# 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

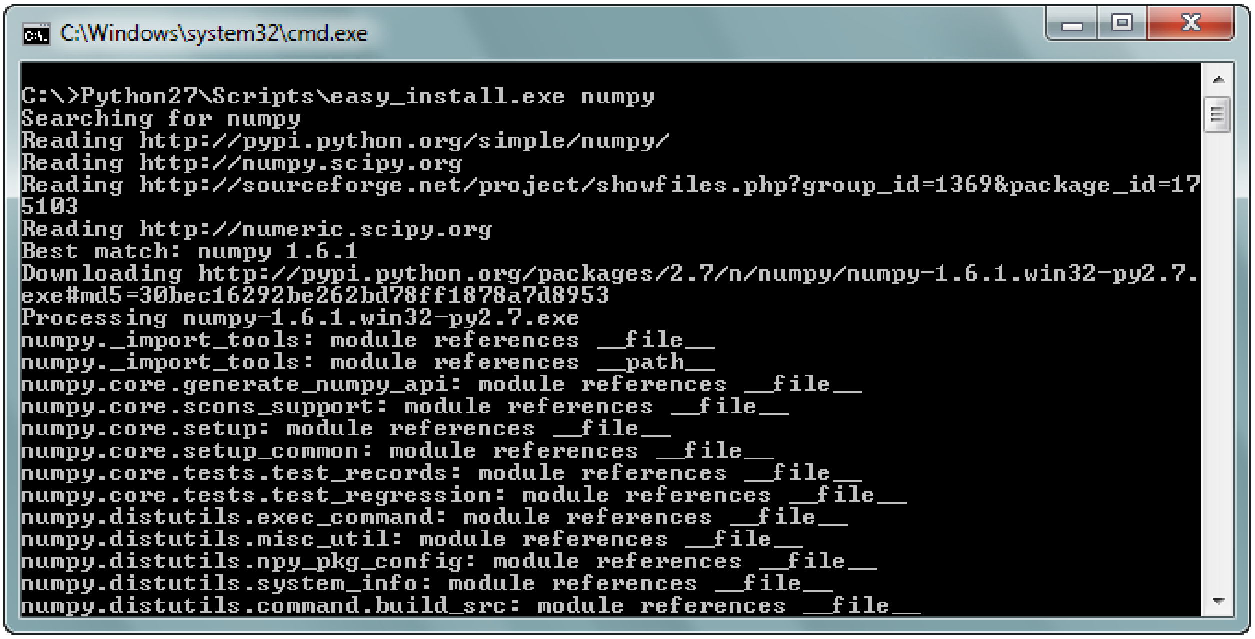
3. Pandas

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

Steps to install the Install the packages:

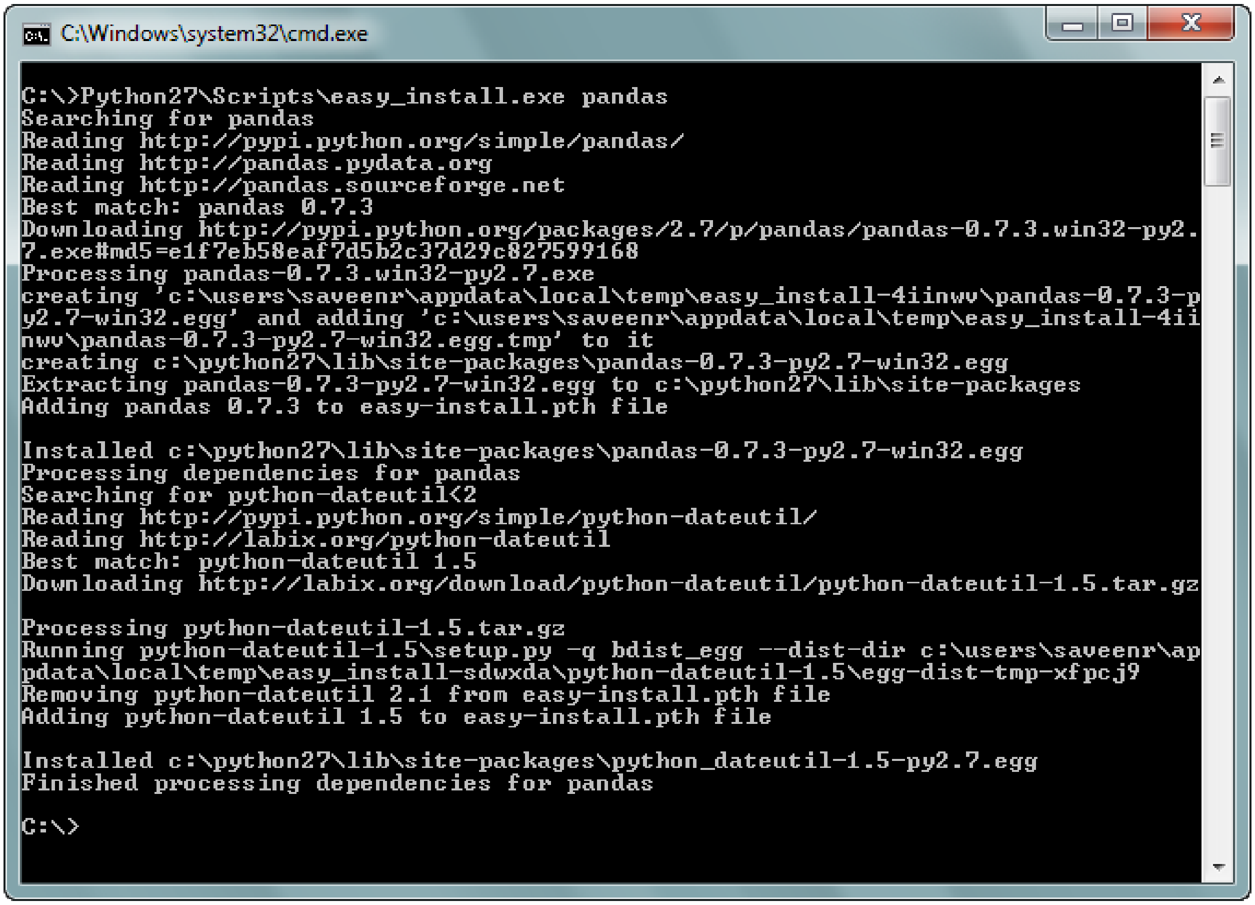
The setup.py file is attached with this project.rar file. Double-click the file and it installs certain tools in python directory.

