EE219 Project 3

COLLABORATIVE FILTERING

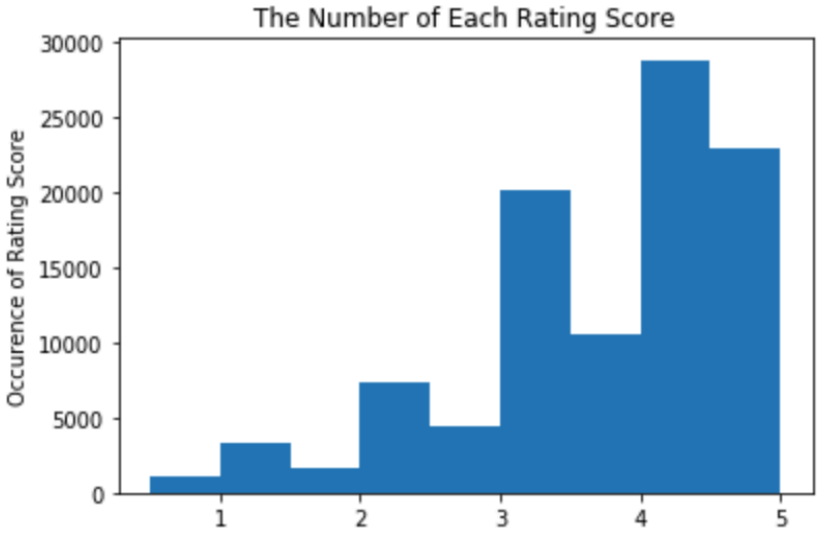
Winter 2018

Di Ma 004945175
Weijie Tang 305029285
Wandi Cui 905024671
Yuanzhi Gao 704145326

Movie Dataset

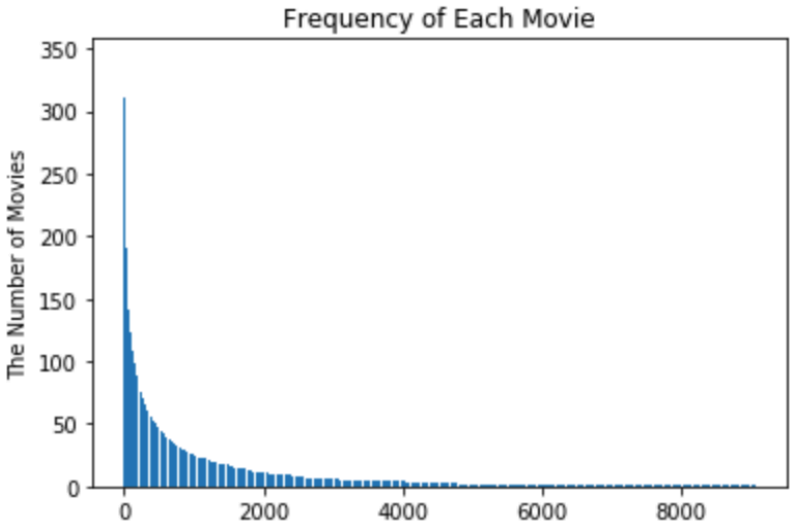
Q1: Sparsity = 0.016439141608663475

Q2:

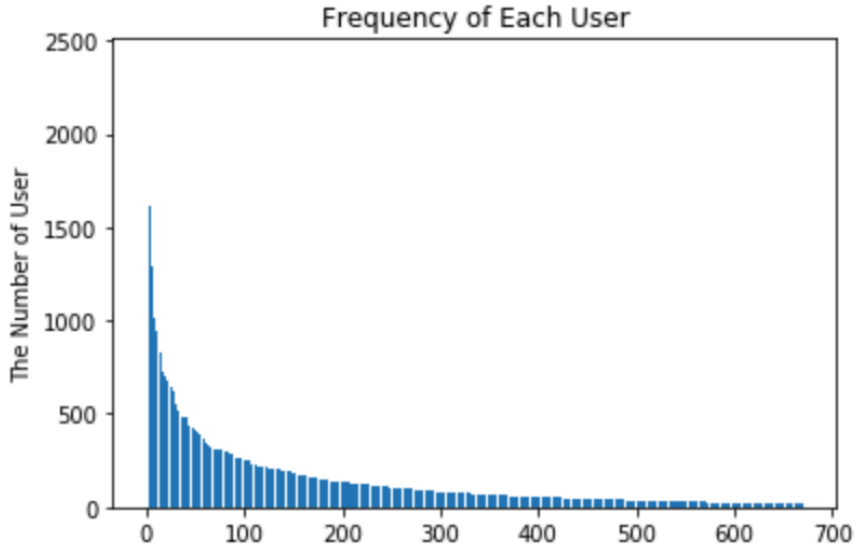


From the histogram, we can see that most of the rating scores gather around 3-5, which means that users tend to give high ratings than low ratings. This could be an indicator for us to adjust the threshold for like/unlike movies for later experiments.

Q3:

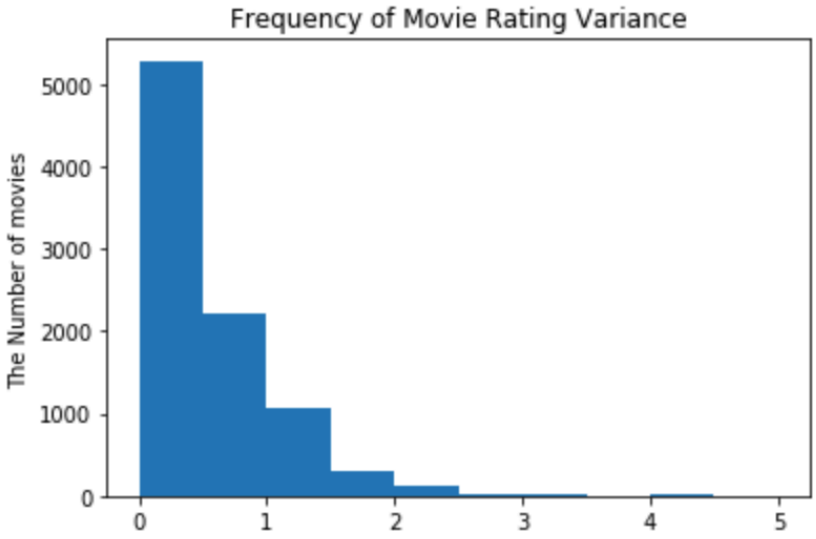


Q4:



Q5: The plot in Q3 illustrates that only a few movies received ratings, so the rating matrix R will be sparse, which will make the model time- consuming and influence the accuracy of recommendation results.

Q6:



From the histogram, we can see that most of the movies are pretty consistent regarding the ratings they receive.

Neighborhood-based Collaborative Filtering

Q7:

$$u_u = \sum_{k=I0}^{Iu} r_{uk} / |I_u|$$

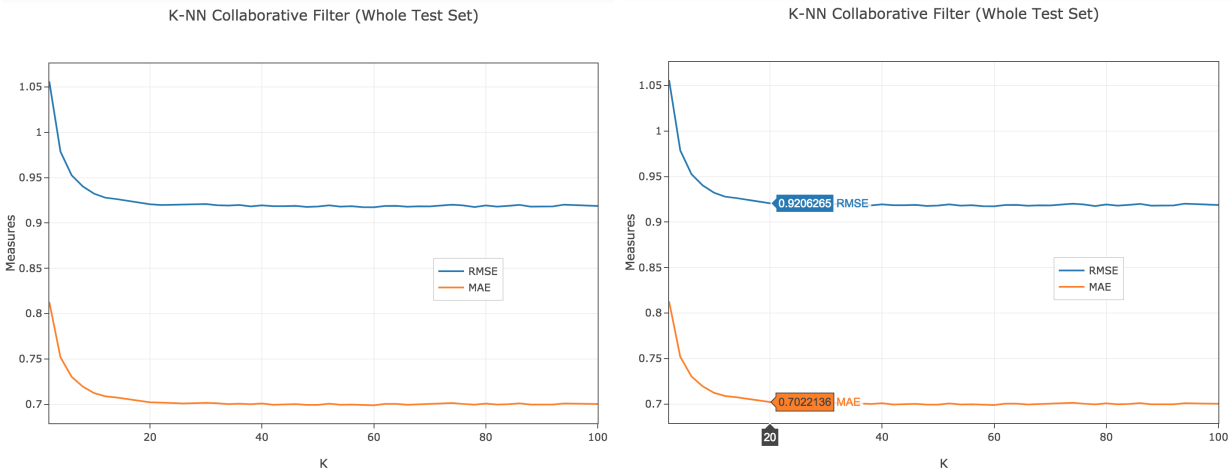
Q8:

means the set of item indices for which ratings have been specified by both user and , in other words, the movies that both users have rated.The intersection could be an empty set, which means the movies rated by two users are totally different. Considering the fact that the rating matrix is sparse, an empty intersection is more likely than usual.

Q9:

The reason for mean-centering is to handle the problem that not all users have the same rating preferences. For instance, users and both rate a movie for 3 stars, but we could not jump to the conclusion that they have same feeling about this movie without other information. If user tends to give 2 stars while user tends to give 4 stars, then it is highly likely that user has better comments on the movie than user . As a result, using mean-centering avoid that bias and only consider the relative rates given by a certain user.

