Item-based Collaborative Filtering for Recommendation

Instructor: Kunpeng Zhang   (kzhang6@uic.edu)
    TA: Minghong Xu   (mxu29@uic.edu)

The Deadline: 6:00PM, Mar. 16, 2015 (Monday)

1) Download the MovieLens 100K data from
http://grouplens.org/datasets/movielens/

2) Write your own Mapper and Reducer to compute similarity scores for all pairs of movies. Similarity measure could be Cosine similarity or Pearson correlation. The input is the provided dataset and the outputs are files in which each line would be like:

movieID_1, movieID_2, <similarity score>

(You don't need to generate movieID_2, movieID_1, <similarity score>, because they are the same. The item-item similarity matrix is symmetric.)

For Cosine similarity of two vectors X and Y, the similarity score S_{XY} is:

$$S_{XY} = \frac{\sum_{i=1}^{n} X_i * Y_i}{\sqrt{\sum_{i=1}^{n} X_i^2} \sqrt{\sum_{i=1}^{n} Y_i^2}}$$

For Pearson correlation similarity of two vectors X and Y, the similarity score S_{XY} is:

$$_{XY} = \frac{\sum_{i=1}^{n} (X_i - \bar{X}) * (Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}}$$

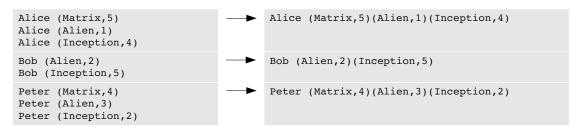**The key-value pairs of two mappers and reducers are shown below:**

Algorithm in Map/Reduce - Pass 1

Map - make user the key

(Alice,Matrix,5)	⟶ Alice (Matrix,5)
(Alice,Alien,1)	⟶ Alice (Alien,1)
(Alice,Inception,4)	⟶ Alice (Inception,4)
(Bob,Alien,2)	⟶ Bob (Alien,2)
(Bob,Inception,5)	⟶ Bob (Inception,2)
(Peter,Matrix,4)	⟶ Peter (Matrix,4)
(Peter,Alien,3)	⟶ Peter (Alien,3)
(Peter,Inception,2)	⟶ Peter (Inception,2)

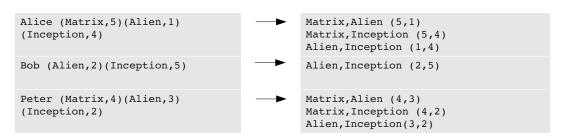## Reduce - create inverted index

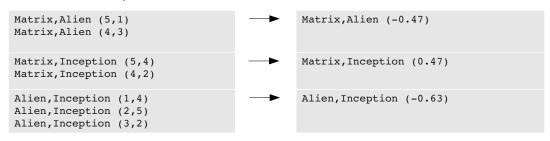|	
Alice (Matrix,5) Alice (Alien,1) Alice (Inception,4)	⟶ Alice (Matrix,5)(Alien,1)(Inception,4)
Bob (Alien,2) Bob (Inception,5)	⟶ Bob (Alien,2)(Inception,5)
Peter (Matrix,4) Peter (Alien,3) Peter (Inception,2)	⟶ Peter (Matrix,4)(Alien,3)(Inception,2)

Algorithm in Map/Reduce - Pass 2

Map - emit all cooccurred ratings

Alice (Matrix,5)(Alien,1) (Inception,4)	⟶ Matrix,Alien (5,1) Matrix,Inception (5,4) Alien,Inception (1,4)
Bob (Alien,2)(Inception,5)	⟶ Alien,Inception (2,5)
Peter (Matrix,4)(Alien,3) (Inception,2)	⟶ Matrix,Alien (4,3) Matrix,Inception (4,2) Alien,Inception(3,2)

Reduce - compute similarities

Matrix,Alien (5,1) Matrix,Alien (4,3)	⟶ Matrix,Alien (-0.47)
Matrix,Inception (5,4) Matrix,Inception (4,2)	⟶ Matrix,Inception (0.47)
Alien,Inception (1,4) Alien,Inception (2,5) Alien,Inception (3,2)	⟶ Alien,Inception (-0.63)

Requirements:

- You need to submit top 100 movie pairs in terms of similarity score.
- **Your codes must be readable and clean.**
- When you submit your codes through blackboard, you need to put all source codes (.java files, NOT jar files), a text file containing 100 movie pairs with their corresponding similarity scores, and some other optional files (e.g., a readme file) into one folder and name that folder as <YOUR UID>_ASSIGN3. Assignments not following this rule will not be graded. In addition, no resubmission after TA grades it. Late submission rule: 10% deduction for one day late. Late submission over a week is NOT acceptable.

DO NOT copy any codes from others. Otherwise, both will be penalized.