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



```
C:\>Python27\Scripts\easy_install.exe numpy
Searching for numpy
Reading http://pypi.python.org/simple/numpy/
Reading http://numpy.scipy.org
Reading http://sourceforge.net/project/showfiles.php?group_id=1369&package_id=17
5103
Reading http://numeric.scipy.org
Best match: numpy 1.6.1
Downloading http://pypi.python.org/packages/2.7/n/numpy/numpy-1.6.1.win32-py2.7.
exe#md5=30bec16292be262bd78ff1878a7d8953
Processing numpy-1.6.1.win32-py2.7.exe
numpy._import_tools: module references __file__
numpy._import_tools: module references __path__
numpy.core.generate_numpy_api: module references __file__
numpy.core.scons_support: module references __file__
numpy.core.setup: module references __file__
numpy.core.setup_common: module references __file__
numpy.core.tests.test_records: module references __file__
numpy.core.tests.test_regression: module references __file__
numpy.distutils.exec_command: module references __file__
numpy.distutils.misc_util: module references __file__
numpy.distutils.npy_pkg_config: module references __file__
numpy.distutils.system_info: module references __file__
numpy.distutils.command.build_src: module references __file__
```

After that, enter "easy_install.exe pandas" to install pandas.



```
C:\>Python27\Scripts\easy_install.exe pandas
Searching for pandas
Reading http://pypi.python.org/simple/pandas/
Reading http://pandas.pydata.org
Reading http://pandas.sourceforge.net
Best match: pandas 0.7.3
Downloading http://pypi.python.org/packages/2.7/p/pandas/pandas-0.7.3.win32-py2.
7.exe#md5=e1f7eb58eaf7d5b2c37d29c827599168
Processing pandas-0.7.3.win32-py2.7.exe
creating 'c:\users\saveenr\appdata\local\temp\easy_install-4iinwv\pandas-0.7.3-p
y2.7-win32.egg' and adding 'c:\users\saveenr\appdata\local\temp\easy_install-4ii
nwv\pandas-0.7.3-py2.7-win32.egg.tmp' to it
creating c:\python27\lib\site-packages\pandas-0.7.3-py2.7-win32.egg
Extracting pandas-0.7.3-py2.7-win32.egg to c:\python27\lib\site-packages
Adding pandas 0.7.3 to easy-install.pth file

Installed c:\python27\lib\site-packages\pandas-0.7.3-py2.7-win32.egg
Processing dependencies for pandas
Searching for python-dateutil<2
Reading http://pypi.python.org/simple/python-dateutil/
Reading http://labix.org/python-dateutil
Best match: python-dateutil 1.5
Downloading http://labix.org/download/python-dateutil/python-dateutil-1.5.tar.gz

Processing python-dateutil-1.5.tar.gz
Running python-dateutil-1.5\setup.py -q bdist_egg --dist-dir c:\users\saveenr\ap
pdata\local\temp\easy_install-sdwxda\python-dateutil-1.5\egg-dist-tmp-xfpcj9
Removing python-dateutil 2.1 from easy-install.pth file
Adding python-dateutil 1.5 to easy-install.pth file

Installed c:\python27\lib\site-packages\python_dateutil-1.5-py2.7.egg
Finished processing dependencies for pandas

C:\>
```

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