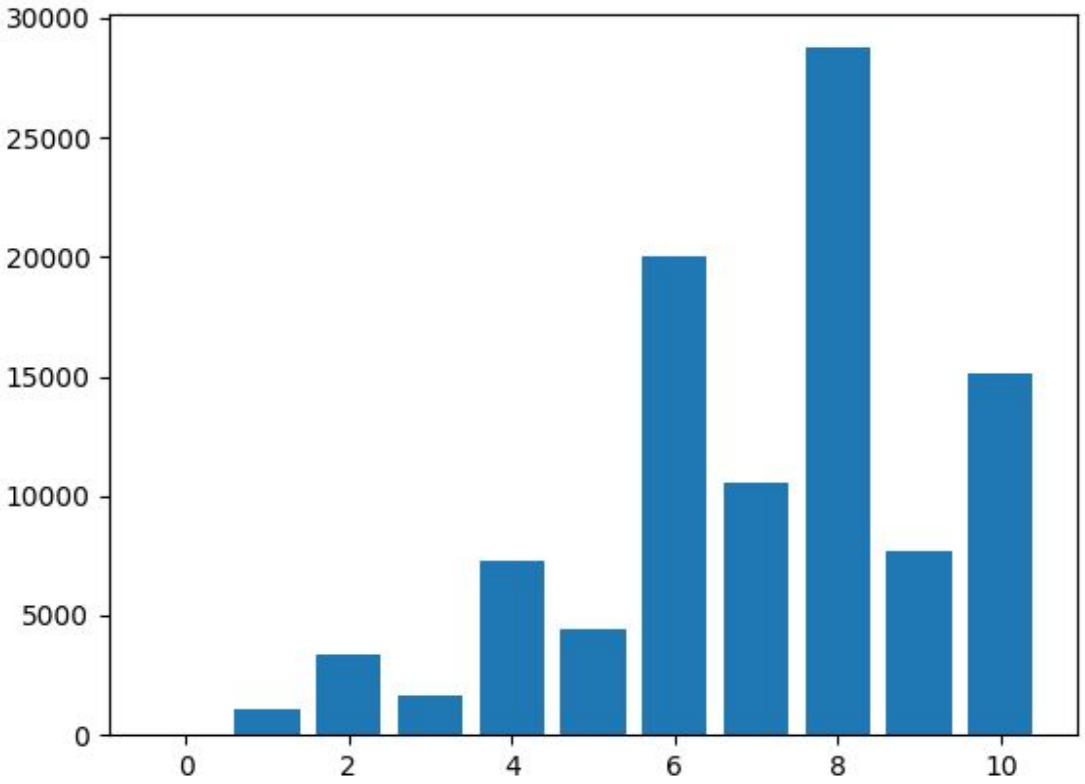
ECE 219 Project3 Report Feb 22. 2018

Don Lee
Zixia Weng 305029822
Shunji Zhan 405030387

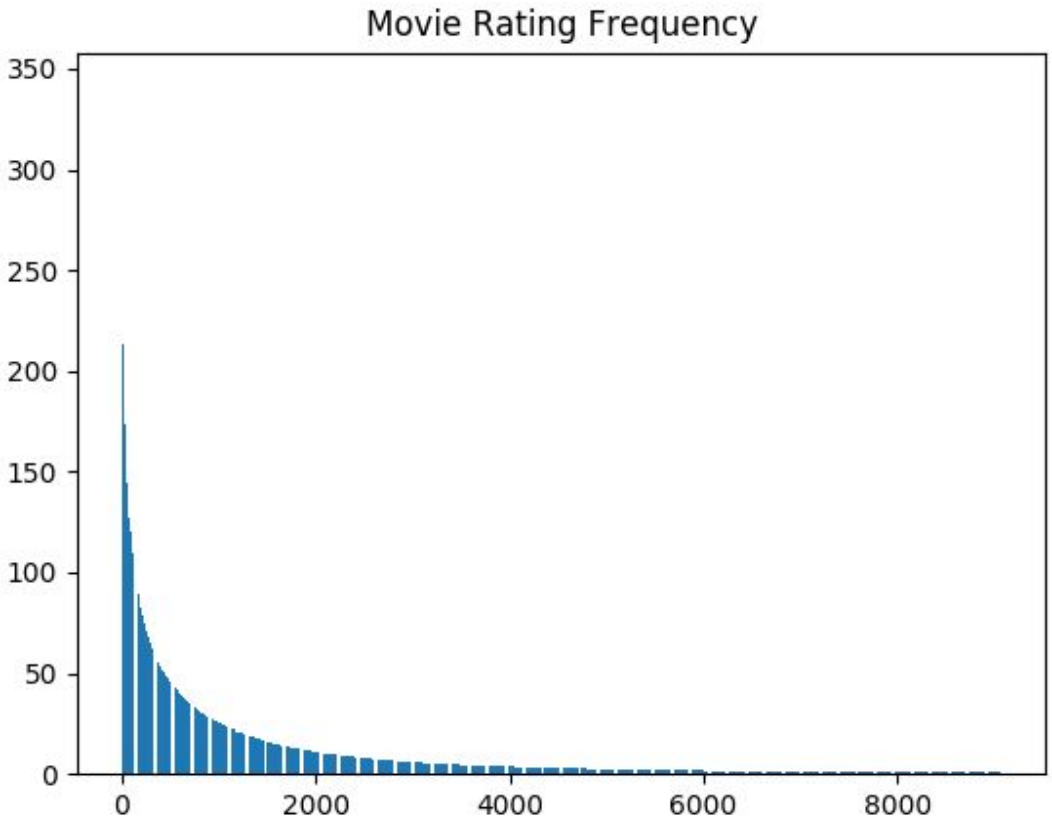
Q1:
Sparsity = 0.0164391416087

Q2:

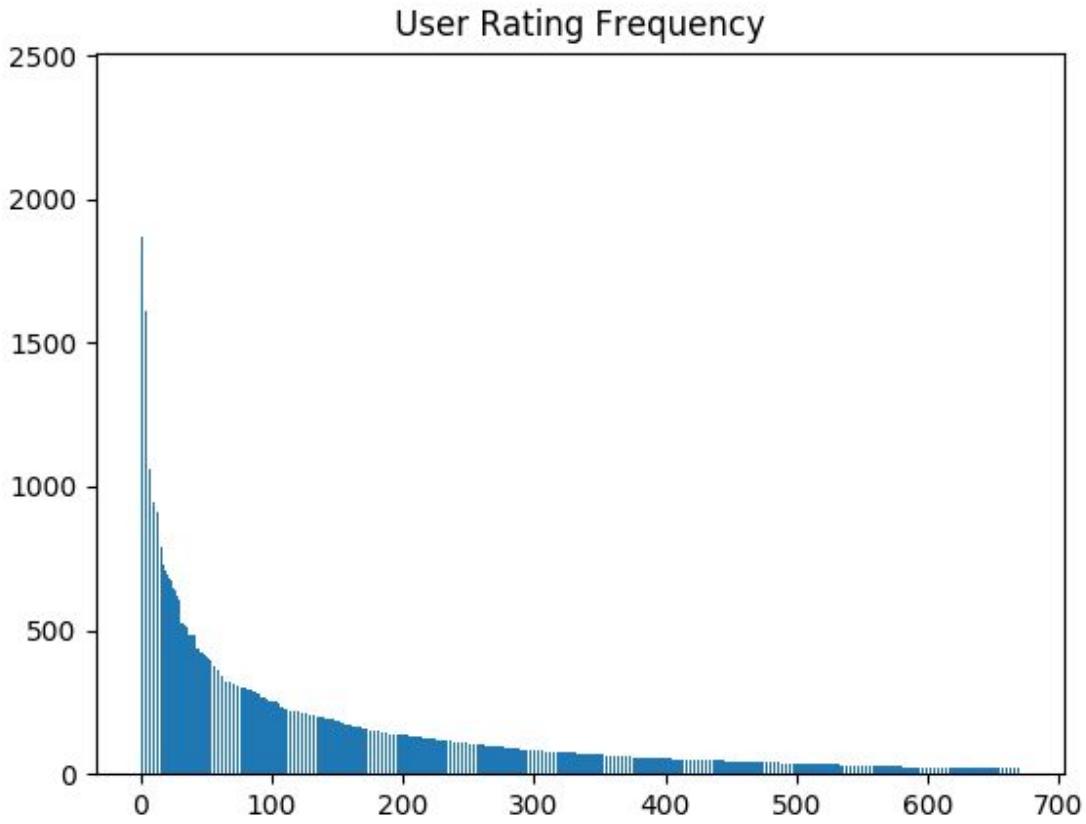


Briefly comment on the shape of the histogram：
The ratings of movies mainly range from 6 to 10 and the histograms are not distributed normally. The Most ratings are in 8/10 points.

Q3:


Movie Rating Frequency
350
300
250
200
150
100
50
0
0
2000
4000
6000
8000

Q4:



Q5:

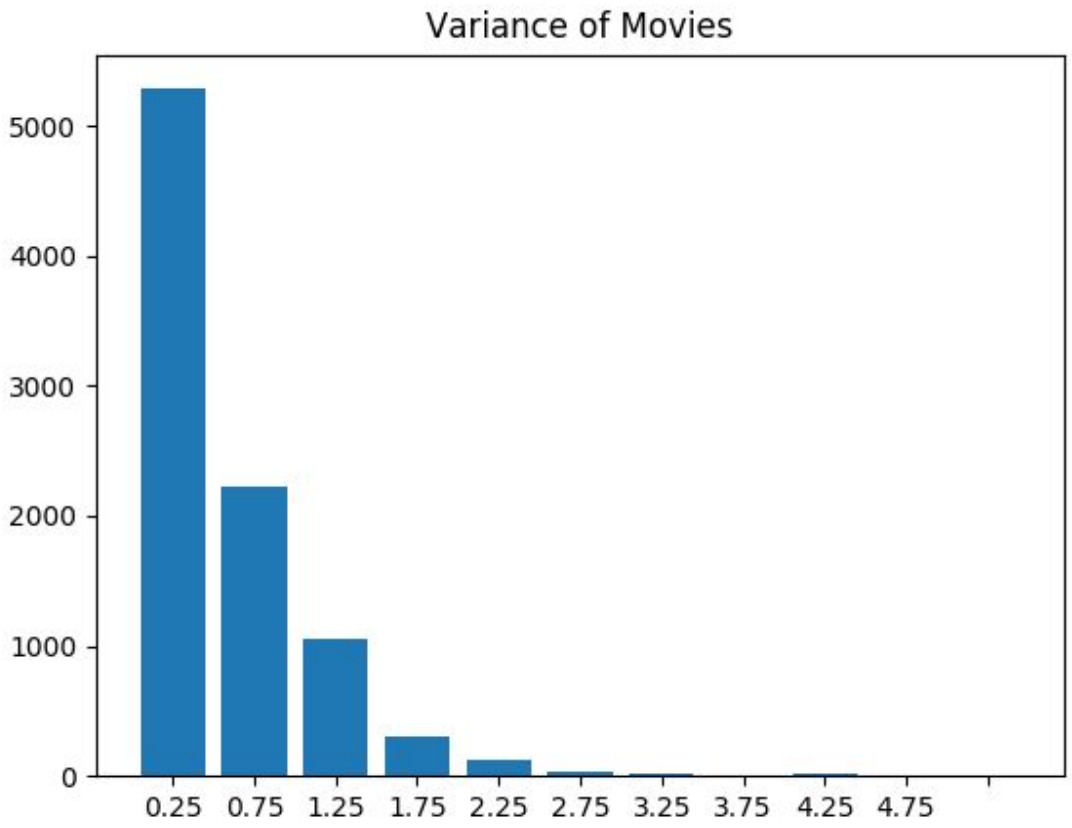
The salient features of the distribution found in question 3 was because there are many movies that are super famous and thus many many people rate it, and some not famous movie only have very few ratings. Since the super famous movies are relatively few, the shape of the frequency graph is salient.This feature implies the sparsity in the rating matrix with some movies having lots of ratings while others have very few.

