Vijayakumar Gedigeri INF553 HW4, April 5, 2017

This assignment is written in Python 2.7 for Spark 2.0.1. Below are details on how to run the code, answers to questions.

1. Command to run the code

   ```
   .\bin\spark-submit --master local[*] --driver-memory 5g
   Vijayakumar_Gedigeri_LSH.py ratings.csv
   C:\Users\Output\Vijayakumar_Gedigeri_SimilarMovies.txt
   ```

   Output path in above command is given for Windows OS. In case, you want to run on Linux machine, instead of front slash '\', use back slash '/' for absolute path.

   Above will create folder **Output** in C:\Users and then the file Vijayakumar_Gedigeri_SimilarMovies.txt under Output folder. You can specify any path.

   Sample Vijayakumar_Gedigeri_SimilarMovies.txt output file is available under OutputFiles directory.

2. Problem 2 Solution

   Precision and Recall of my implementation were calculated comparing with Ground Truth file that has 384233 similar movies with jaccard similarity > 0.5

   Number of True positives (tp)= 337817
   Number of False positives (fp) = 0
   Number of False Negatives (fn) = 46416

Precision = tp / (tp + fp) = 337817/ (337817+ 0) = 1.0
Recall = tp / (tp + fn) = 337817/ (337817+ 46416) = 0.8791

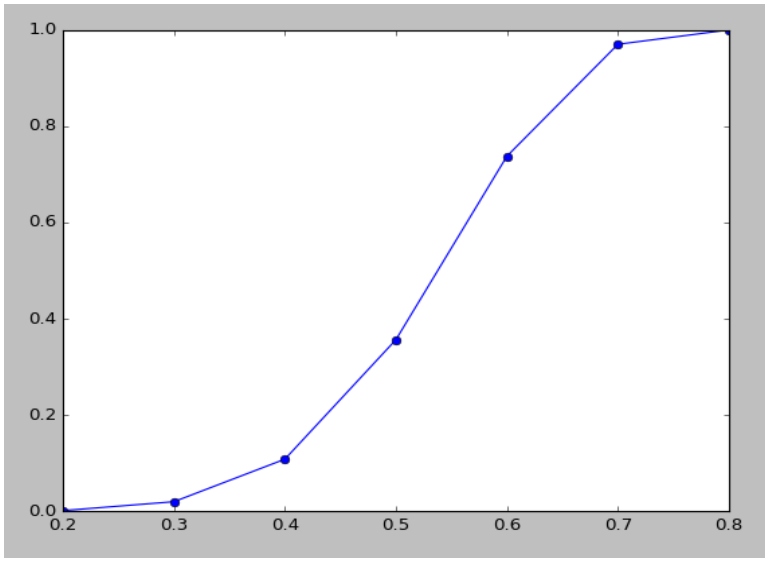
3. Table showing the values for S-Curve for the below values of b and r used in my implementation

   # of bands = **b** = 28
   # of rows per band = **r** = 6
   # of signatures = 168

S	**$1-(1-s^r)^b$**
0.2	0.00179
0.3	0.02021
0.4	0.10856
0.5	0.35657
0.6	0.73758
0.7	0.96994
0.8	0.99979

4. How does the value of *b* and *r* affect the value of the threshold for the movies dataset?

   Answer:



S-curve for above table.

In above graph, we can see that the curve hasn't risen significantly until the threshold of slightly more than s = 0.5 is reached (Around s = 0.55 to be exact). Slope in the middle is significant and raises by more than 0.6 when going from threshold 0.4 to 0.6 or 0.5 to 0.7.

For a threshold of 0.8, 1- $(0.8)^6$ = 0.7378 where r = 6. When we raise this by 28 (bands), we get 0.000201. Subtracting this from 1 gives 0.99979. This tells that for a movie pair from movies dataset with jaccard similarity of 0.8, there is 27% chance of agreeing on all 6 rows in any band to become a candidate pair. As there are 28 bands, there are 28 chances of becoming a candidate pair.

If we consider there are 100,000 movie pairs with jaccard similarity 0.8, then as less as 20 will fail to become candidates and hence false negative.