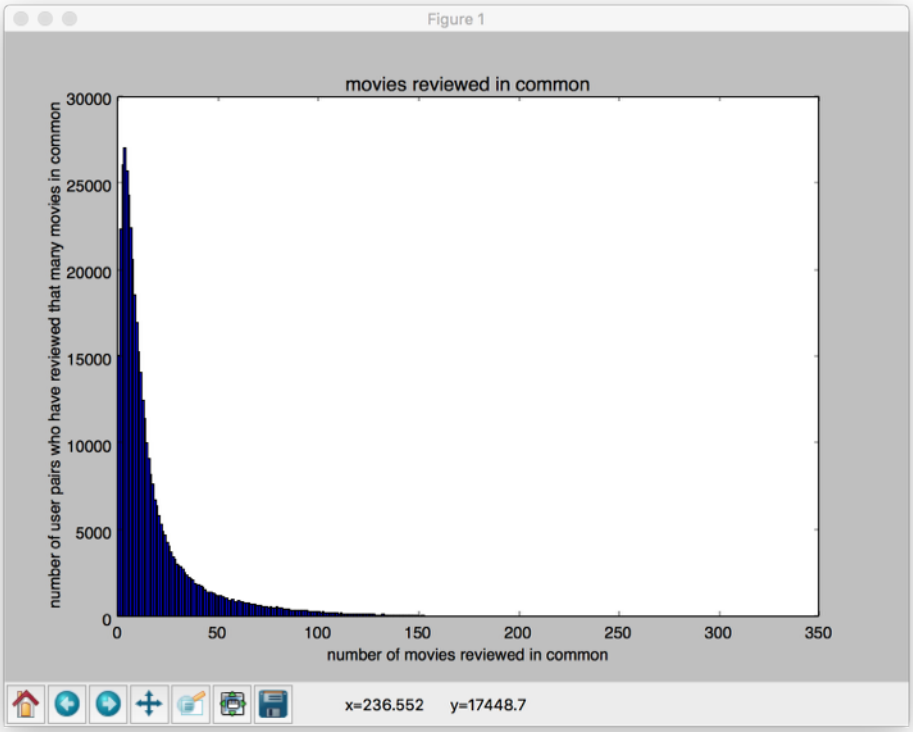
EECS 349 (Machine Learning) Homework 4

Xinyi Chen

Problem 1

A)

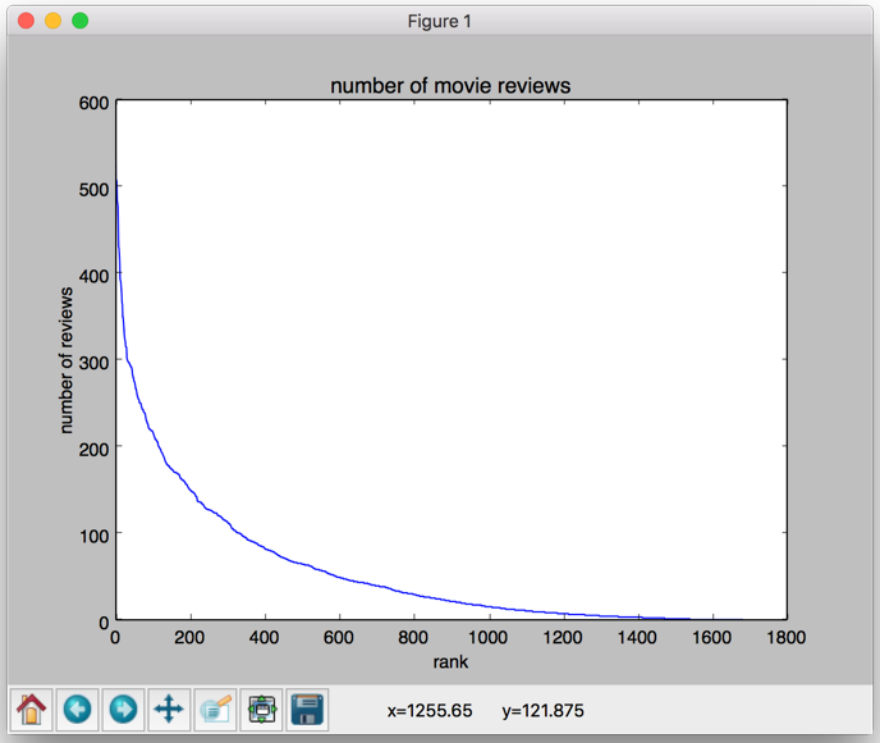
The mean number is 18.81, median number is 10. I choose pot.bar() to generate this histogram, and since when number of movies reviewed in common is greater than 350, the number of pairs all are 0, so my histogram didn't draw the part after x=350.(Run pairsUserMovieRate_Problem1.py to see)