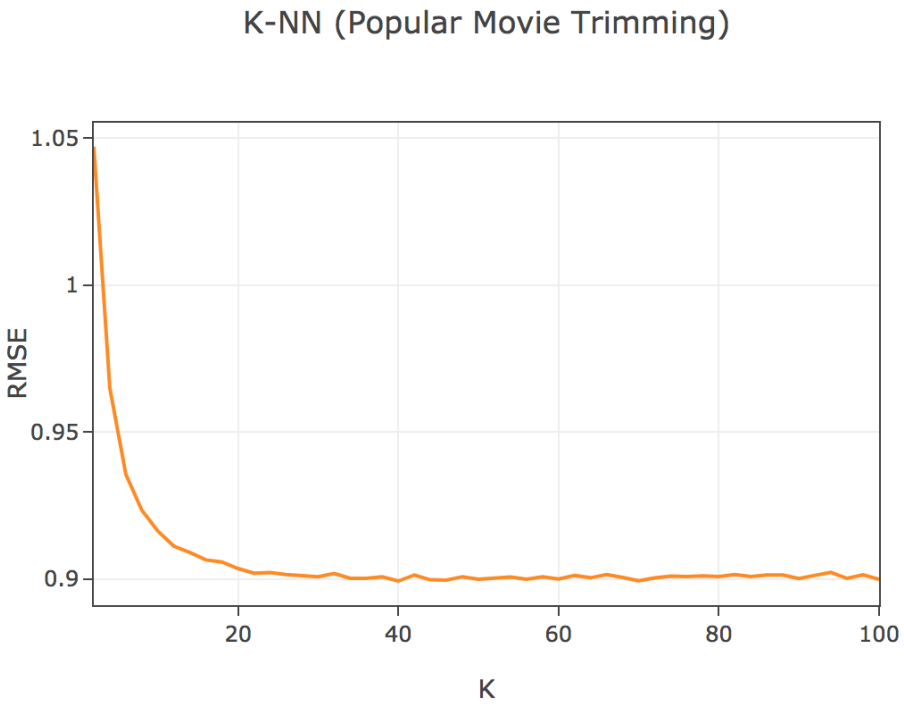
Q10:



Q11:

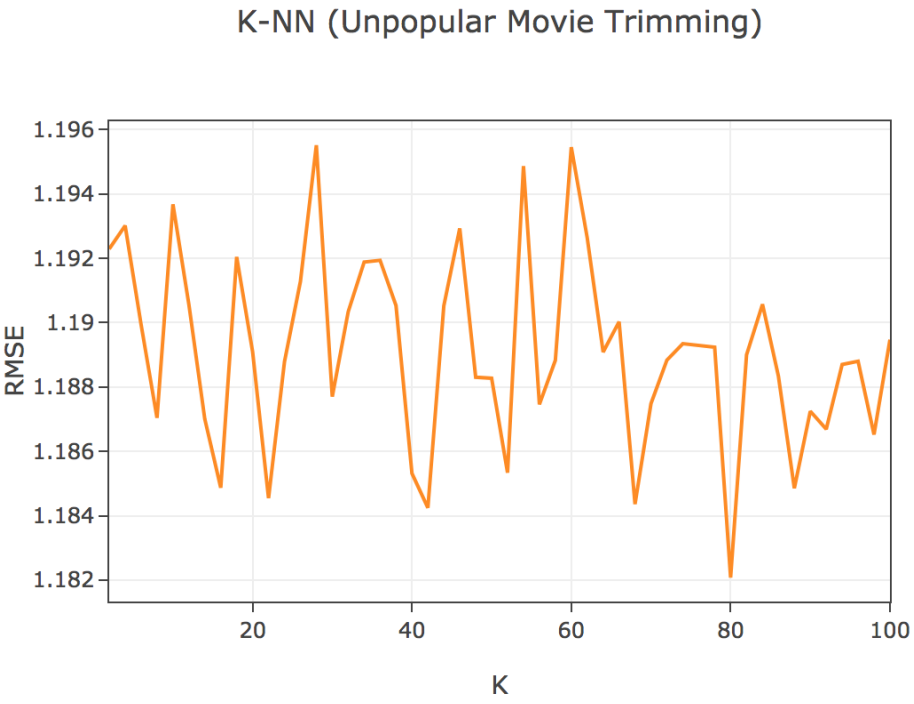
Minimum (RMSE = 0.9206, MAE = 0.7022)

Q12:



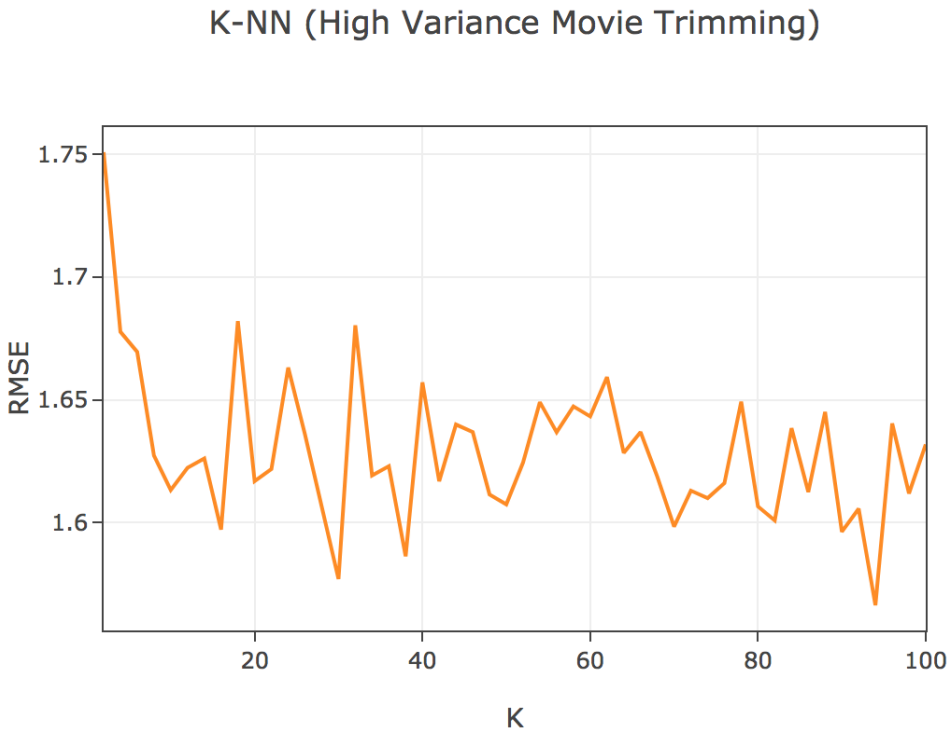
Minimum average RMSE: 0.8994

Q13:



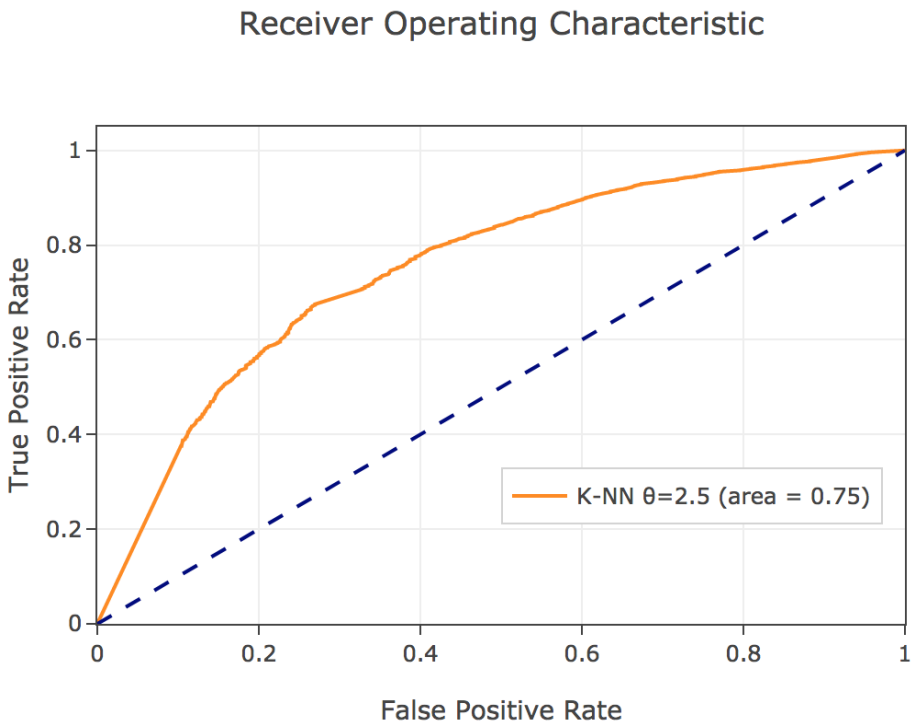
Minimum average RMSE: 1.1821

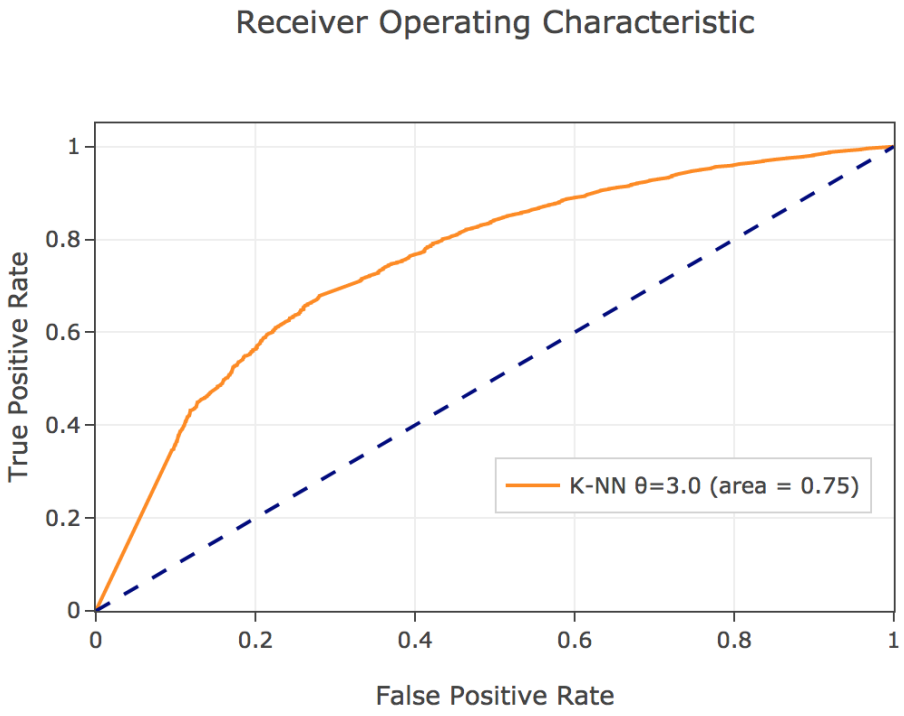
Q14:

