텍스트 마이닝과 데이터 마이닝

Part 09. 추천 시스템

정 정 민

programmers

Chapter 24. 추천 시스템 알고리즘

1. 컨텐츠 기반 필터링
2. 협업 필터링

콘텐츠 기반 필터링

콘텐츠 기반 필터링 적용 과정

- 콘텐츠 기반 필터링의 주요 단계는 아래와 같음

1. 아이템 프로파일을 구성

- 프로파일 : 아이템을 설명할 특성의 모음집.
- 예를 들어, 영화 A : [감독, 장르, 배우, 제작 년도, 등등]

2. 각 특성 당 정보 추출

- 프로파일을 구성하는 특성을 숫자의 형태로 변경
- 각 특성 데이터에 적합한 embedding을 진행

3. 아이템들의 프로파일을 생성

- 사용자의 아이템
- 추천에 사용될 아이템

4. 유사도 계산 및 추천

- 사용자 프로파일과 아이템 프로파일 간의 유사도를 계산
- 코사인 유사도, 유클리드 거리, 피어슨 상관계수 등이 사용
- 상위 N개를 추천에 활용

1. 아이템 프로파일 생성

- **프로파일**이란
  - 특정 아이템을 설명하는 특성들의 집합
  - 풀어야 하는 문제에 따라 다르게 설정 가능

- 특성은 서로 다른 이종 데이터의 집합도 가능
- 예를 들어, 유튜브 영상이라면
  - 영상의 주제(text), 조회수(number), 좋아요(number), 인기 프레임(image) 등등
  - 이것들을 순차적으로 모아서 구성

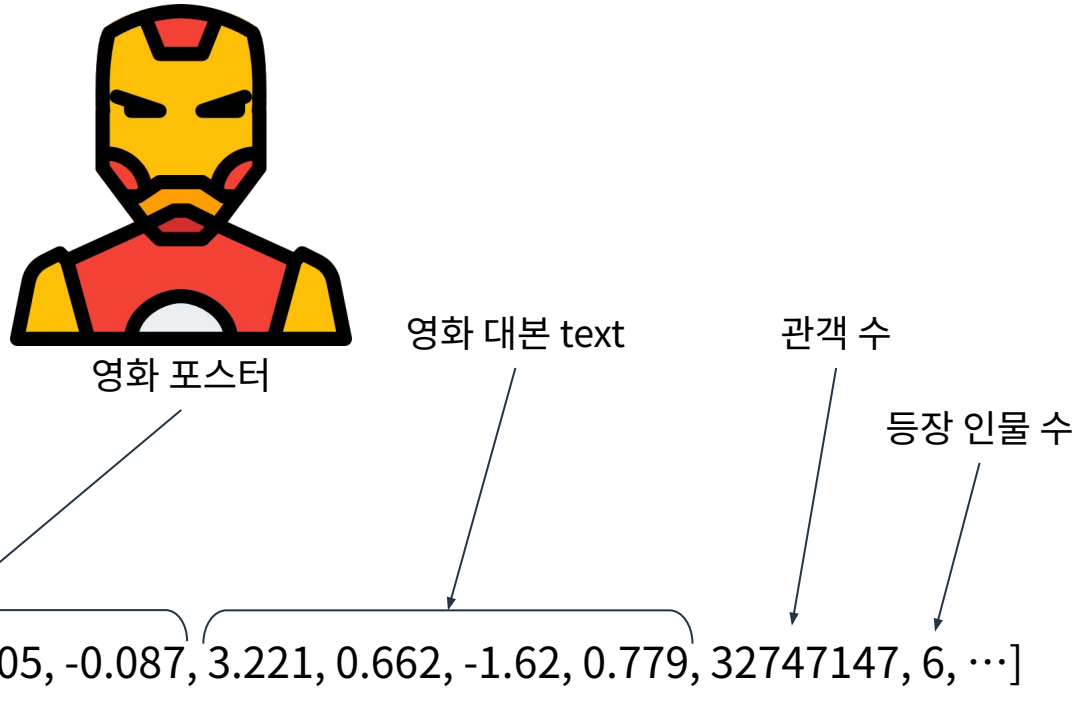
- 이러한 특성을 어떻게 구성하는지에 따라
- 추천 시스템의 성능이 달라짐

