Assignment 2

File을 이용한 데이터 입출력, 데이터 변환, Euclidean Distance를 이용한 Item-based Collaborative Filtering을 실습한다.
Due : 2018.11.16 (금) 24:00

주어진 형식과 같이 각 사용자가 평점을 메긴 상품들의 목록이 입력파일로 주어졌을 때, Euclidean Distance를 이용하여 각 상품-상품 사이의 Similarity를 구한 후, 그 결과를 주어진 형식의 출력 파일과 같이 저장한다.

Input Data File (파일명: in.txt)

- Data Field : User, Item, Score
- 필드 구분 : '\t'

Output Data File (파일명: out.txt)

- Data Field : Item, Item, Similarity
- 필드 구분 : '\t'

Similarity(p, q)
= 1 / (1 + distance(p,q))

distance(p, q)
참고 : *https://en.wikipedia.org/wiki/Euclidean_distance*

$$d(\mathbf{p},\mathbf{q}) = d(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

주의사항

- 입력 데이터 및 출력 데이터는 반드시 파일로 처리해야 한다. (파일명 준수)
- 입출력 데이터 파일의 형식은 반드시 주어진 형식을 따른다.
    - 필드 구분은 반드시 탭('\t') 문자 사용
    - 데이터 파일의 필드 순서를 바꿀 수 없음
- 프로그램은 하나의 파일로 작성 및 제출한다.
- 모든 상품에 대해 서로간의 Similarity를 구하되, Similarity가 0인 것은 제외한다.

채점기준

- 제출하지 않거나 주어진 과제에 맞지 않는 것을 제출할 경우 0점
- 오류 메시지 없이 코드가 실행되어 결과를 출력하면 5~9점 (기본)

- 실행 결과의 값이 의미적으로 정확하면 10점 (만점)
- 코드의 품질이 상대적으로 우수하거나 추가적인 데이터 처리가 돋보이면 1점 추가
- 늦게 제출하면 하루에 1점씩 감점
- 다른 사람의 프로그램을 Copy했을 경우 최하 점수

Input Data Example (예, Ronaldo는 Apple에 대해 별 3개 부여)

```
Ronaldo Apple 3
Ronaldo Orange 5
Messi Apple 2
Messi Orange 4
Messi Mango 4
Mbappe Banana 4
```

Output Data Example (예, Apple과 Orange의 Similarity는 0.2612038749637414)

```
Apple Orange 0.2612038749637414
Apple Mango 0.3333333333333333
Orange Mango 1.0
Orange Apple 0.2612038749637414
Mango Orange 1.0
Mango Apple 0.3333333333333333
```

Similarity Example

Similarity(Apple, Orange) = 1 / (1+distance(Apple, Orange))

Apple의 평점 : {'Ronaldo':3, 'Messi':2}
Orange의 평점 : {'Ronaldo':5, 'Messi':4}

distance(Apple, Orange)

= $\sqrt{(3-5)^2 + (2-4)^2} = \sqrt{8}$

Similarity(Apple, Orange) = 1 / (1+ $\sqrt{8}$) = 0.2612038749637414

생각해 봅시다!

* 큰 입력 데이터를 사용하면 어떻게 될까?
* Similarity를 구할 수 있는 다른 Measure들은 무엇이 있을까?
* 주어진 Similarity 예시에서 추가로 고려할만한 포인트는 무엇이 있을까?