추천 알고리즘 (Recommendation Algorithm)

* 콘텐츠 기반 필터링 (Content-based Filtering)
* 협업 필터링 (Collaborative Filtering)
* 콘텐츠 기반 필터링 (Content-based Filtering)
* 콘텐츠(아이템)의 특성과 사용자의 선호도를 비교해 추천하는 방식
* 사용자가 선호하는 아이템을 기반으로 해당 아이템과 유사한 아이템을 추천
* Item Profile
* 추천 대상이 되는 아이템들에 대해서 Profile을 작성
* 단어의 중요도를 알아내기 위해 기본적인 방법으로 TF-IDF 활용
* TF-IDF
* TF = Term Frequency, 특정 단어가 특정 문서에 등장하는 횟수
* IDF = Inverse Document Frequency, 전체 문서 중 특정 단어가 등장한 문서의 비율의 역수
* TF-IDF = TF \* IDF
* User Profile
* 사용자가 선호하는 아이템의 벡터를 통해 작성
* Simple 방식 vs Variant 방식
  + Simple 방식 : 사용자가 선호한 Item Vector들의 평균 값 사용
  + Variant 방식 : 사용자가 아이템에 내린 선호도로 Normalize하여 평균 값 사용
* Item Profile 과 User Profile을 이용하여 추천
* 코사인 유사도 (Cosine Similarity) 활용
* 장점
  + 다른 유저의 데이터가 필요하지 않음
  + 추천할 수 있는 아이템의 범위가 넓음 = 새로운 아이템이나 인기 없는 아이템도 추천이 가능
  + 추천하는 이유를 제시 가능 = 아이템의 특징으로 분석하기 때문에 어떤 특징이 추천의 이유가 되었다고 설명 가능
* 단점
  + 적절한 특징을 찾기가 어려움
  + 새로운 유저를 위한 추천이 어려움 = 새로운 유저의 데이터가 존재하지 않거나 데이터가 부족하다면 해당 유저에게 추천 시스템을 적용하기 어려움
  + 선호하는 특성을 가진 항목을 반복 추천

▶︎ 사용자 정보를 활용해 희망 기업의 요구or우대 기술 스택에 대한 데이터와 비교하여 추천하도록/사용자가 희망하는 기술 스택에 대한 프로젝트를 사용자들의 선호도 집계를 통해 높은 점수를 가진 프로젝트를 사용자에게 추천

* 협업 필터링 (Collaborative Filtering)
* 유사도를 기반으로 아이템을 추천하는 방식.
* 특정 집단에서 발생하는 ‘유사한 사용 행동’을 파악하여, 비슷한 성향의 사람들에게 아이템을 추천하는 방식
* 성향이 비슷하면, 선호하는 것도 비슷할 것이라는 가정을 전제

1. 사용자 기반 협업 필터링 (User-based CF)

* 나와 성향이 비슷한 사람들이 사용한 아이템을 추천해 주는 방식

ex. 사용자 A가 온라인 몰에서 치킨, 피자, 콜라를 구매하고 사용자 B가 치킨과 콜라를 구매했다고 가정할 때, 알고리즘은 구매 목록이 겹치는 이 두 사용자가 유사하다고 판단해, 사용자 B에게 피자를 추천

1. 아이템 기반 협업 필터링 (Item-based CF)

* 사용자가 구매하려는 물품과 함께 구매된 경우가 많은 아이템을 추천해 주는 방식

ex. 공책과 볼펜을 함께 구매하는 소비자가 많다면, 공책을 구매한 사용자에게 볼펜을 추천. + 이때 두 상품의 특징은 파악하지 않음

즉, 공책과 볼펜이 사무용품인지, 서로 같이 관용되는 관계인지 등은 고려하지 않음. 다만 두 제품이 같이 구매된 기록이 많기에 새로운 사용자에게 추천할 뿐

* 고객의 선호도 정보를 활용하는 면에서는 사용자 기반의 협업 필터링과 동일하나, 인접이웃이 아닌 아이템 간의 유사도를 이용하여 추천을 해주는 점에서 차이가 있으며, 고객이 구매한 상품과 목표 상품 간의 유사도만 계산하여 추천에 반영
* 장점
  + 다수의 사용자에게 얻은 정보로 새로운 아이템을 추천
  + 직관적으로 이해하기 쉽고 합리적으로 보임
* 단점
  + 콜드 스타트 (Cold Start) : 새로운 아이템이나 사용자가 추가되면, 충분한 사용기록이 확보될 때 까지는 적절한 투표를 하기 어려움
  + 롱테일 (Long Tail) : 인기 편향성의 문제라고도 하며, 사용자가 소수의 아이템만 선호하여 대다수의 비인기 아이템들은 추천을 위한 충분한 정보가 쌓이지 못함
  + 계산 효율 저하 : 협업 필터링은 계산량이 비교적 많은 알고리즘 이기 때문에 사용자 수가 많은 경우 계산이 몇 시간에서 길게는 며칠까지 소요되기도 함.
* 선호도 추정 방식
* 직관적으로 A라는 아이템에 대한 선호 또는 동의를 표현했다면, 이와 유사한 다른 상품에 대해서도 높은 관심을 가질 수 있음을 가정해야 함.

▶︎ 사용자 기반 협업 필터링 방식 활용 시 사용자들끼리 기술 스택 or 활동 프로젝트 성향이 비슷한 내용으로 프로젝트를 추천하여 멘토링/아이템 기반 협업 필터링 방식을 활용하려면 사용자들이 기술 스택의 프로젝트 내용 데이터들의 양이 많아야 함.

