Java Collection Framework

1. 주제: 추천 시스템

○ 추천 시스템이란?
- 사용자의 기존 행위(콘텐츠 시청, 물건 구입 등)를 분석하여 사용자가 흥미를 느낄 수 있는 콘텐츠나 제품을 추천
- 응용 분야: OTT, 전자 상거래

○ 추천 시스템의 두 가지 유형
- 콘텐츠 기반의 추천: 사용자가 기존에 구매한 콘텐츠의 특성을 분석하여 이와 유사한 다른 콘텐츠를 추천
- Collaborative Filtering: 사용자 A의 구매 행위와 유사한 다른 사용자 B를 찾은 다음, B의 구매 내용을 바탕으로 A에게 추천

○ 입력: 사용자-콘텐츠 행렬
- 사용자는 자신이 구매한 콘텐츠에 대해 1점 ~ 5점 사이의 점수를 부여

콘텐츠 / 사용자	A0	A1	B0	B1	C0	C1
0	5	4	1			
1	3		4	4		
2			1		3	2
3	5			5	3	
4		3	2			5
5		5	3	5		

- 사용자마다 점수를 주는 성향이 다르므로, 위의 데이터를 그대로 사용하는 것은 불합리하다. (예: 2번 사용자는 전반적으로 부정적인 평가, 5번 사용자는 긍정적)
- 점수의 정규화 과정을 수행: 사용자가 부여한 점수의 평균을 구한 다음, 각 점수에서 평균을 뺀 값으로 수정. 그 결과로 각 사용자의 점수의 합은 0

콘텐츠 / 사용자	A0	A1	B0	B1	C0	C1
0	5/3	2/3	-7/3			
1	-2/3		1/3	1/3		
2			-1		1	0
3	2/3			2/3	-4/3	
4		-1/3	-4/3			5/3
5		2/3	-4/3	2/3		

○ Collaborative Filtering에서 두 사용자 간에 유사도 계산

- 사용자의 특징을 행벡터로 표현할 수 있으므로, Cosine 유사도 사용

- Cosine 유사도 = $\frac{A \cdot B}{||A|| \times ||B||} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$

  - 사용자 0과 1의 유사도 = $\frac{\frac{5}{3} \times (-\frac{2}{3}) + (-\frac{7}{3}) \times \frac{1}{3}}{\sqrt{(\frac{5}{3})^2 + (\frac{2}{3})^2 + (-\frac{7}{3})^2} \times \sqrt{(-\frac{2}{3})^2 + (\frac{1}{3})^2 + (\frac{1}{3})^2}}$

    = -0.7858252779857412

  - 사용자 0과 나머지 사용자 사이의 유사도

    (1, -0.7858252779857412) (2, 0.560448538317805) (3, 0.23112508176051225)

    (4, 0.45425676257949804) (5, 0.7396002616336388)

○ 추천 콘텐츠 선정

- Target 사용자와 유사도가 가장 높은 n 명의 사용자(참고인)을 선택
- 참고인이 구매한 콘텐츠 중에서 target 사용자가 구매하지 않은 콘텐츠들에 대해
  - (참고인의 유사도) * (참고인이 부여한 콘텐츠 점수)를 합산
- 합산 결과가 가장 높은 k 개의 콘텐츠를 추천
- 예: Target 사용자가 0이며, n = 2일 경우
  - 참고인 = 사용자 5와 2
  - 대상 콘텐츠의 합산 점수 = (C0, 0.560448538317805),

    (B1, 0.49306684108909277), (C1, 0.0)
  - k = 1일 경우, C0를 추천

○ 입력:

- 파일 이름, Target 사용자, 참고인 수, 추천 항목 수? small.txt target n k
  - 입력 파일의 구성:

    사용자 수

    사용자 콘텐츠 점수 ← 사용자는 0부터 (사용자 수 - 1)까지의 정수

    ...
  - Target 사용자: target ← int
  - 참고인 수: n ← int
  - 추천 항목 수: k ← int

2. 제출 내용: HW1.java 하나의 파일만 제출

① public class HW1 (파일 내에 나머지 클래스들은 public이 아님)
② default package 사용
③ 프로그램 내에 주석은 모두 삭제
④ 한글 encoding은 MS949로 설정
← **위의 조건들을 만족하지 않는 과제물은 심사하지 않음!!**