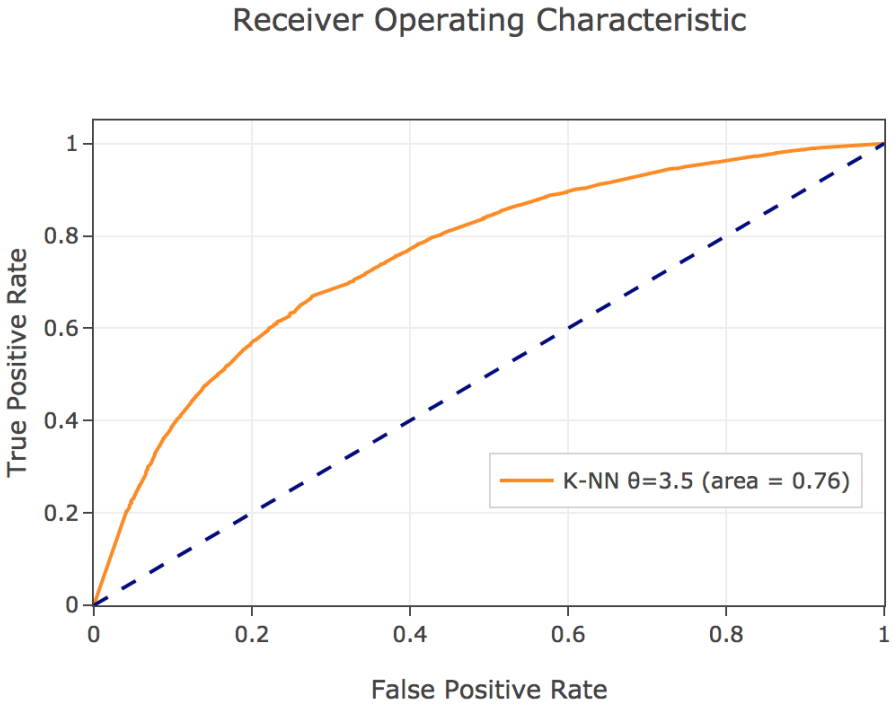


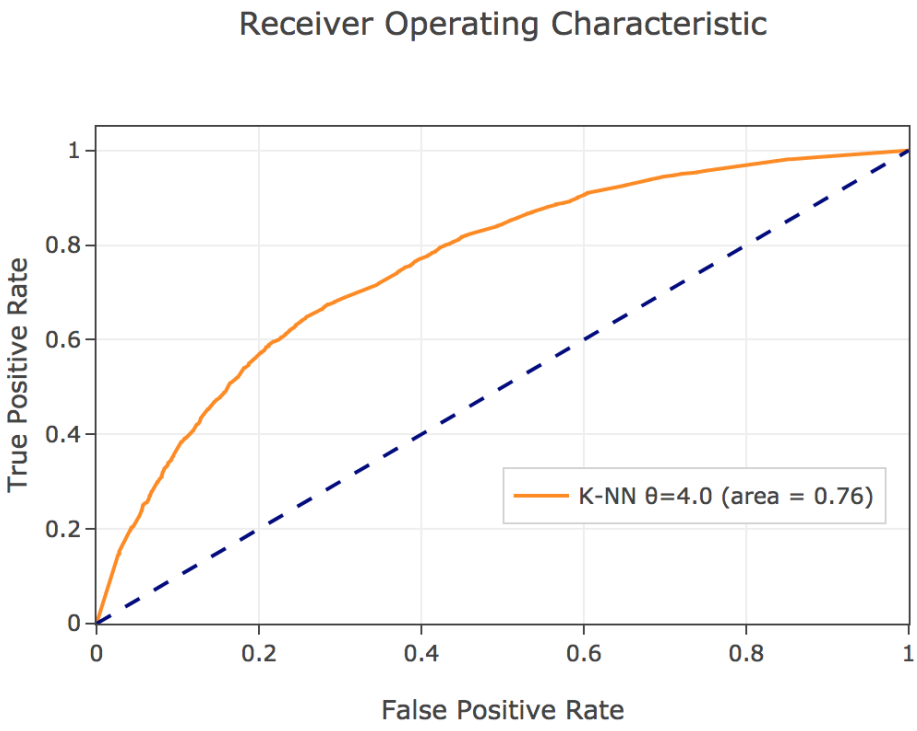
Minimum average RMSE: 1.5664

Q15:




Receiver Operating Characteristic
True Positive Rate
False Positive Rate
K-NN θ=3.0 (area = 0.75)
0
0.2
0.4
0.6
0.8
1


Receiver Operating Characteristic
True Positive Rate
False Positive Rate
K-NN θ=3.5 (area = 0.76)
0
0.2
0.4
0.6
0.8
1



Non-negative matrix factorization (NNMF)

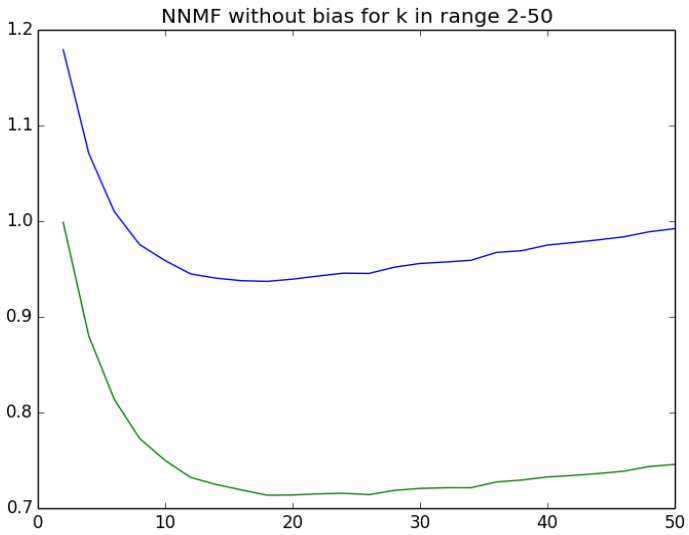
Q16: Is the optimization problem given by equation 5 convex? Consider the optimization problem given by equation 5. For U fixed, formulate it as a least-squares problem.

No, the optimization problem given by equation 5 is NOT convex, because we are aiming to learn two types of variables, variables of U and variables of V, and since both U's and V's values are unknown variables makes this cost function non-convex. However if we fix U and look for V and vice versa, we will get a convex optimization problem. It is a two-step iterative optimization process. In every iteration it first fixes V and solves for U, and following that it fixes U and solves for V. Least square problem is guaranteed to converge only to a local minima, and is ultimately depends on initial values for U or V.

$$\forall U_i = \sum_{i=1}^{m} \sum_{j=1}^{n} W_{ij} \left(r_{ij} - U_i V^T \right)^2$$

$$\forall V_j = \sum_{i=1}^{m} \sum_{j=1}^{n} W_{ij} \left(r_{ij} - U \cdot V_j^T \right)^2$$

Q17:



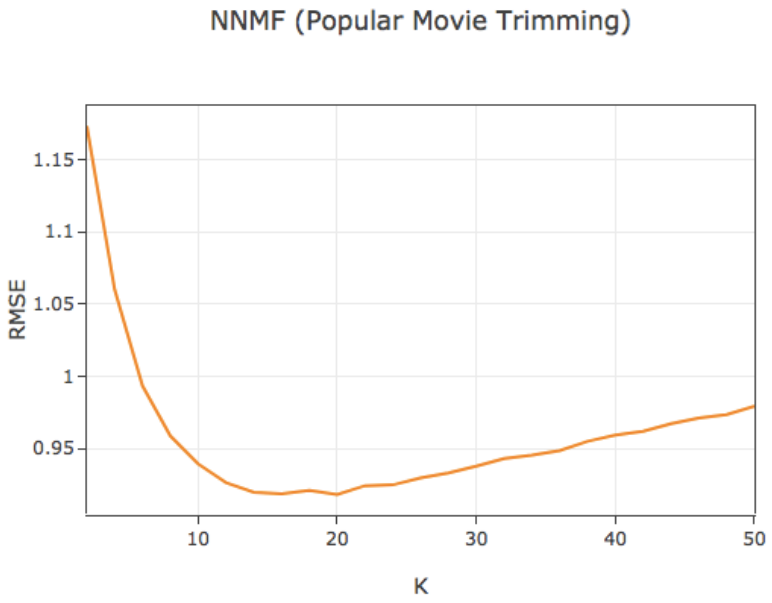
The minimum average RMSE is 0.9371590551266061 at k = 18
The minimum average MAE is 0.713447250506337 at k = 20

Q18:
The best latent factor we get from the plot of question 17 is 18 or 20 for best MSE or RMSE.
Genres are in total 18 different genres as below:

- Action
- Adventure
- Animation
- Children's
- Comedy
- Crime
- Documentary
- Drama
- Fantasy
- Film-Noir
- Horror
- Musical
- Mystery
- Romance
- Sci-Fi
- Thriller
- War
- Western

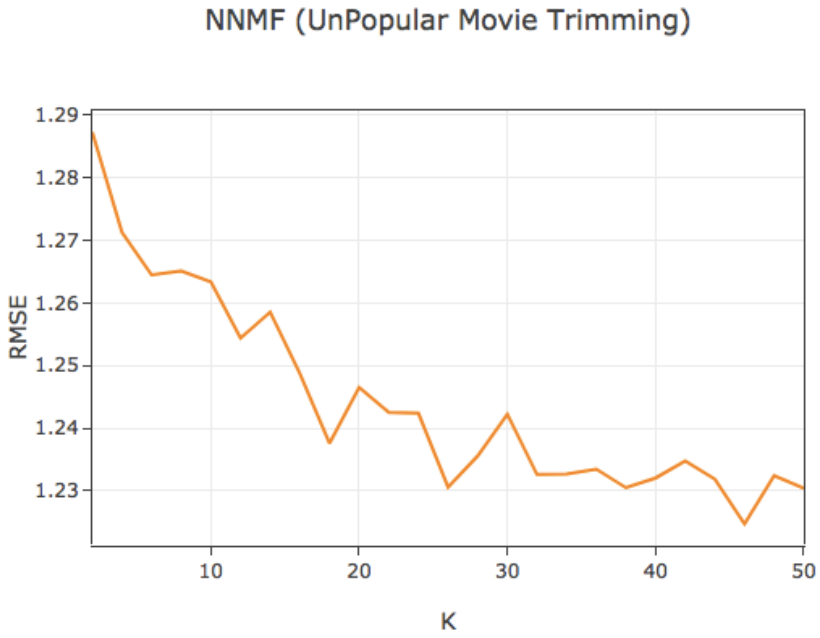
So if we use RMSE to decide the optimal latent factor, then YES it is the same as the movies genres amount.

