지 임베딩(ResNet50, 2048 차원)
- 줄거리 임베딩(KoBERT, 768 차원)
- 장르/서브장르 범주형 벡터 Jaccard 일치 기반
- 각 피처별 유사도 계산 후 가중치 합산

3.1.2 세부 파이프라인

- **입력:** 추천 기준 콘텐츠 선택
- **임베딩 추출:**
    - 썸네일 → ResNet50(0.3)
    - 장르(0.2), 서브장르(0.2) → 1 또는 0(Jaccard/일치 판단)
    - 줄거리 → KoBERT(0.3, 의미 임베딩)
- **유사도 계산:** 코사인 유사도 및 범주 일치 점수 산출
- **최종 점수 공식:**
    - final_score = 0.3 * thumb_sim + 0.2 * genre_sim + 0.2 * subgenre_sim + 0.3 * summary_bert_sim
- **Top-K 추천:** 점수 상위 N 개 콘텐츠 선정

3.1.3 특징 및 장점

- 이미지 기반·장르·텍스트 의미 정보를 복합 반영해, 표면적/내면적 유사성 동시 평가 가능
- TF-IDF 방식의 한계 극복(BERT 임베딩의 문맥 이해 덕분)
- 가중치 조정을 통한 추천 다양성 및 품질 튜닝 가능
- 실시간/배치 시스템에 손쉽게 적용 가능

3.1.4 실험 결과 및 개선 방향

- 단일 이미지 임베딩만 사용할 경우에는 시각적 유사성 위주, 실제 맥락 반영 어려움
- 하이브리드(장르, KoBERT 임베딩) 추가로 추천 정확도·다양성 대폭 향상

- 줄거리 데이터 품질이나 장르 편중이 추천 다양성에 영향, BERT 비중 조정 및 추가 메타 피처(감독, 출연진 등) 확장 고려 필요

3.2 콘텐츠 기반 하이브리드 추천 모델

3.2.1 구조 개요

- 텍스트 유사도(TF-IDF, KoBERT), 장르·서브장르, 사용 행동 특성 등 복합 피처 사용
- 행동 데이터, 시청률 등 메타데이터 결합(see 아래 표)
- 분류기(LR 등) 기반 top-N 추천

3.2.2 세부 파이프라인

- **특징 벡터 생성:**
    - TF-IDF 줄거리 유사도
    - KoBERT 줄거리 유사도
    - genre_sim, subgenre_sim(일치/Jaccard)
    - 시청 횟수, 완료율, 출연진·감독 일치, 인기/실시간 지표 등
- **피처 엔지니어링:**
    - StandardScaler 등 전처리
- **분류 모델:**
    - Logistic Regression 등 예측 모델
    - fit(X_train, y_train) → predict_proba(X_test)
- **최적 추천 리스트 생성:**
    - 사용자 기준 Top-N 출력

피처명	설명
watch_count	사용자의 콘텐츠 시청 횟수
content_sim	TF-IDF·KoBERT 기반 줄거리 벡터 유사도
genre_sim	장르 일치 여부(1/0)
subgenre_sim	서브장르 Jaccard 유사도
view_percentage	최대 시청 완료 비율
cast_overlap	사용자와 중첩 출연진 수
director_match	감독 일치 여부(1/0)
popularity	실시간/전체 시청률 평균

3.2.3 평가 및 운영

- Leave-One-Out 방식의 사용자별 교차 검증(실전 신규 추천 상황 유사)
- HR@10, NDCG@10, MRR@10 등 Top-N 지표로 성능 평가
- 단순 콘텐츠 피처 대비, KoBERT 임베딩·행동 데이터 결합 시 성능 급상승(최고 18 배 이상 향상)

- 하이퍼파라미터 튜닝(GridSearchCV), SHAP 등 피처 중요도 분석 지원

3.2.4 한계 및 발전 방향

- cold-start, 결측 시 에러 핸들링(사용자 이탈 최소화)
- 장르·서브장르 과다 편중 시 추천 다양성 한계 : 가중치·피처 확장 필요
- 출연진/감독, 사용자 실시간 피드백, 만족도/신뢰성 중심 신규 피처 도입 및 실시간 모델 적응