2. 각 특성 당 정보 추출

- 특성 그 자체는 사람이 지정하는 요소로
- 추천 과정에서 필요한 계산을 진행할 수 없음

- 계산이 가능하도록 특성을 숫자로 변경해야 함 : Embedding
  - 텍스트라면 NLP를 활용해
  - 이미지라면 Image 처리를 활용해
  - 숫자 데이터라면 필요에 따라 범주형으로 혹은 그대로 사용
  - 등등

- 이 과정은 모든 아이템에 동일하게 적용되어야 함

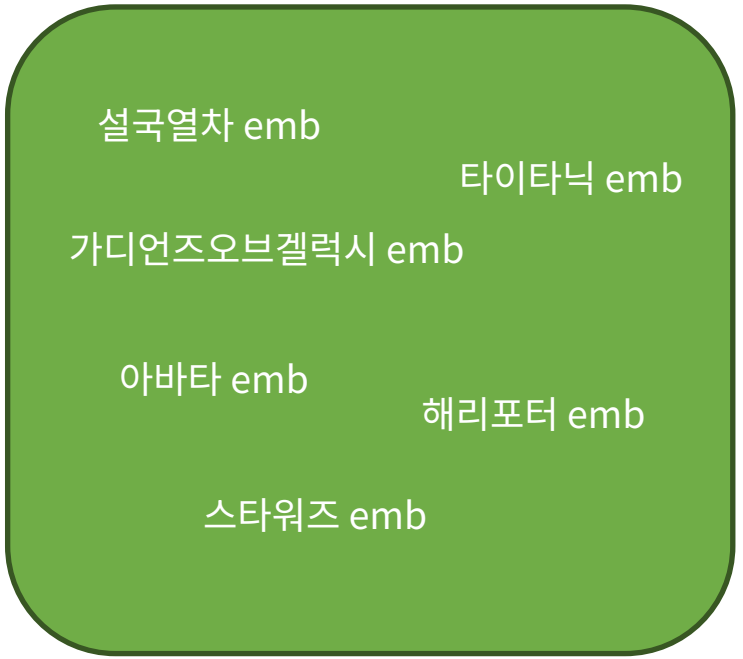


3. 아이템들의 프로파일 생성

- 사용자가 사용한 아이템의 프로파일과
- 사용하지 않은 아이템의 프로파일을 생성하는 과정

- 동일한 embedding 과정을 거쳐야 함

사용자 프로파일



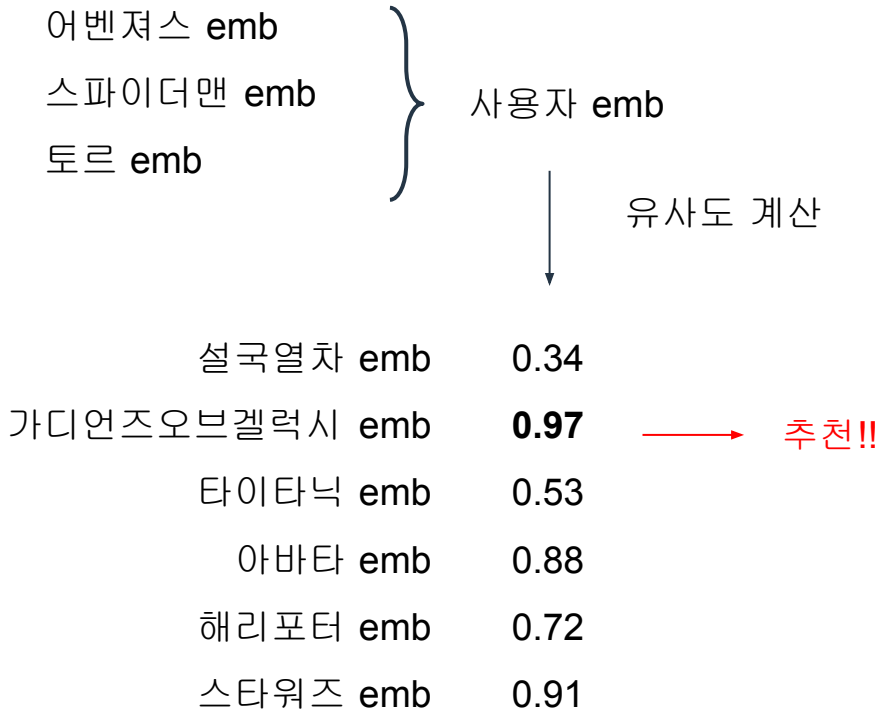
아이템 프로파일

4. 유사도 계산 및 추천

- 사용자가 갖고 있는 embedding 값과
- 다른 아이템들의 embedding 값과 유사도를 계산

- 만약 사용자가 갖고 있는 아이템이 다수라면
  - 단순 평균
  - 사용자가 남긴 다른 메타 정보로 가중합
- 의 과정으로 사용자의 최종 embedding 값 도출