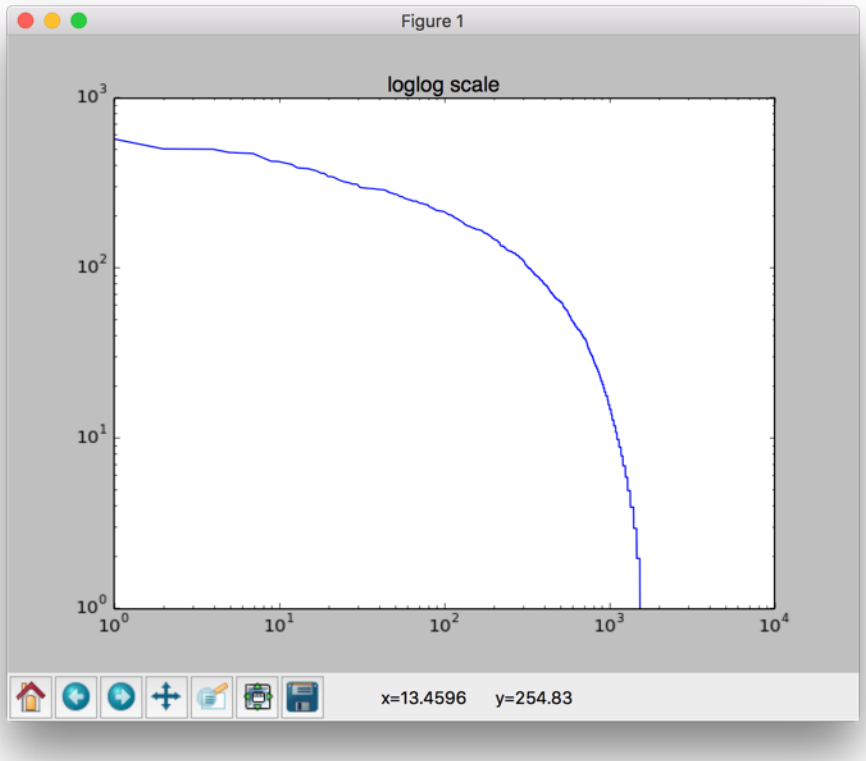
B)

Movie 50 has the most reviews(583 reviews), the minimum review number is 1,and many movies has only 1 review:1653, 1655, 1621, 1627, 1626, 1625, 1624, 1579, 1343, 1507, 1505, 1618, 1619, 1616, 1614, 1613, 1363, 1543, 1548, 1601, 1603, 1604, 1606, 1156, 1557, 1525, 1526, 814, 1492, 1493, 1494, 1498, 1122, 1235, 1236, 1482, 1486, 1130, 1572, 1571, 1570, 1577, 1576, 1575, 1574, 830, 1582, 1583, 1580, 1581, 1586, 1373, 1584, 1457, 1458, 1587, 1595, 1596, 1593, 1599, 1447, 1452, 1453, 857, 852, 1476, 1309, 1559, 1681, 1680, 1682, 1461, 1460, 1310, 1533, 1536, 1678, 1679, 1674, 1675, 1676, 1677, 1670, 1671, 1673, 677, 1414, 1320, 1325, 1329, 1520, 711, 1669, 1668, 1667, 1666, 1665, 1663, 1661, 1660, 599, 1546, 1339, 1352, 1659, 1650, 1651, 1657, 1654, 1633, 1349, 1348, 1340,

1341, 1645, 1647, 1641, 1640, 1649, 1648, 1515, 1510, 1563, 1564, 1566, 1567, 1630, 1632, 1634, 1635, 1636, 1637, 1638, 1366, 1568, 1569, 1561, 1562, 1565, 1201, 1364.(run pairsUserMovieRate_Problem1.py to see)



I don't think the number of reviews per movie follows Zipf's law, because as we can see its log-log scale is not a line.