3. 평가: 50점 만점

- Target 사용자의 (콘텐츠, 정규화된 점수)를 콘텐츠의 오름차순으로 출력 (10점)
- Target과 유사도가 가장 높은 n명에 대해 (id, 유사도)를 유사도의 내림차순으로 출력 (15점)
- Target과 유사도가 가장 높은 n명이 구매한 항목 중에서 target이 구매하지 않은 콘텐츠에 대해 점수의 내림차순으로 (콘텐츠, 점수)를 k개 출력. 점수가 같을 경우에는 콘텐츠의 이름이 작은 것을 먼저 출력. (25점)

- 앞부분이 틀리면, 그 이후는 0점 처리
- JDK 8로 컴파일하며, 프로그램 구성이나 성능에 심각한 문제가 있으면 실행 결과와 관계없이 감점 처리함
- 무작위 샘플링을 통하여 선택된 학생에 대해서는 대면 평가를 수행하며, 자신의 코드를 제대로 설명하지 못하면 감점 (과제 0점 처리 및 최종성적 한 등급 하향 조정까지 가능) 처리함
- Copy로 판단되는 과제에 대해서는 성적을 한 등급 하향 조정

4. 동작의 예

실행의 예 1:

```
파일 이름, target 사용자, 참고인 수, 항목 수? s.txt 0 2 2

1. 사용자 0의 콘텐츠와 정규화 점수:
   [(A0, 1.667), (A1, 0.667), (B0, -2.333)]

2. 유사한 사용자 id와 유사도 리스트
   사용자 id:     5, 유사도: 0.739600
   사용자 id:     2, 유사도: 0.560449

3. 사용자 0에게 추천할 콘텐츠와 추천 점수
   [(C0, 0.560), (B1, 0.493)]
```

```
6
0 A0 5
0 A1 4
0 B0 1
1 A0 3
1 B0 4
1 B1 4
2 B0 1
2 C0 3
2 C1 2
3 A0 5
3 B1 5
3 C0 3
4 A1 3
4 B0 2
4 C1 5
5 A1 5
5 B0 3
5 B1 5
```

실행의 예 2:

```
파일 이름, target 사용자, 참고인 수, 항목 수? l.txt 5 10 5
1. 사용자 5의 콘텐츠와 정규화 점수:
   [(A5, -1.640), (A19, -1.640), (A44, -0.640), (A60, -2.640), (A166, -1.640), (B4, 1.360), (B5, 1.360), (B7, 1.360), (B17, 1.360), (B20, -0.640), (B24, 1.360), (B26, 1.360), (B31, 1.360), (B34, 1.360), (B50, 0.360), (B57, 0.360), (B135, -0.640), (B170, -0.640), (B192, 1.360), (C10, 0.360), (C172, -0.640), (D30, 1.360), (E28, -2.640), (E30, -1.640), (E64, 0.360)]

2. 유사한 사용자 id와 유사도 리스트
    사용자 id: 9771, 유사도: 0.282633
    사용자 id: 6112, 유사도: 0.262331
    사용자 id: 6713, 유사도: 0.253528
    사용자 id: 3738, 유사도: 0.249615
    사용자 id: 3926, 유사도: 0.241778
    사용자 id: 7300, 유사도: 0.238162
    사용자 id: 2308, 유사도: 0.237533
    사용자 id: 2891, 유사도: 0.231276
    사용자 id: 6901, 유사도: 0.229736
    사용자 id: 4809, 유사도: 0.226214

3. 사용자 5에게 추천할 콘텐츠와 추천 점수
   [(B28, 0.550), (B29, 0.495), (B14, 0.380), (B56, 0.380), (A32, 0.360)]

파일 이름, target 사용자, 참고인 수, 항목 수? l.txt 675 4 4
1. 사용자 675의 콘텐츠와 정규화 점수:
   [(A15, 1.200), (A18, 1.200), (A133, -0.800), (C22, 0.200), (C142, -1.800)]

2. 유사한 사용자 id와 유사도 리스트
    사용자 id: 6056, 유사도: 0.376823
    사용자 id: 9053, 유사도: 0.352342
    사용자 id: 1820, 유사도: 0.348465
    사용자 id: 4356, 유사도: 0.313112

3. 사용자 675에게 추천할 콘텐츠와 추천 점수
   [(A0, 0.723), (A30, 0.723), (A105, 0.723), (A16, 0.626)]
```