Description Assignment 3: LSH and Collaborative Filtering

1. Spark Version: 2.2.1 & Python Version: 2.7

I) Task1: Create a new folder and add the following files Anmol_Chawla_task1_Jaccard.py, ratings.csv

Command line input: Anmol_Chawla_task1_Jaccard.py ratings.csv

ii) Task 2:

1)Part1:

a) Create a new folder and add the following files Anmol_Chawla_task2_UserBasedCF.py, ratings.csv and testing_small.csv

b) Open the command prompt and write the following command: spark-submit Anmol_Chawla_task2_ModelBasedCF.py ratings.csv testing_small.csv

c) A file called Anmol_Chawla_task2_ModelBasedCF.txt will be created in the same folder.

2) Part 2:

a) Create a new folder and add the following files Anmol_Chawla_task2_UserBasedCF.py, ratings.csv and testing_small.csv

b) Open the command prompt and write the following command: spark-submit Anmol_Chawla_task2_UserBasedCF.py ratings.csv testing_small.csv

c) A file called Anmol_Chawla_UserBasedCF.txt will be created in the same folder.

2. Jaccard LSH Task 1
   I) Precession: 1.0
   II) Recall: 0.903714673128

3. Base Line Table

	Task 1		Task 2
	Small	Large	Small
>=0 and <1	13109	3242200	14961
>=1 and <2	4183	700566	4424
>=2 and <3	1066	90506	721
>=3 and <4	293	12001	139
>=4	82		11
RMSE	1.12398211617	0.819062412312	0.954352976157

4. Improvements:

   Pearson co-relation was used to calculate, so was jaccard similarity