Q19 Trimmed set for NNMF Popular:



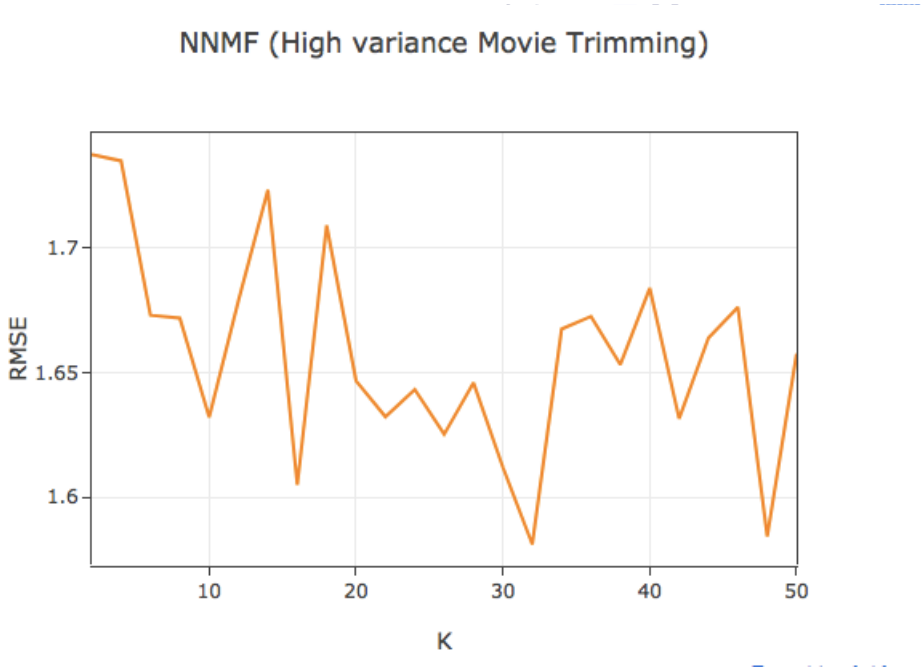
Minimum RMSE: 0.9182

Q20 Trimmed set for NNMF UnPopular:



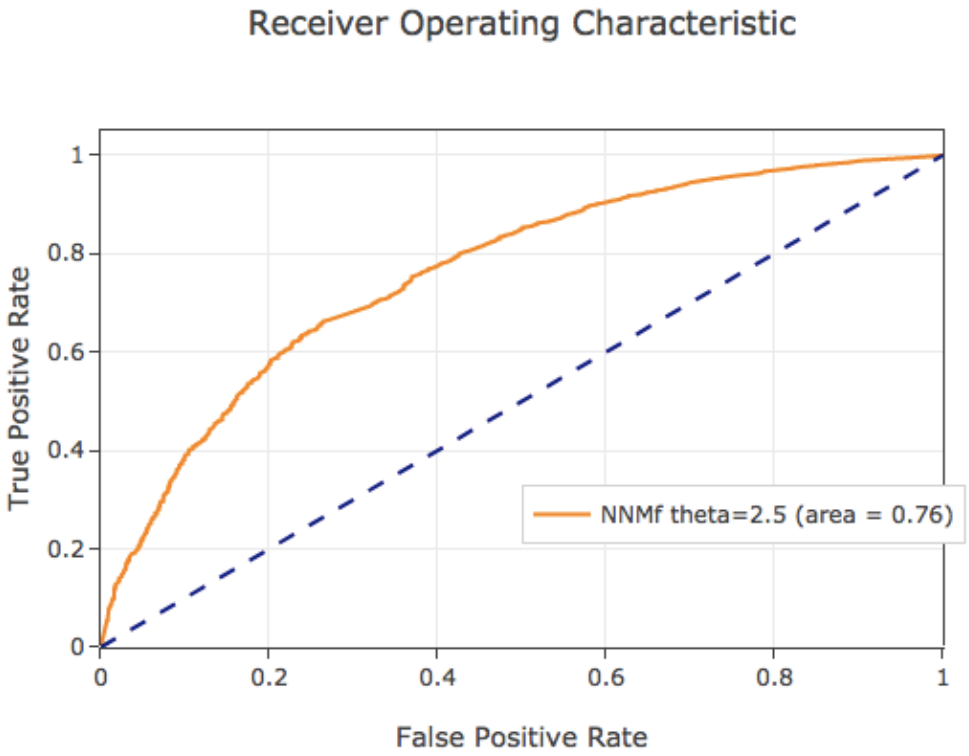
Minimum RMSE: 1.2247

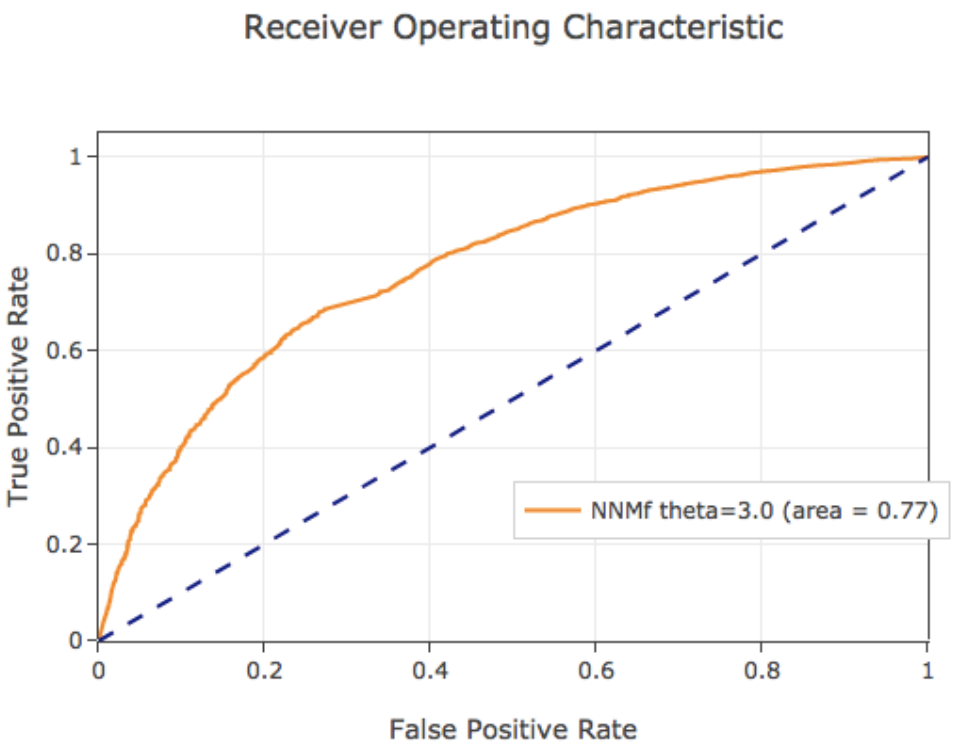
Q21 Trimmed set for NNMF High Variance:

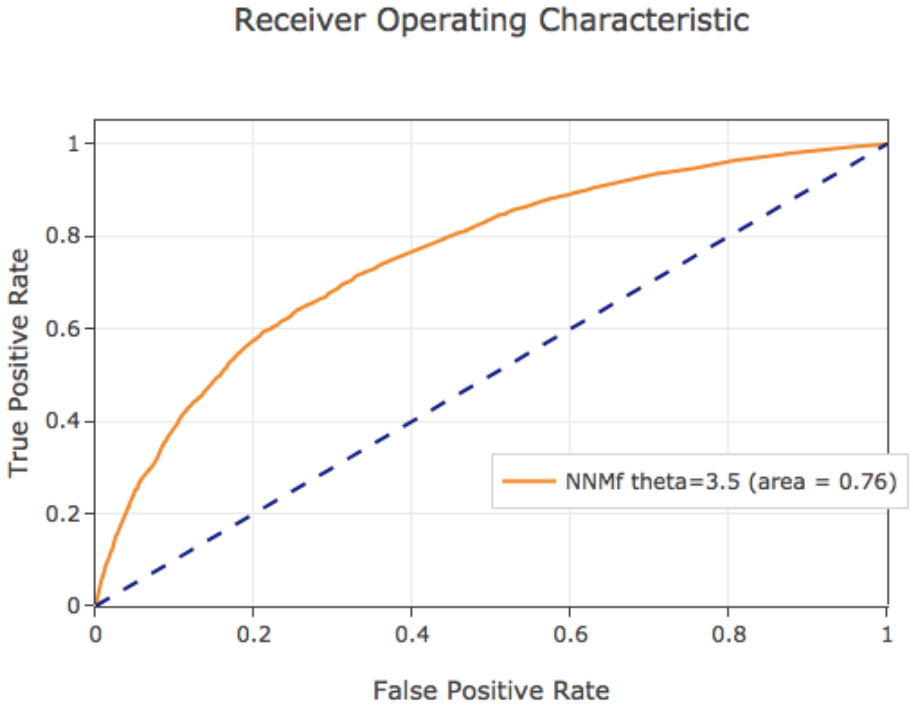


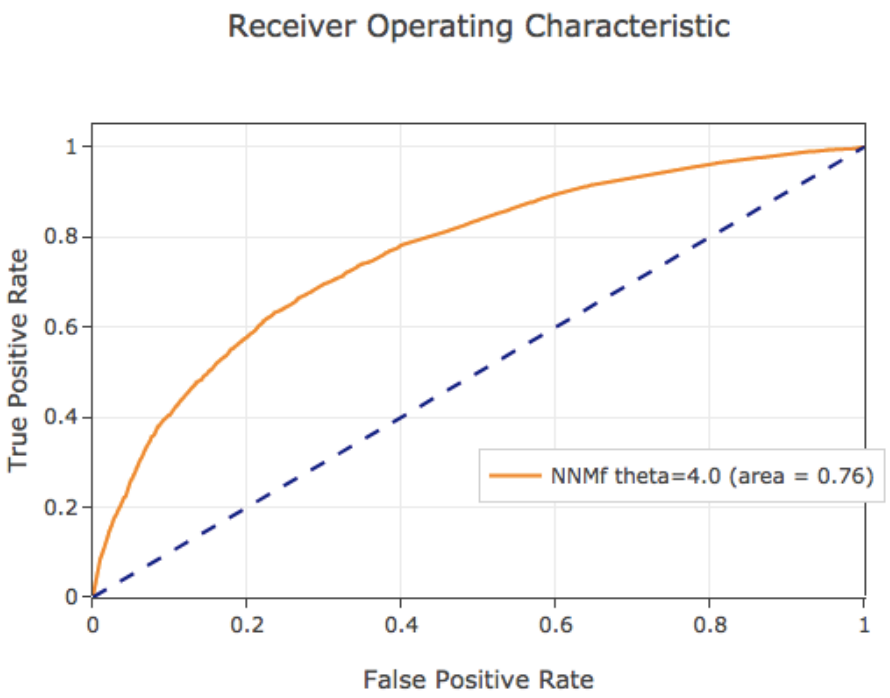
Minimum RMSE: 1.5811

Q22 ROC curve and AUC value




Receiver Operating Characteristic
True Positive Rate
False Positive Rate
NNMf theta=3.0 (area = 0.77)
0
0.2
0.4
0.6
0.8
1


Receiver Operating Characteristic
True Positive Rate
False Positive Rate
NNMf theta=3.5 (area = 0.76)
0
0.2
0.4
0.6
0.8
1

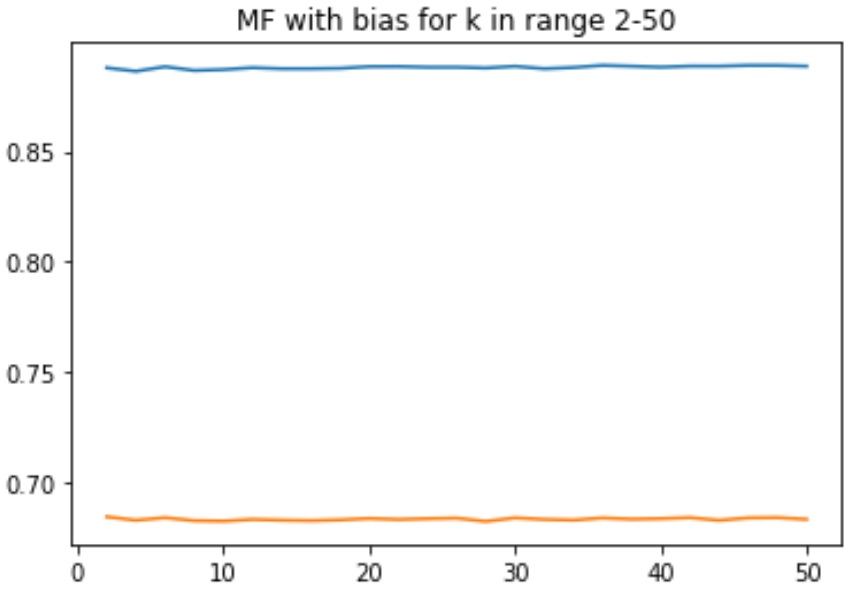


Q23:
The top 10 movies for each k does belong to a small collection of genres. And they have a lot of similarities. The connection between the latent factor and the genres we observed is that when k is increasing, the more similarity between the genres. For example, in the table below, k=8 has 13 different genres whereas when k=10 has 9 and k=12 has 7 and k=18 has 10 and k=19 has 7.

K = 8	K = 10	K = 12	K = 18	K = 19
Mystery|Romance|Sci-Fi|Thriller	Comedy	Comedy	Action|Horror|Sci-Fi|IMAX	Comedy|Drama
Crime|Drama|Thriller	Comedy	Comedy|Mystery	Action|Adventure|Sci-Fi	Drama
Comedy|Drama|Romance|War	Comedy|Romance	Comedy|Drama|Romance	Drama|Romance	Drama|Thriller
Comedy|Musical|Romance	Comedy|Drama|Romance	Comedy	Documentary	Adventure|Drama
Drama	Documentary	Comedy|Musical	Comedy|Crime|Romance	Drama
Crime|Drama|Thriller	Drama|Musical	Comedy|Drama	Drama	Action|Comedy|Drama
Action|Romance	Comedy|Romance	Drama|Romance	Action|Crime|Drama|Thriller	Comedy|Drama|Musical
Adventure|Children|Drama	Drama	Comedy|Drama	Documentary|Horror	Drama
Action|Adventure|Comedy|Fantasy	Drama	Fantasy|Romance|Thriller|IMAX	Adventure|Children|Fantasy	Drama|Thriller
Romance	Action|Crime|Drama|Thriller	Drama|Thriller	Action|Crime|Thriller|IMAX	Drama

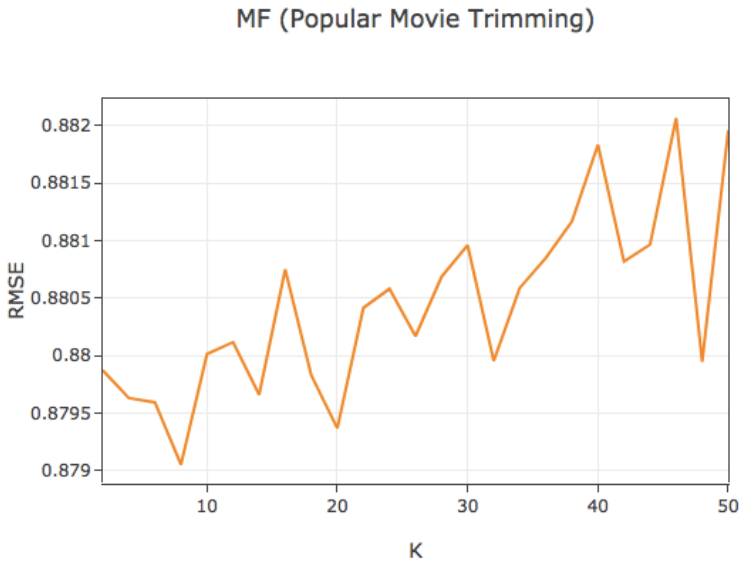
Model-based Collaborative Filtering

Q24: The difference is very very small so it is hard to tell from the picture, but the value does change:



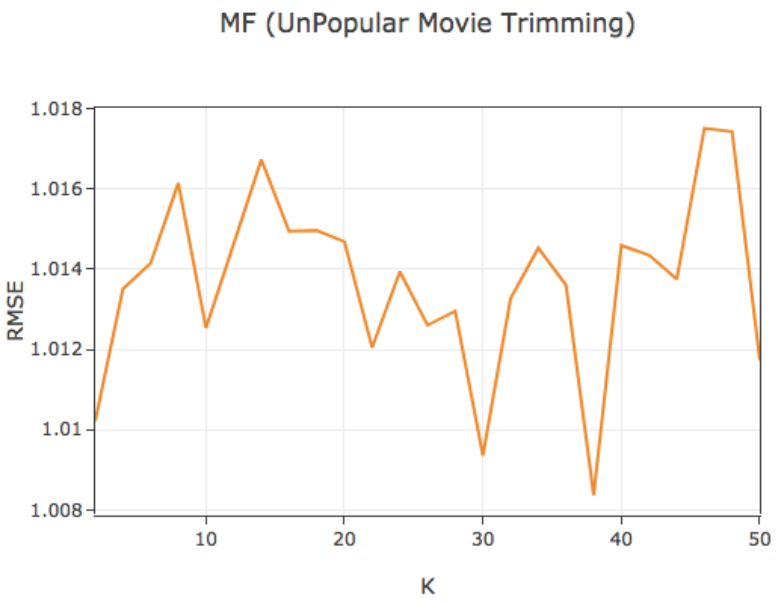
Q25:
minmae happened at k = :16
minrmse happened at k = 24
The best latent factor we used is 16.

