Assignment 1: User-based and Item-based Collaborative Filtering Recommendations

Group members: Tuomas Porkamaa, Mueed Irfan

We Implemented solutions for all subtasks.

How to run the code:

The submission consist four python files:

- *assignment1_1.py, assignment1_1_spyder.py, assignment1_2.py, assignment1_2_spyder.py*

Files with *1.py* and *1_spyder.py* prefixes implement user-based collaborative filtering approach. Correspondingly files prefixed by 2.*py* and *2_spyder.py* implement the item-based collaborative filtering approach.

Prefix "spyder" in the filename means that the program is recommend to run in Spyder-editor cell-by-cell manner. This gives the user the possibility to tweak hyperparameters without needing to run whole program again. Code is commented so that hyperparameters can be modified easily.

Files without "spyder"-prefixes are safe to run eg. in console environment. For example, *assignment1_1.py* can be executed by giving a command <*python assignment1_1.py*>.

NOTE BEFORE RUNNING THE CODE:

Modify the *root* variable to point to the directory (explicitly) where the program data is located (ratings.csv and movies.csv). Current setting is *root = '/home/tuomas/Python/DATA.ML.360/ml-latest-small/'*. This variable can be found in following lines:

- *assignment1_1.py,* line 163.
- *assignment1_1_spyder.py,* line 96.
- *assignment1_2.py,* line 39
- *assignment1_2_spyder.py,* line 6

Also, if you want to change the current user under examination, change the following lines:

- *assignment1_1.py,* line 190.
- *assignment1_1_spyder.py,* line 130.
- *assignment1_2.py,* line 53
- *assignment1_2_spyder.py,* line 23

Assumptions

Basically, we had following assumptions:

- If similarity score between two users cannot be calculated (eg. no similar movies exist between them), set it to 0.
- User have an opportunity to normalize the similarity scores between two users.
    - Code is initialized to run with normalized values. This can be changed by modifying the call *calculate_sim_matrix()* to *calculate_sim_matrix(False)* in lines 182 and 115 (*assignment1_1.py, assignment1_1_spyder.py*)
- User have an opportunity to "hardcode" some similarity scores to 0 if the score was

calculated from too few samples.

    - Code is initialized to accept all sample sizes. This can be changed by modifying the assignment *filt = cases_matrix < 1*. For example, *filt = cases_matrix < 5* forces all similarity scores which were calculated from less than five samples to zero.

- User-id's run from 0 to 609 instead of from 1 to 610.

Personal notes

Files *assignment1_1.py* and *assignment1_1_spyder.py* are overly complicated due to poor choice of basic data structure for data fetched from ratings.csv and movies.csv. This poor design was corrected in *assignment1_2.py* and *assignment1_2_spyder.py.*