모델 정의서
&
성능 평가 방법

프로젝트명 : 추천시스템 기반 영화 스트리밍 사이트(PICK & FLIX)

2024. 12. 30.

1. 모델 개요

모델이름	▪ Contents-based filtering 추천 알고리즘
목적	▪ 사용자 시청 영화의 장르, 키워드 데이터를 기반으로 유사 영화 리스트 출력

2. 비즈니스 요구사항

문제정의	▪ 사용자가 시청한 영화 시청 이력을 기반으로 높게 평가하고 시청 완료율이 높은 영화와 유사한 영화 추천
성공기준	▪ 콘텐츠 재생 여부 ▪ 사용자 평점 ≥ 4.0/5

3. 데이터 정의

입력 데이터	▪ 유저 ID(user_id) ▪ 사용자 시청 기록(시청시간, 평점) ▪ 사용자 시청 영화 메타데이터(장르, 키워드)
테이터 전처리	▪ 최근에 시청했던 영화의 장르, 키워드 단어 통합 ▪ CountVectorizer : **장르** 집합에서 단어 토큰을 생성하고 각 단어의 수를 세어 BOW(Bag Of Words) 인코딩 벡터를 만듬. ▪ TfdifVercorizer: **키워드** 집합에서 TF-IDF 방식으로 단어의 가중치를 조정한 BOW 인코딩 벡터를 만듬
출력데이터	▪ 추천 영화 리스트 ▪ 추천 영화 유사도 결과

4. 모델 설명

알고리즘 유형	▪ (CountVectorizer or TfidfVertorizer) + 코사인 유사도 기반 추천
모델 구조	▪ 영화 메타데이터를 벡터화 하여 영화 간 유사도 계산 ▪ 사용자 선호 벡터를 생성하고 유사 영화 추출

5. 실험 및 결과

데이터셋 크기	▪ 20개년 한국 박스오피스 Top50(963개의 영화)
평가	▪ Precision@K - 상위 K개 추천 항목 중에서 사용자에게 실제로 유용한 항목의 비율을 측정 (클릭, 재생여부) 그림1. 성능평가 지표 $$\text{Precision@K} = \frac{\text{Relevant Items in Top K}}{\text{K}}$$
업데이트 요소	▪ 영화 콘텐츠가 증가함에 따라, 장르, 키워드 데이터가 커질수록 계산 리소스가 커질 수 있는 문제가 있음 → 사용자가 선호하지 않는 영화를 예측해 계산과정에서 제외가능 ▪ 내용기반 필터링과 협업기반 필터링 중 위의 평가 방식을 기준으로 더 높은 추천 알고리즘 채택이나 하이브리드 방식 적용 가능

※ 본 프로젝트는 배포 단계를 제외했기에 실제 성능평가에 한계가 있어 방법론 까지만 제시