Q26 Trimmed set for MF Popular:



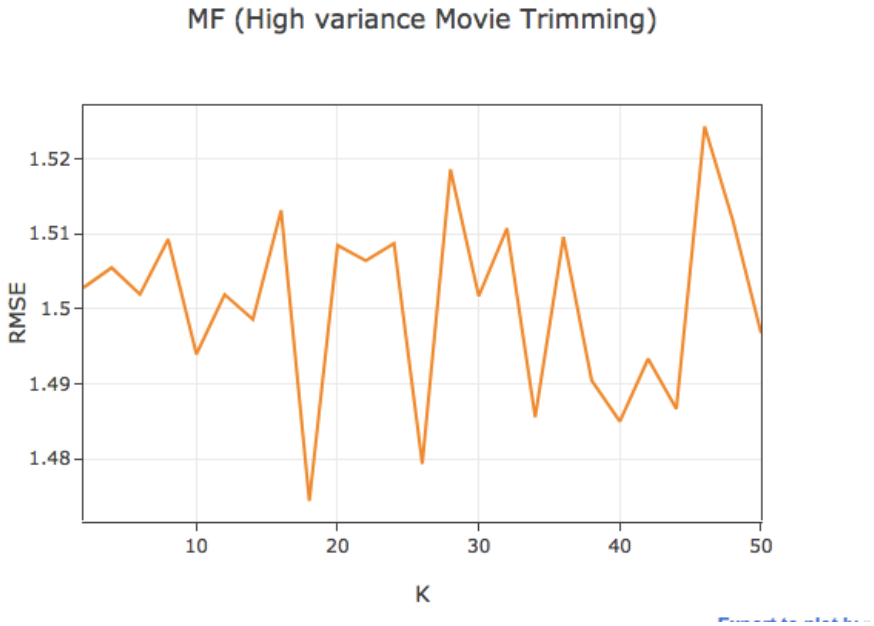
Minimum RMSE: 0.8791

Q27 Trimmed set for MF UnPopular:



Minimum RMSE: 1.0084

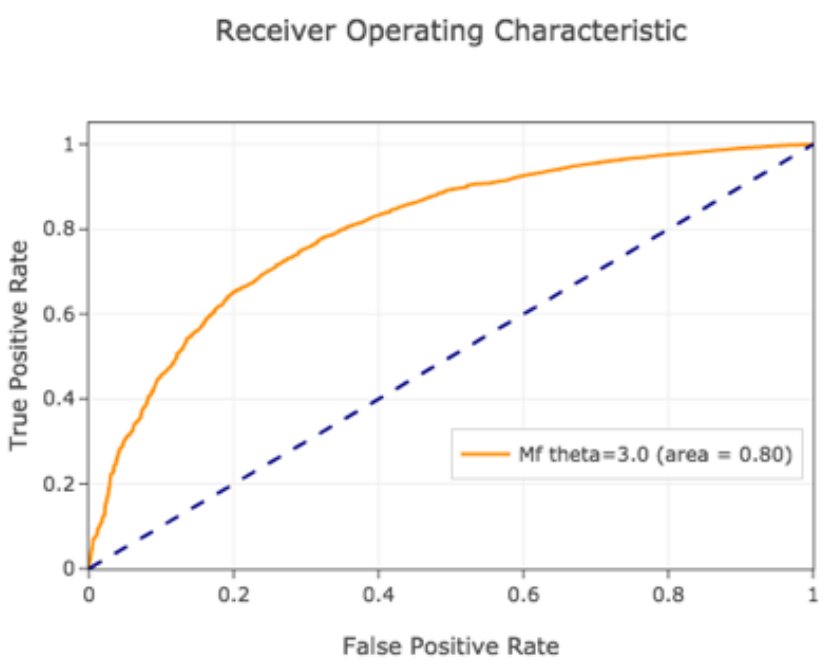
Q28 Trimmed set for MF High Variance:

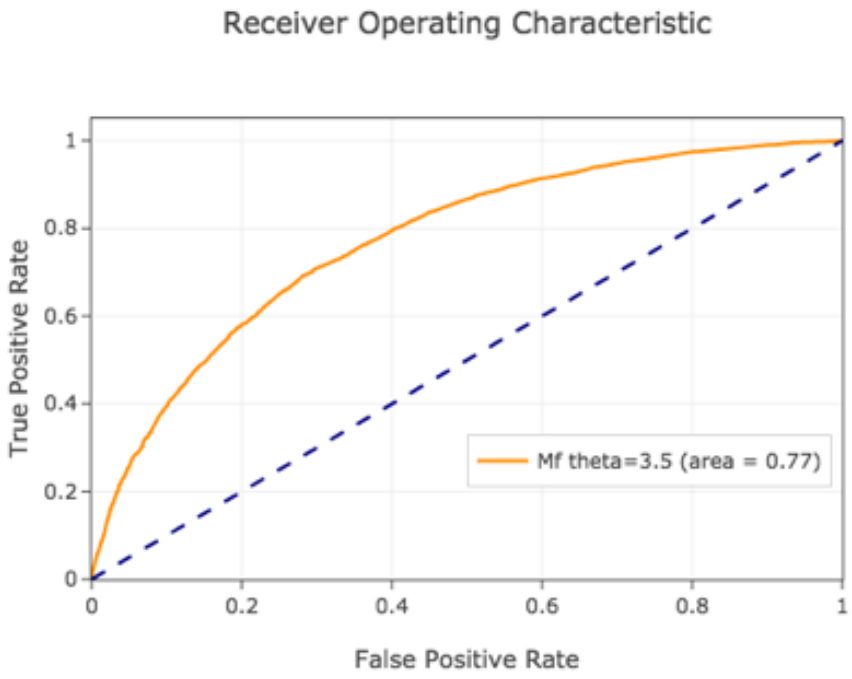


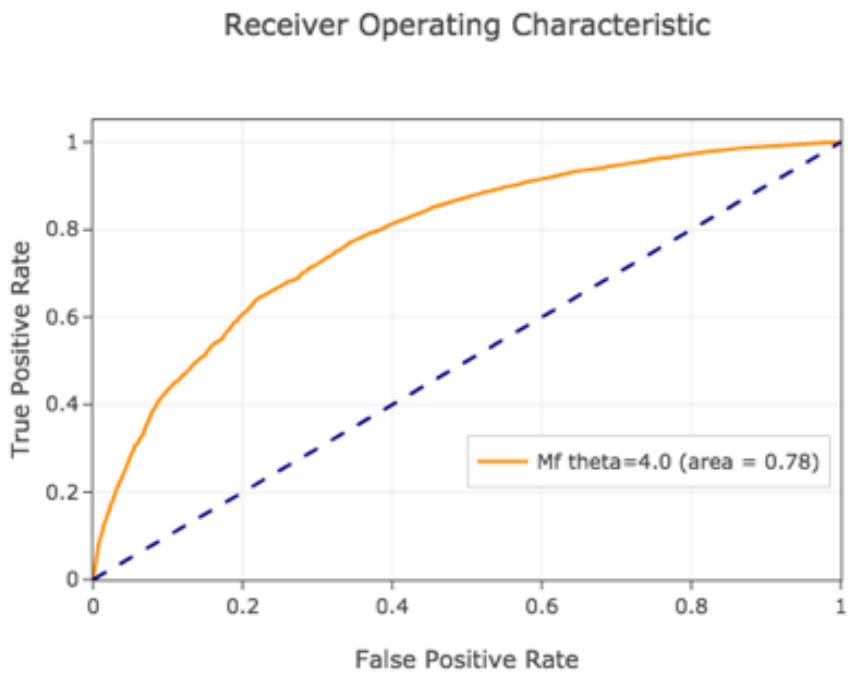
Minimum RMSE: 1.4744

Q29 ROC and AUC:


Receiver Operating Characteristic
True Positive Rate
False Positive Rate
Mf theta=2.5 (area = 0.77)
0
0.2
0.4
0.6
0.8
1


Receiver Operating Characteristic
True Positive Rate
False Positive Rate
Mf theta=3.0 (area = 0.80)
0
0.2
0.4
0.6
0.8
1


Receiver Operating Characteristic
True Positive Rate
False Positive Rate
0
0.2
0.4
0.6
0.8
1
Mf theta=3.5 (area = 0.77)


Receiver Operating Characteristic
True Positive Rate
False Positive Rate
0
0.2
0.4
0.6
0.8
1
Mf theta=4.0 (area = 0.78)

Naive Collaborative Filtering

Q30:
The Average RMSE for Original Test Set:
0.954963227285

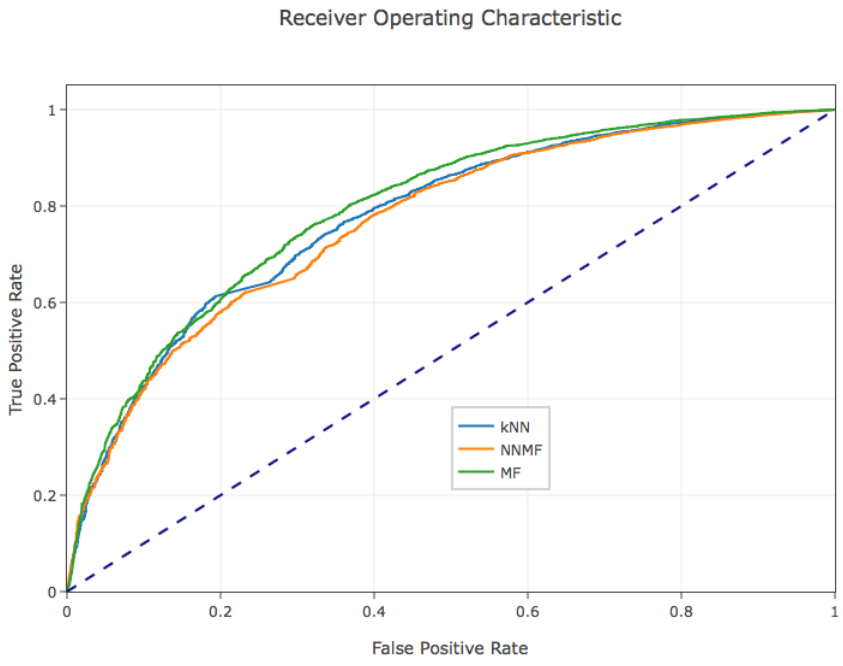
Q31:
The Average RMSE for Popular Movie Trimmed Test Set:
0.951702566585

