Collaborative Filtering

Collaborative Filtering is one of the recommender systems, which predicts the missing values. In this project, the learning algorithm predicts the rating of a movie which would be given by a user and recommends the movie to the user.

Collaborative filtering with Pearson Correlation co-efficient was chosen to predict the user’s movie rating on the test data set. Pearson Co-efficient computes the weights (similarity co-efficients) between 2 users based on the common movies over which the 2 users have voted.

Training Set: Training.txt file which contains about 3.2 million records of user votes of the format (Movie ID, User ID, Vote)

Test Set: Testing.txt file which contains about 100000 records of user entries with movies for which the vote prediction has to be made. The test set also contain the actual vote that the user. The format is (Movie ID, User ID, Vote)

The Collaborative Filtering is implemented using Python. The projects uses certain external packages to implement the algorithm.

The external Packages are:

1. Latest Python Version (python 3.3)

2. Numpy

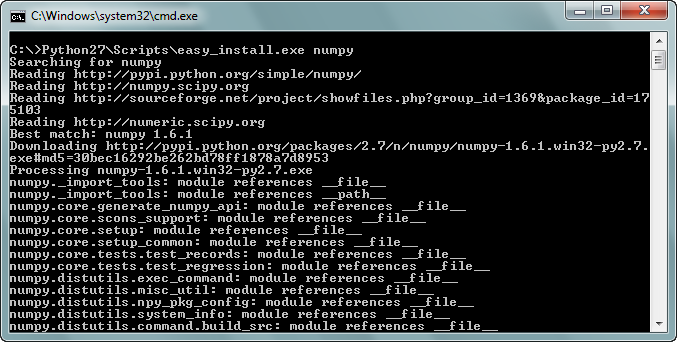
3. Pandas

The projects needs the external packages to be installed before executing the program.

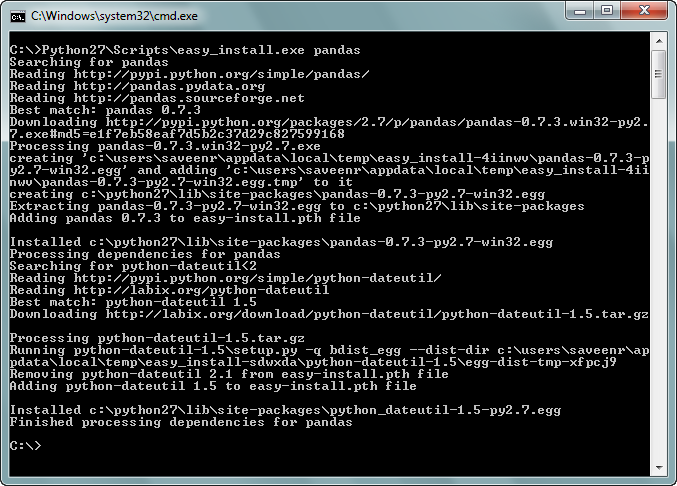
Steps to install the Install the packages:

The setup.py file is attached with this project.rar file. Double-click the file and it installs certain tools in python directory.

After that, in command prompt, change the directory to C:\Python33\Tools\Scripts and enter "easy\_install.exe numpy". This will install numpy package on the system



After that, enter "easy\_install.exe pandas" to install pandas.



The project - "SVM\_Perceptron" contains following files:

1. prob4data - Directory for training and testing files.

2. Collab\_Filter.py - Implementation of Collaborative Filtering

3. SVM\_Perceptron.py - Implementation of Perceptron (Part1 of Assignment 4)

4. training.train - Training file for SVM\_Perceptron.py file (Part 1 of Assignment 4)

5.validation.train - Test File for SVM\_Perceptron.py fiel (Part2 of Assignment 4)

**Only the " Collab\_Filter.py" should be run for Collaborative Filtering Project.**

The testing and training files for collaborative filtering should be place be in **prob4data** folder.

**Structure of the Program:**

1. main() - Reads the testing and training files using pandas - data analytics tool.

2.Data\_Matrix() - Predicts the movie for the given user.

3.indentify\_error\_rate() - identifies the absolute mean error and root mean square error.

Since this collaborative algorithm is a memory-based one, therefore the calculation of training and prediction are data intensive operations, considering the volume of the training and testing Data sets(3.2 million and 100,000 records  respectively). The computation of correlation coefficient for each user entry in the test set with corresponding entries in the training set takes significant amount of time.

Each such weight calculation and prediction takes about 200 milliseconds approx and the prediction for the entire dataset of 100000 records takes upto 10 hours with the current implementation.

**The program requires an execution time of less than 5-10 hours to predict the ratings for movie and printing the final absolute mean error and root mean square error in console.**

**Output of the program:**

The Mean Absolute Error - 0.67450696

The Root Mean Square Error - 0.789876

To assure that the program runs perfectly, a sample data is attached with project file. The sample data contains files:

1.TrainingRatings-Sample.txt - Set of 4 users and 5 movies rated by each user

2.TestingRatings-Sample.txt - Set of 5 users with movie id, user id and ratings