CSE 6240 - Spring 2015

Web Search & Text Mining

Homework 5

02/16/2015

Due: 03/08/2015 23:59

Collaborative Filtering

Description

Implement a movie recommending system using collaborative filtering.

The following files are provided and their detailed descriptions are in README.txt:

(1) ratings.csv: contains ratings matrix
(2) users.csv: contains users' information
(3) movies.csv: contains movies' information
(4) toBeRated.csv: contains the cells that you need to fill

For similarity, you must implement all three types of similarity below:

(1) Jaccard similarity
(2) Pearson correlation similarity
(3) Cosine similarity

Results:

(1) Use the user-based method and ratings.csv only to fill the ratings matrix to predict ratings for required cells in toBeRated.csv.
(2) Use the item-based method and ratings.csv only to fill the ratings matrix to predict ratings for required cells in toBeRated.csv
(3) Combing users' and items' information, using the best similarity type you think, to fill the ratings matrix to predict ratings for required cells in toBeRated.csv.

Testing:

You should use multi-fold cross validation to find which type of similarity is the best. Use RMSE as the metric.

Output Format

20,000 lines, line *i* has the rating for the cell denoted by line *i* in toBeRated.csv

For example, line 1 should contain the rating for user 3374 and movie 673.

Deliverable

The deliverable should contain two folders with 7 files, please put all the files in to a directory named "HW5-{YOUR FIRST NAME}-{YOUR LAST NAME}":

HW5-{YOUR FIRST NAME}-{YOUR LAST NAME}

\ -- code (50%)

| -- recommender.py/ recommender.cpp/ recommender.java (code) (40%)

| -- README.txt (showing how to run your code, your code should take types of similarity and input files as input arguments.) (10%)

\ -- results (50%)

| -- result1.csv (10%) (Results with best RMSE from three types of similarity)

| -- result2.csv (10%) (Results with best RMSE from three types of similarity)

| -- result3.csv (10%) (Results with best RMSE from three types of similarity)

| -- results.pdf (10%)

| -- a. Use a bar char to show RMSE for each method (2 item/user based method * 3 similarity measures + 1 your method = 7 methods). Describe your observations. (10%)

| -- b. Briefly explain your result1/result2. Explain how you get the result3 using what methods and which type of similarity and what information. Explain why using these settings will give the best result. (10%)

Note: Your can submit two version of your program, one for cross validation and the other for predication. Or, you can put them all in one program without triggering cross validation by default. Or handle this in anyway you feel comfortable about, as long as you explain it in README.txt

Please archive the folder and name it as "HW5-{YOUR FIRST NAME}-{YOUR LAST NAME}.zip". and upload it to T-square.