[AI 사업 내 추천 모델링 관련] - 베이스라인 추천 모델 설명

서울대학교 인간중심컴퓨팅 연구실
2020/10/05 작성

- **추천 알고리즘 적용 대상 데이터**
  - 2016 - 2019년 (4개년) 에 해당하는 'MBN 종합뉴스' 의 보도 동영상 (약 1분 30초 단위 원천 동영상) 가운데 '경제' 카테고리로 분류 가능한 것
  - 미디어 카테고리 (MDA_CGR_NM) 기준 -- 경제, 증권, 부동산

카테고리	분류 코드	데이터 수	총 데이터 수
경제	mbn00003	4,521	4,527
증권	mbn00004	1	
부동산	mbn00005	5	

- 베이스라인 모델 (공통)
  - 보도 동영상의 메타데이터 만을 활용하여 각 동영상 컨텐츠 간 유사도 계산
  - 제공받은 메타데이터 (Analyzed_뉴스_MBN종합뉴스_2016(~2019).csv) 내 포함되어 있는 각 기사 스크립트에 대한 품사 태깅 (POS Tagging) 정보 활용
  - 전처리 (단어 필터링)
    - 1차 필터링: 모든 품사를 이용하지 않고 주요 품사 (아래 8개) 에 해당하는 단어만 활용

품사	일반 명사	고유 명사	동사	형용사
태그	NNG	NNP	VV	VA
품사	관형사	일반 부사	접속 부사	감탄사
태그	MM	MAG	MAJ	IC

    - 2차 필터링: 전체 단어 중 2회 이상 등장한 단어만 활용

    → 1차 필터링 후 690,295 단어 ⇒ 2차 필터링 후 최종적으로 19,348 단어 활용
  - 최종적으로 필터링 된 단어를 vocabulary set 으로 간주하고 각 동영상 컨텐츠에 해당하는 기사를 TF-IDF 벡터화하여 표현
  - 유저가 선택한 각 동영상 컨텐츠를 기준으로 가장 유사한 10개 동영상을 추천

- 베이스라인 모델 (1) reco_v1
  - 뉴스 스크립트 (텍스트) 만을 활용한 모델
  - 각 기사 스크립트를 TF-IDF 벡터로 표현 후, 이들 벡터 간 코사인 유사도 (Cosine similarity) 를 바탕으로 각 동영상 컨텐츠(기사) 간 유사도 계산하여 랭킹 부여

- 베이스라인 모델 (2) reco_v2
    - 뉴스 스크립트 (텍스트) 외, 보도 일자 간 인접성을 추가로 고려한 모델
    - reco_v1 모델과 동일한 방식으로 계산한 각 동영상 컨텐츠(기사) 간 유사도 점수 (1) 에 추가적으로 보도 일자 간 인접성 점수(페널티) (2) 를 더하여 랭킹 부여
    - 보도 일자 간 인접성 점수는 (두 기사의 보도 일자 차이) 를 제곱한 후, 이를 해당 기사 기준 보도 일자 차이의 최대값으로 나누어 표준화한 값으로 정의하여 사용함
      → penalty = - [(days_diff)**2 / max(days_diff)] (-1 <= score <= 0)
    - 기사 스크립트 간 유사도 점수가 높게 나오더라도, 해당 기사와 보도된 시기가 크게 차이날수록 큰 페널티 부여하여 상대적으로 시기적으로 인접한 기사를 우선 추천하게 됨