Problem 2

A)

I think approach B is better, because average chosen by user's movie rating he/she did rate can reveal the user's behaviour in some way(how he/she tend to rate a movie), but put 0 for every missing data has nothing to do with a specific user's feature.

	Movie 1	Movie 2	Movie 3	Movie 4	Movie 5
A	2	3	1	2	?
B	2	4	1	2	4
C	3	3	2	3	1

Toy Example: As we can see in the example, for movie 1 to 4, A and B are almost the same, only in movie 2 A and B have 1 difference, so B should be calculated closer to A than C. If we set A movie 5 = 2(2 is average), Manhattan Distance(A,B) = 3, MD(A,C) = 4, shows that A is closer to B than C, which is reasonable. But if we set A movie 5 = 0, MD(A,B) = 5, MD(A,C) = 4, shows that A is closer to C than B, this isn't reasonable. So average is better.

B)

I think Pearson 's correlation is better than Euclidean distance. Because Pearson' correlation can calculate the rating trend(goes up or goes down), which is very important to the difference calculation of item based prediction, because items have the same trend should be thought closer. Euclidean distance only calculates the distance between two ratings, but didn't consider the trend.

	User 1	User 2	User 3	User 4
Movie A	1	2	3	4
Movie B	2	3	4	5
Movie C	2	2	2	3

Toy example: As we can see in the table, B should be considered closer to A than C because A and B have the same trend and each rating is very close, C's trend is totally different with A. Pearson(A,B) = 1, Pearson(A,C) = 0.77, which shows B is closer to A than C, it is reasonable. However Euclidean Distance(A,B) = 2, ED(A,C) = 3^(1/2), which shows that C is closer to A than B, it isn't reasonable. So Pearson's correlation is better.

Problem 3

python *user_cf.py* <datafile> <userID> <movieID> <distance> <k> <i>

python *item_cf.py* <datafile> <userID> <movieID> <distance> <k> <i>

Problem 4

A)

For each prediction, Error Measure = (predicted rating – true rating)^2. So 0 is the best, it means the prediction is the same as true rating. The larger error measure, the worse the prediction is. I choose this because it can measure both right/wrong and how wrong the prediction is.

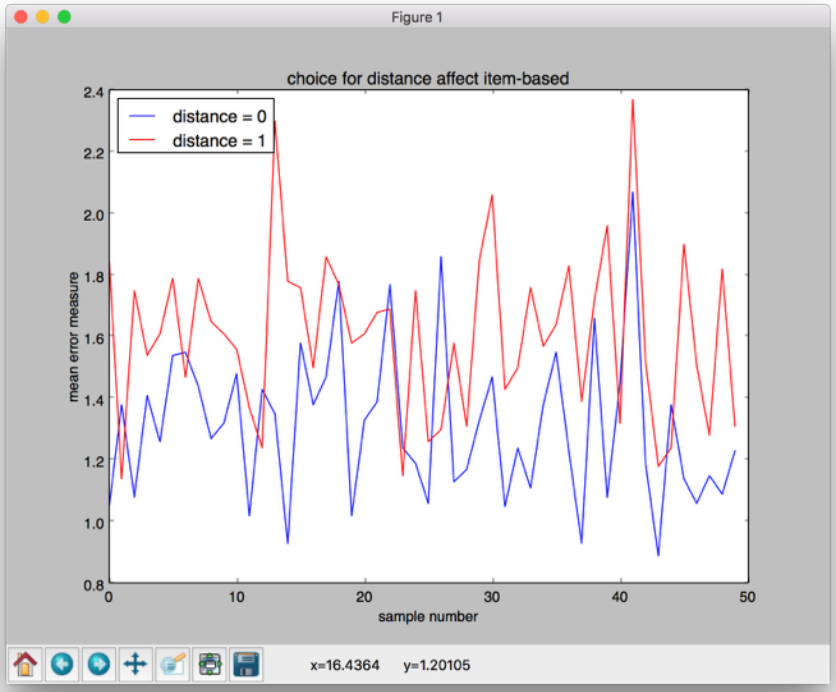
For question B, C, D, E and F, I'm going to use each samples' mean error measure to evaluate every filter. E.g, there are 50 samples, 100 data in each sample. To compare filter A with filter B, I'm going to calculate each samples' mean error measure(for each sample, run its 100 data points and get 100 error measure's mean value), so I have 50 mean error measure for each filter, then make a graph to show these two sets of mean error measure, the lower line(smaller mean error measure), the better the filter is.