- 유사도를 기반으로 상위 N개를 추천
- 아직 경험하지 않은 아이템 중 유사도가 높은 아이템을 추천



협업 필터링

협업 필터링

- 사용자들 사이의 상호작용 데이터 혹은 선호도 패턴 데이터를 기반으로 추천을 수행

- 크게 두 가지 주요 방식
  - **사용자 기반 (User-based)**
    - 사용자들의 아이템 선호 데이터를 활용
    - 비슷한 선호도 또는 행동 패턴을 보이는 사용자의 선호를 추천
  - **아이템 기반 (Item-based)**
    - 사용자들이 아이템을 평가한 데이터를 활용
    - 특정 사용자가 사용한 아이템의 평가와 비슷한 아이템을 추천

사용자 기반 협업 필터링

- 이를 위해서는 아래의 주요 과정이 필요

1. 상호 작용 데이터 준비

- 사용자로부터 얻은 데이터 활용 (예: 평점, 클릭, 구매 등)
- 기존 데이터가 없다면 사용할 수 없음 : 콜드 스타트 (Cold Start)

2. 사용자(행) – 아이템(열) 상호 작용 테이블 생성

- 상호 작용이 없는 경우 누락값으로 표시

3. 사용자 간 유사도 추출

- 유사도란 사용자들이 아이템에 대해 얼마나 비슷한 반응을 보였는지를 의미
- 코사인 유사도, 피어슨 상관계수 등이 활용

4. 유사도 기반 아이템 추천

- 유사도를 기반으로 “이웃”을 선정
- 이웃이 높게 평가한 아이템을 추천!

사용자 기반 협업 필터링 예시

- 영화A ~ 영화E, 사용자 a~f에 대한 사용자 협업 필터링 예시
- 영화를 보고 난 후의 평점(1~5) 데이터를 활용

- a의 이웃은 : c > d, e
- a가 본 영화가 아닌 것 중, 이웃이 높게 평가한 영화 : **영화E**

	영화A	영화B	영화C	영화D	영화E
a	4	4	4		
b	1		5		1
c	2	2		1	
d			1	4	1
e	4	4			4
f			3	5	1

	a	b	c	d	e	f
a	1.000	0.667	0.770	0.136	0.667	0.293
b	0.667	1.000	0.128	0.272	0.222	0.520
c	0.770	0.128	1.000	0.314	0.770	0.282
d	0.136	0.272	0.314	1.000	0.136	0.956
e	0.667	0.222	0.770	0.136	1.000	0.098
f	0.293	0.520	0.282	0.956	0.098	1.000

아이템 기반 협업 필터링

- 아이템 간의 유사성을 분석해 추천을 수행하는 방법
- 아래의 과정으로 구성

1. 상호 작용 데이터 준비

2. 아이템 간 유사도 계산

- 수집된 상호 작용 데이터를 기반으로 아이템 유사도를 계산
- 코사인 유사도, 피어슨 상관계수 등

3. 추천 점수 계산

- 아이템 유사도와 사용자 상호 작용 데이터를 조합해 추천 점수 계산 (뒷장에 설명!)
- 사용자가 평가한 아이템과 유사한 아이템들의 대해 가중 평균

4. 최종 아이템 추천

- 추천 점수를 기반으로 추천 진행

아이템 기반 협업 필터링 예시

- 동일한 예시 데이터

- 사용자 a가 아직 시청하지 않은 영화 : D, E
- D의 추천 점수 : $4 \times 0.0507 + 4 \times 0.0514 + 4 \times 0.4105 = 0.0504$
- E의 추천 점수 : 동일한 방법으로 = 6.1684

	영화A	영화B	영화C	영화D	영화E
a	4	4	4		
b	1		5		1
c	2	2		1	
d			1	4	1
e	4	4			4
f			3	5	1

	영화A	영화B	영화C	영화D	영화E
영화 A	1.0000	0.9864	0.4834	0.0507	0.6412
영화 B	0.9864	1.0000	0.3734	0.0514	0.6118
영화 C	0.4834	0.3734	1.0000	0.4105	0.2891
영화 D	0.0507	0.0514	0.4105	1.0000	0.3186
영화 E	0.6412	0.6118	0.2891	0.3186	1.0000

E.O.D