* 협업 필터링 (Collaborative Filtering) 적용 기업
* Amazon, Spotify, Netflix, Kakao 등
* Amazon
  + Recommended for you
  + Customers who viewed this also viewed
  + Customers who bought this also bought
  + 위와 같은 영역을 구성하여 고객에게 추천 상품을 제공
* YouTube
  + 개인 사용자의 시청 이력
  + 채널 영상의 실적
  + 계절성 요인 외부 요인을 통해 개인별로 맞춤 추천
* Netflix
  + 유사 사용자 기반 알고리즘
  + 유사 아이템 기반 알고리즘
  + 잠재 모델 기반 알고리즘ss
  + 콘텐츠 기반 알고리즘
* 아이템 기반 협업 필터링 (Item-based Collaborative Filtering)
* 사용자 기반 협업 필터링 (User-based Collaborative Filtering)의 단점과 한계를 극복하고자 미국의 이커머스 기업인 아마존이 개발한 방법
* 코사인 유사도 (Cosine Similarity) 를 활용하여 추천 아이템 찾기

1. 비슷한 아이템 찾기
   1. 아이템 간의 유사성을 계산 (벡터의 유사성을 구하는 기술을 활용)
2. 아이템 별 평점 예측
   1. 평점 데이터가 있다고 가정
   2. 사용자가 이미 평점을 준 아이템들을 검색
   3. 코사인 유사도를 활용하여 각 아이템 간의 유사도를 계산
   4. 유사도를 가중치로 사용하여 가중평균을 내면 해당 아이템에 대한 예상 평점이 나옴
3. 예상 평점이 높은 아이템 추천
   1. 사용자가 아직 접하진 못한 상품들의 예상 평점을 내림차순으로 정렬하여 가장 상위에 있는 아이템을 추천 상품으로 제시

* 추천 시스템에서 사용하는 데이터
* 사용자 정보
* 사용자의 프로필 정보를 의미
* ex. 사용자의 나이, 성별, 지역, 학력 등 개인적인 신상정보 혹은 사용자의 웹페이지 방문 기록, 클릭 패턴, 페이지 체류 시간 등 행동 정보 로그 등
* 사용자의 평점, 리뷰와 같은 직접적인 피드백
* 사용자가 웹페이지 or 앱에 다녀간 흔적 로그 등과 같은 간접적인 피드백
* 위 피드백들을 통해 데이터를 수집
* 아이템 정보
* 사용자에게 제공하려는 아이템은 서비스마다 다르게 정의할 수 있음
* 아이템의 프로필과 같은 고유 정보를 활용하거나 아이템을 사용한 사용자의 리뷰/후기와 같은 집계 정보 등을 활용
* 아이템 추천 점수
* 사용자가 해당 아이템을 얼마나 좋아했는가, interaction이 있었는가(있었다면 얼마나 자주 있었는가) 등을 점수화(scoring)한 값
* 사용자들이 얼마나 좋아할지에 대한 정량화된 기준이 필요하기 때문에 해당 데이터가 필요

▶︎ 사용자 정보 : 사용자가 입력한 학력, 기술 스택, 활동 프로젝트, 대외 활동 등을 입력하면 해당 정보 로그들을 저장, 사용자가 남긴 후기/평점 등의 로그들 저장

▶︎ 아이템 정보 : 기술 스택 별 추천할 프로젝트 데이터 저장해야 됨.

▶︎ 아이템 추천 점수 : 사용자가 추천 받은 해당 기술 스택의 프로젝트가 얼마나 좋았는가, interaction이 있었는가(있었다면 얼마나 자주 있었는가) 등을 점수화한 값 / 사용자들이 얼마나

* 유튜브 영상 알고리즘 원리
* 몇 가지 근거를 기반하여 사용자의 취향을 탐색하고, 그에 따른 추천 동영상을 유튜브 홈에서 보여줌
* 가장 먼저 시청자의 시청 이력과 선호도를 기반으로 개인화 데이터를 수집
* ex) 함께 자주 보는 동영상이나 주제 관련 동영상, 사용자가 과거에 본 동영상 등을 추천 동영상으로 큐레이팅
* 사용자와 비슷한 시청자들이 관심 있을 법한 소재의 동영상을 찾아내는데, 이것을 ‘알고리즘 선택’이라고 함
* 동영상의 성공 여부에 따른 추천
* 사용자들의 시청 시간이 오래 되었거나, 꾸준히 좋은 조회수 지표를 기록한 동영상을 추천
* 개인의 취향과 더불어 사용자들이 전반적으로 만족했던 동영상이 함께 제공된다면, 알고리즘에 대한 신뢰나 만족도가 훨씬 더 높아지게 됨
* 유튜브 알고리즘 구조
* 플랫폼에서 사용자의 활동에 대한 데이터를 수집하는 것
* 시청하는 동영상, 시청 시간, 동영상과 상호 작용하는 방식이 포함됨
  + ex) 좋아요, 댓글, 공유나 구독 등의 행위
* 사용자의 활동 데이터에서 관련 특징을 추출
* 제목, 설명, 태그와 같은 비디오에 대한 메타 데이터와 위치 및 장치 유형과 같은 사용자에 대한 정보가 포함
* 추출된 특징으로는 임베딩이라는 프로세스를 사용하여 벡터 표현으로 변환됨.
* 벡터 표현은 신경망에 입력되어 사용자가 주어진 비디오에 참여할 확률 (ex. 시청, 좋아요 등)을 예측하는 방법을 학습
* 신경망은 네트워크의 가중치를 조정하여 예측 오류를 최소화하는 역전파라는 기술을 사용하여 사용자 활동 데이터의 대규모 데이터 세트에서 훈련됨
* 훈련이 끝난 신경망은 각 비디오와 사용자 선호도, 평균 조회율 간의 유사성 점수를 매기는 데에 사용할 수 있음.