B)

I choose paired samples t-test, because variants are independent and identically distributed(chose from the same data set independently and randomly), paired(change parameters to generate paired response on same data) and variants have same variance(use the same way to calculate) and also fit Gaussian distribution(can tell in distribution graph). The null hypothesis: the difference between two system variants are not statistically significant. I choose p = 0.05, because it is convenient to take this point as a limit in judging whether a deviation ought to be considered significant or not.

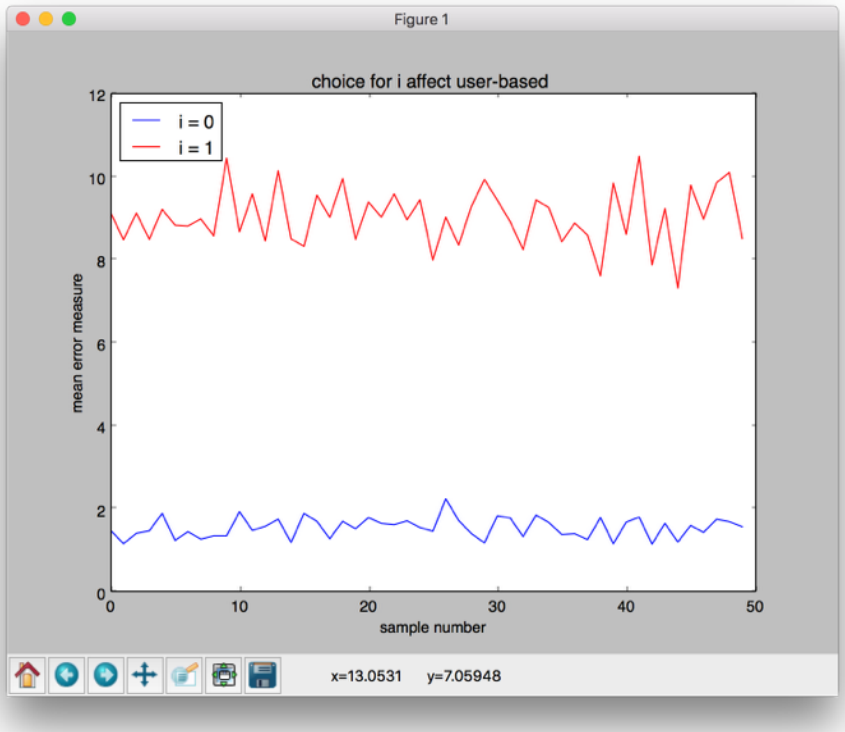
C)

As we can see in the graph, when distance = 0(Pearson's correlation) the mean measure error is smaller than distance = 1(Manhattan distance) (other parameters are the same, i = 0, k = 5), and by using paired samples t-test, p = 2.87 *10^(-8), it is lower than 0.05, so the null hypothesis is not true, the difference between two system variants are statistically significant. So Pearson's correlation is better, and this result is the same as the intuition I developed in problem 2.B.