Q32:
The Average RMSE for Unpopular Movie Trimmed Test Set:
1.00587357969

Q33:
The Average RMSE for High Variance Movie Trimmed Test Set:
1.51820049369

Performance Comparison

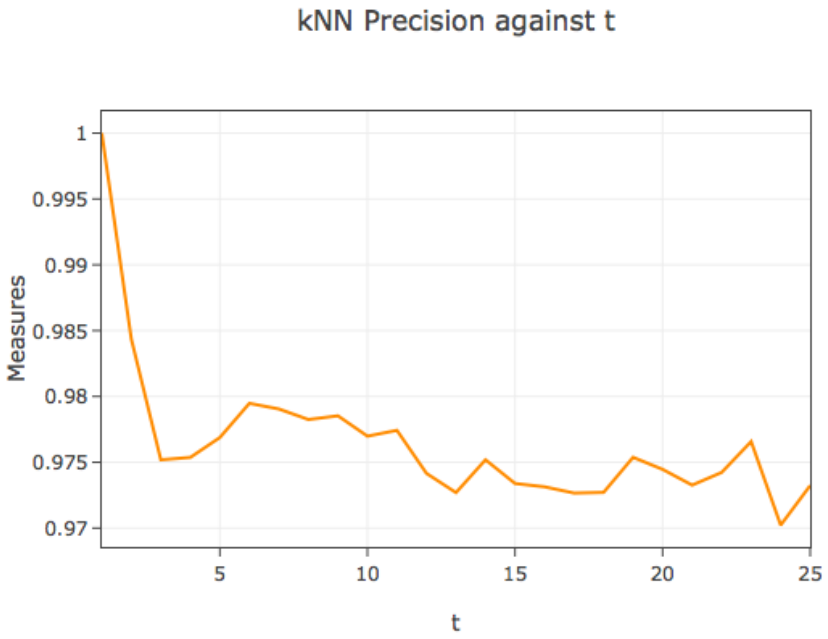
Q34:
We plotted the three ROC curves from kNN, NNMF and MF together to compare them:

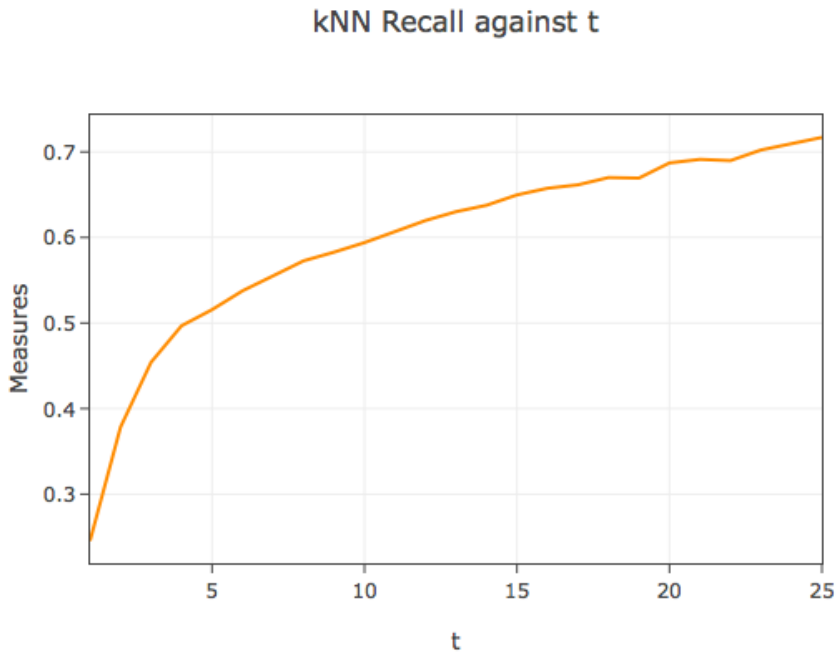


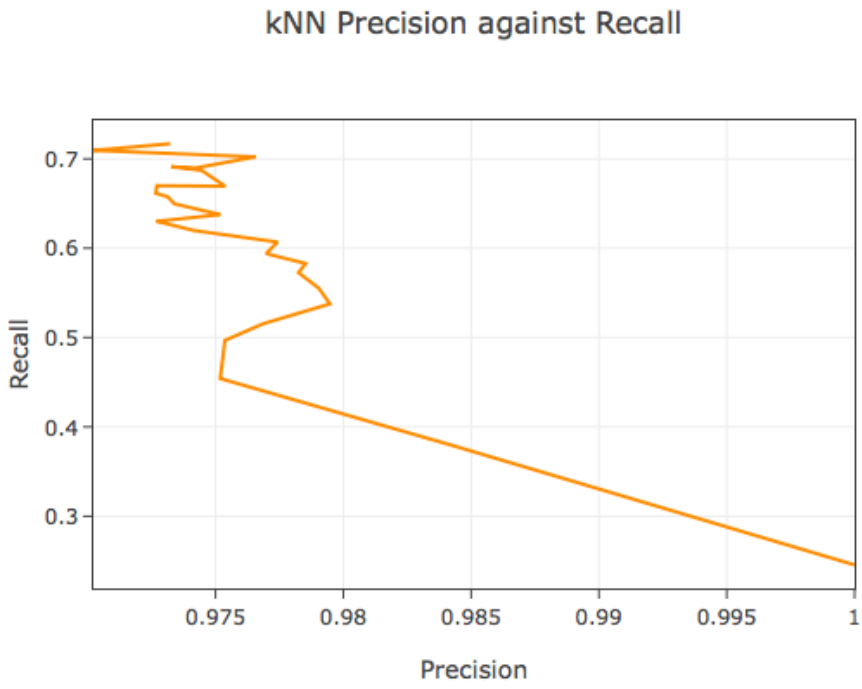
We can see that MF seems to have the highest area under curve.

Q35: Precision is basically out of the results you predicted, what is the percentage of the correct results. Recall is that out of all the ground truth positive, what is the percentage of the covering of the true positive. So in other words, precision is accuracy whereas recall is coverage.

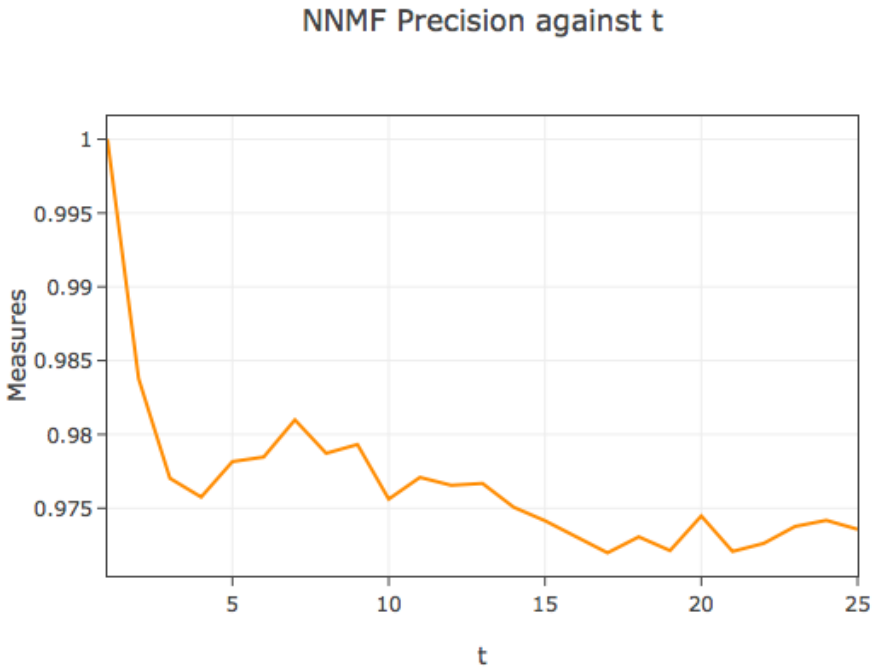
Q36:

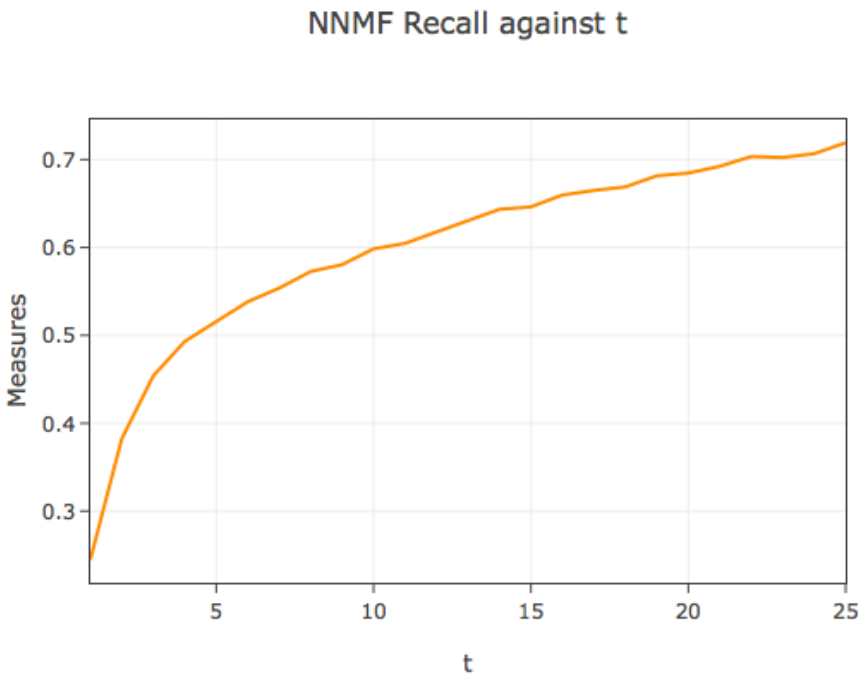


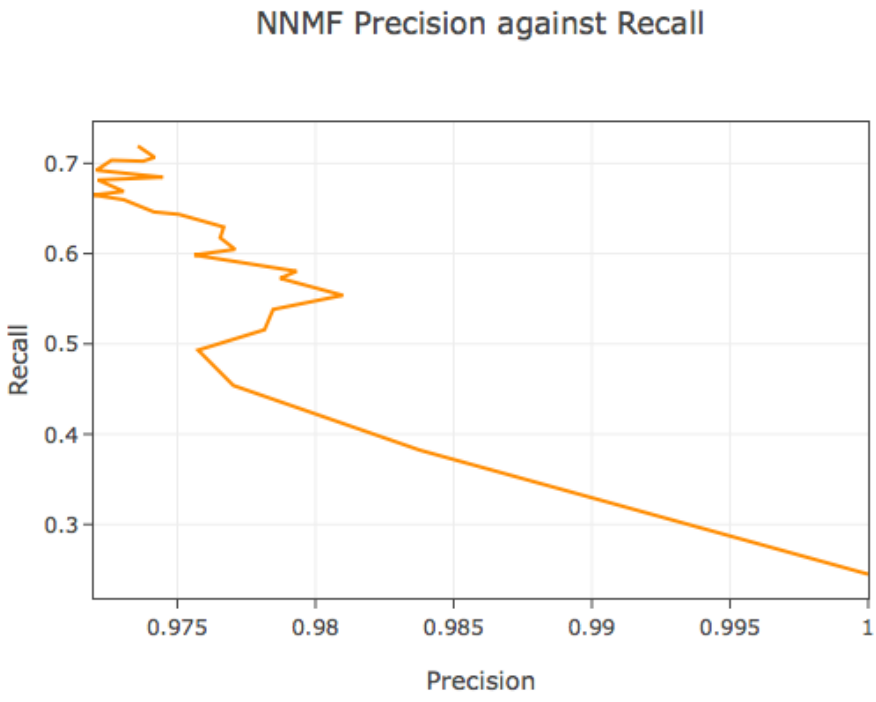



kNN Precision against Recall
Recall
0.7
0.6
0.5
0.4
0.3
0.975
0.98
0.985
0.99
0.995
1
Precision

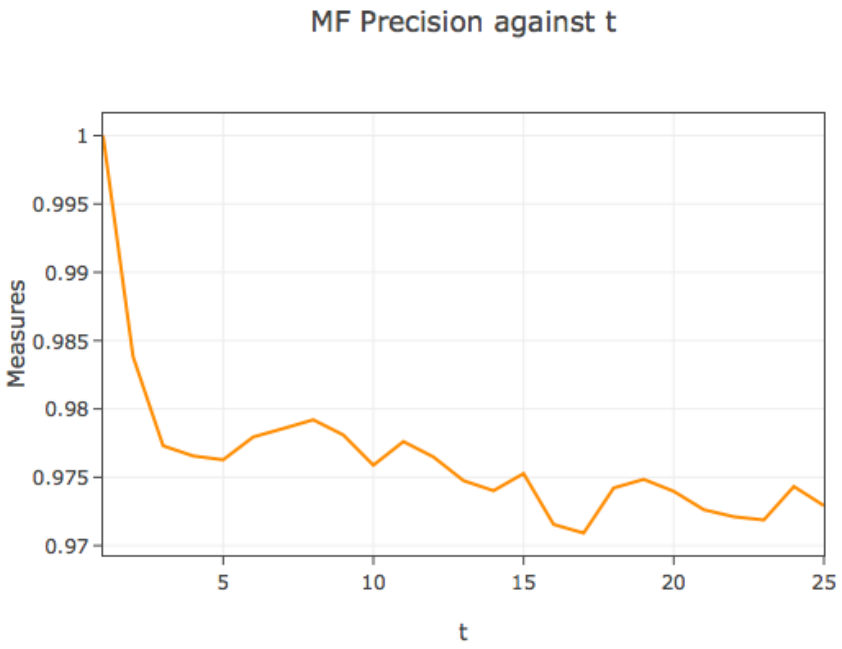
Q37:

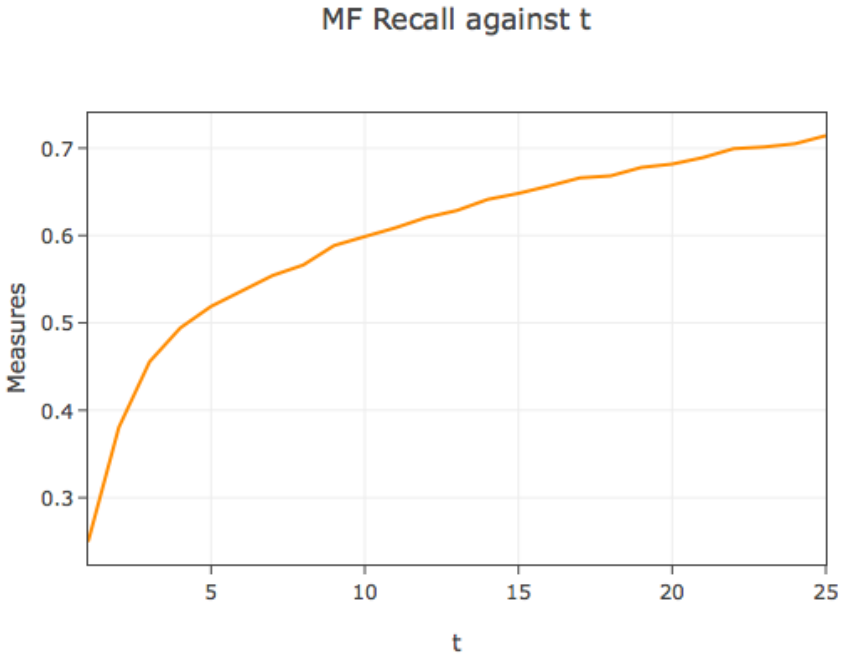

NNMF Precision against t
Measures
1
0.995
0.99
0.985
0.98
0.975
5
10
15
20
25
t

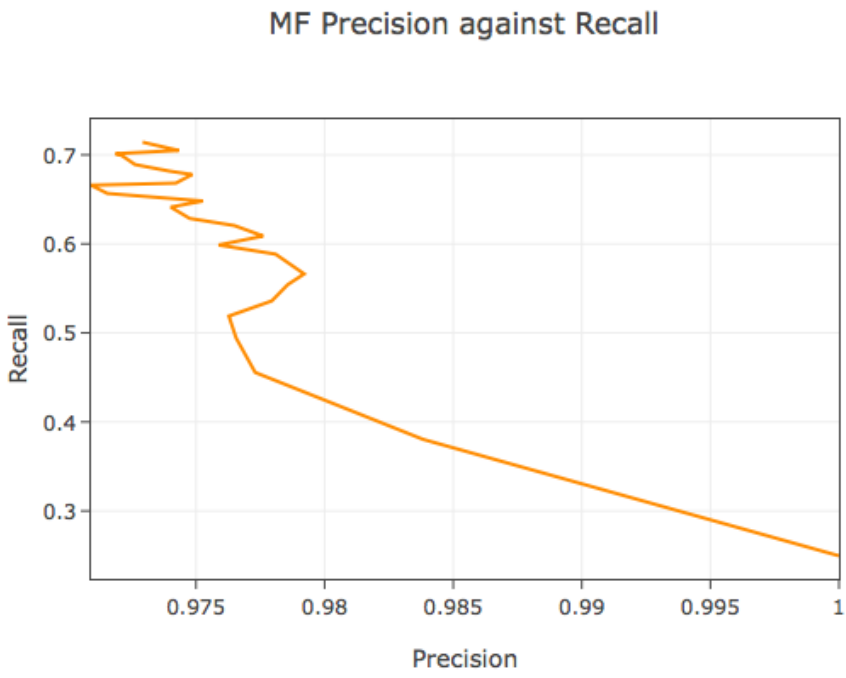

NNMF Recall against t
Measures
0.7
0.6
0.5
0.4
0.3
5
10
15
20
25
t


NNMF Precision against Recall
Recall
0.7
0.6
0.5
0.4
0.3
0.975
0.98
0.985
0.99
0.995
1
Precision

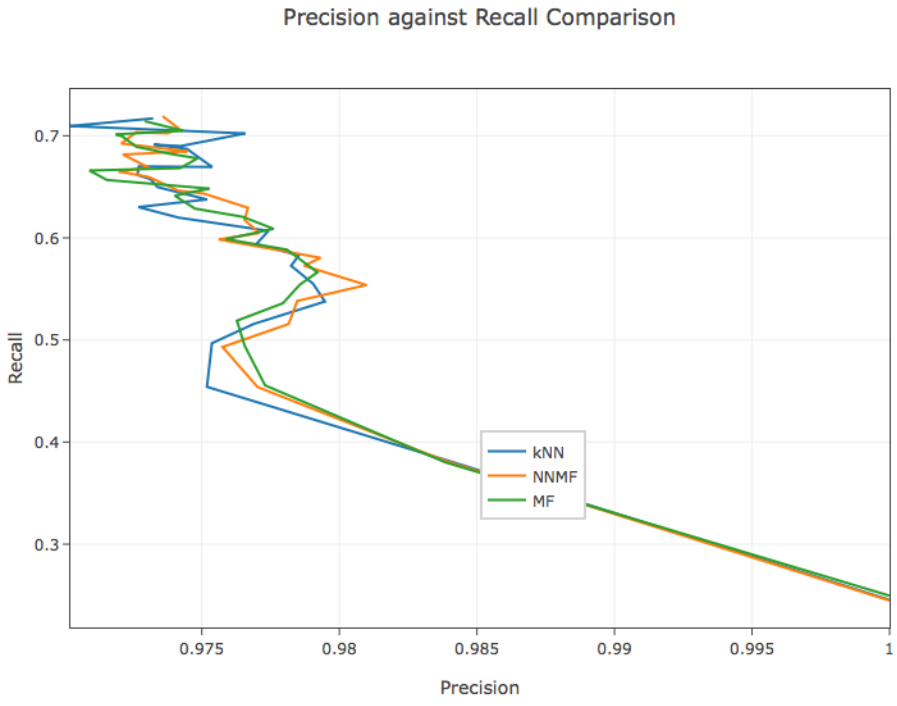
Q38:


MF Precision against t
Measures
1
0.995
0.99
0.985
0.98
0.975
0.97
5
10
15
20
25
t


MF Recall against t
Measures
0.7
0.6
0.5
0.4
0.3
5
10
15
20
25
t



Q39:



The precision and recall curves for these three methods show similar performance.