Content-based Filtering 기반 공공도서관 인기도서 추천 시스템

TEAM: 에폭 히어로즈

팀원: 한우림, 송태원

프로젝트 개요

- 본 프로젝트는 EPOCH 3기 멤버들의 홈페이지에 기재된 interests 정보를 바탕으로, 공공도서관에서 보유 중인 인기 도서들 중, 관심 분야와 유사도가 높은 상위 도서를 추천해주는 시스템 구현을 목표로 한다.
- 저희 시스템은 콘텐츠 기반 추천(CBF)으로, 책 소개에 입력 키워드와 유사한 설명을 가진 도서들을 추출하여 상위 대출 수 도서를 추천합니다.
- 나아가 외부/리뷰 데이터를 크롤링하여 더 고도화된 맞춤형 추천 시스템으로 활용될 것을 기대한다.

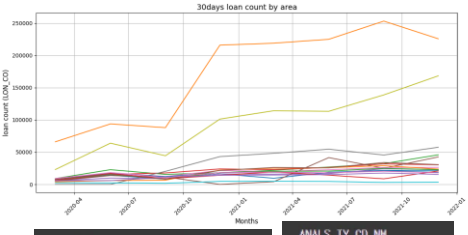
수행 과정

- 1. 인기대출(loan_information)데이터셋 정제
- 2. 도서별 고유 식별키 생성, 분석기간 "30일" 기준 필터링
- 3. 도서별 데이터셋 내 가장 최신 대출 정보를 기준으로, 지역별 대출 수를 전처리 mapping한 뒤, 총 대출 수를 집계
- 4. [콘텐츠 기반 필터링 모델 적용]
  유사도 계산: TF-IDF + 코사인 유사도
- 5. 유사 도서 中 대출 수 기준 상위 n개 추천

ANALS_PD_CD_NM	
90일	502000
30일	501000
7일	495532

Gangwon	Gyeonggi	Gyeongnam	Gyeongbuk	Gwangju	...	Busan	Seoul	Sejong	Ulsan	In
26	507	52	32	28	...	50	215	4	13	
49	936	32	0	11	...	79	75	0	0	

상세 내용



ANALS_PD_CD_NM	
90일	502000
30일	501000
7일	495532

ANALS_TY_CD_NM	
연령 및 성별	861542
지역별	396990
연령별	216000
전체	24000

dtype: int64

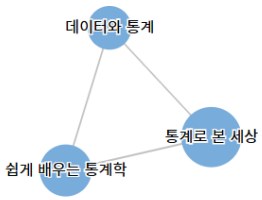
기간별 분석 대출 수를 분석했을 때, 30일 단위가 50만개 존재

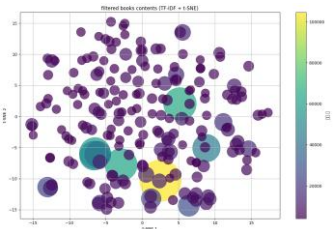
지역별 30일 기준 대출 수 추이 확인 – '경기도' 지역 내 대출 빈도가 모든 시계열 포인트에 대해 최상위 기록

30일 기준 > 지역별 > 도서별 가장 최신 기준의 데이터 필터링

지역별 대출 수를 column으로 전환하여 지역별 데이터 반영

ISBN 번호가 식별키 역할을 X -> [BOOK_TITLE_NM, BOOK_INTRCN_CN]





e.g. "통계" 키워드에 대한 콘텐츠 기반 필터링 모델 추천 결과

$$\text{cosine\_sim}(u, v) = \frac{\sum_{i \in I_{uv}} r(u,i) \cdot r(v,i)}{\sqrt{\sum_{i \in I_{uv}} r(u,i)^2} \cdot \sqrt{\sum_{i \in I_{uv}} r(v,i)^2}}$$

- 불용어 리스트 할당 및 불용어 제거
- 형태소 분석, 토큰화
- 정규화(정규표현식 사용)
- 벡터화 및 유사도 계산

결과 및 기대효과

- 입력되는 키워드별로 다르지만, "수학" 키워드에 대해서는 0.4 이상의 유사도 측정
- 추천 시스템의 보다 나은 예측 정확도를 위해 "협업 필터링 모델"을 활용한 시스템까지 확장할 필요가 있어 보인다
- 객관적인 실제값과 성능 평가를 위해 도서에 대한 리뷰/별점과 같은 외부 데이터를 추가로 활용
- 제목 + 저서 내용을 동시에 활용한 복합 유사도 분석에 대한 구현