D)

As we can see in the graph, when i = 0(only use users that have actual ratings) the mean measure error is smaller than i = 1(regardless of whether actual or filled-in ratings)(other parameters are the same, distance = 0, k = 5), and by using paired samples t-test, p = 7.32 *10^(-53), it is lower than 0.05, so not as what I have expected, the null hypothesis is not true, the difference between two system variants are statistically significant, so i = 0 is better.



E)

As we can see in the graph, when k = 32 the mean measure error is the smallest, (other parameters are the same, distance = 0, i = 0), and by using paired samples t-test, all the p value calculated by using other k and 32 is lower than 0.05, so the null hypothesis is not true, the difference between k = 32 and all other system variants are statistically significant, so k = 32 is the best.

p values:

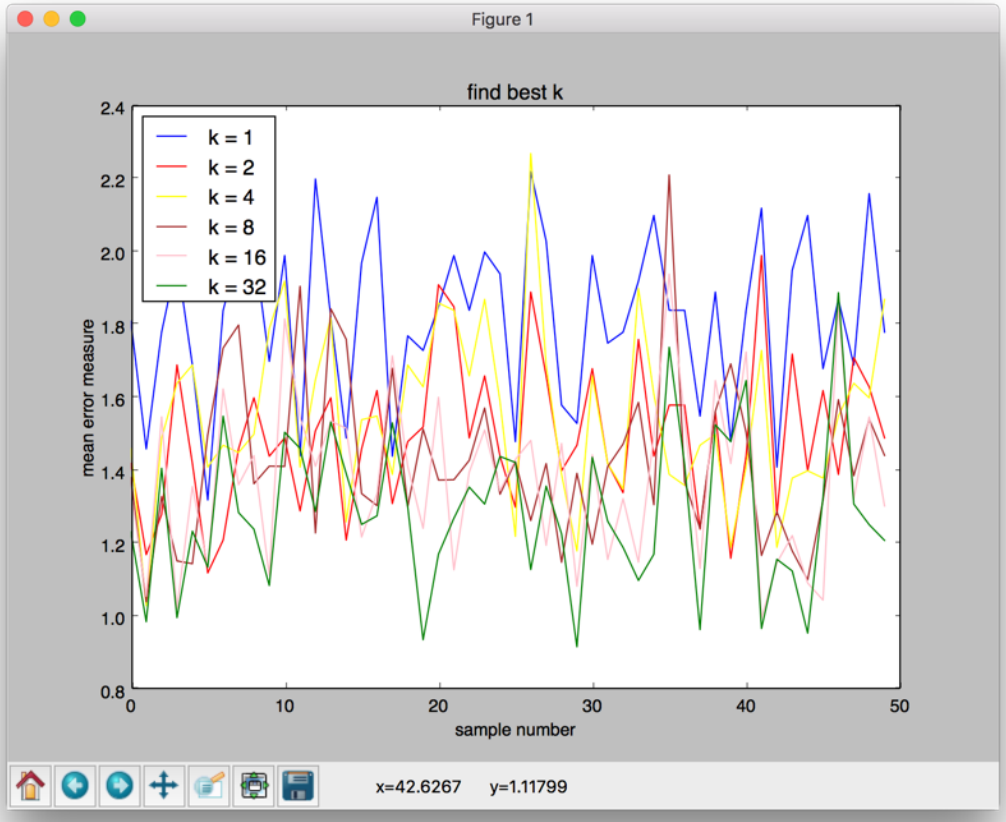
p(1 , 32) = 4.13832048375 * 10^(-16)

p(2 , 32) = 1.93790181435 * 10^(-5)

p(4 , 32) = 6.07108989611 * 10^(-7)

p(8 , 32) = 5.20691623022 * 10^(-6)

p(16 , 32) = 2.03242006114 * 10^(-5)



F)

As we can see in the graph, the mean measure error of item based filter is the smaller than user-based filter(all the parameters are the same, distance = 0, k = 32, i = 0), and by using paired samples t-test, the p value(confidence) = 2.27*10^(-7), it is lower than 0.05, so the null hypothesis is not true, the difference between item based filter and user-based filter variants are statistically significant, so item-based filter is better than based filter.