Thus if we are required to use limited samples of the dataset, the reduction method can help us to filter the unpopular(the movies with less rating frequency) movies and only use those popular data in the dataset because they are more representative in the features we are looking for.

Q6:

Briefly comment on the shape of the histogram:

This histogram shows the variances of ratings received by movies. Each histogram represents a range 0f 0.5 starting from 0. And our result shows that most variance ranges in the bin of 0-0.5, which is very normal and reasonable because, all the ratings of a movie should be very concentrated and not diverged, those variance bigger than 0.5 should be relatively less because some movies may be controversial and has a relatively large range of ratings.

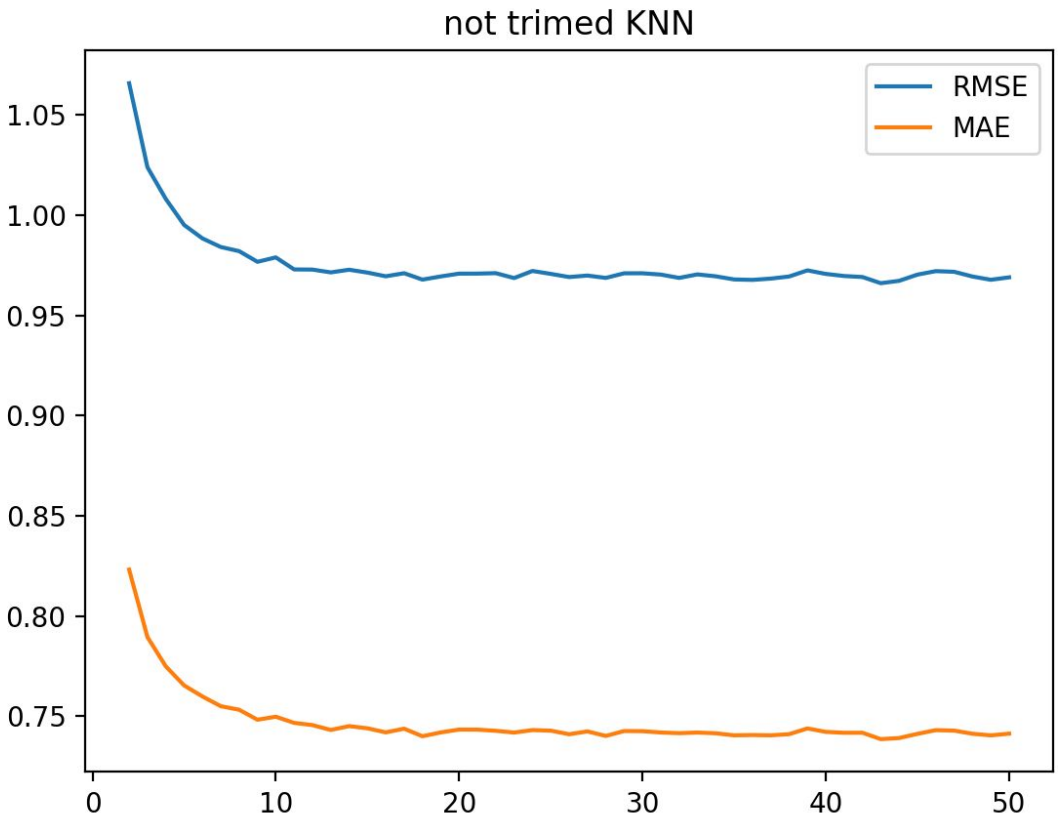


Q7 :
Uu = (sum all ruk where u in the set Iu) / num of elements in Uu

Q8:
The intersection is the set of items for which ratings have been specified by user u and v. This can be empty since rating matrix R is sparse and it is totally possible that two users didn't rate any same movie.

Q9:
Since some users tends to rate every movie really low, some other user tends to rate every movie really high, in order the predict more accurately, we will do mean center first so that the above situation will not affect the prediction.

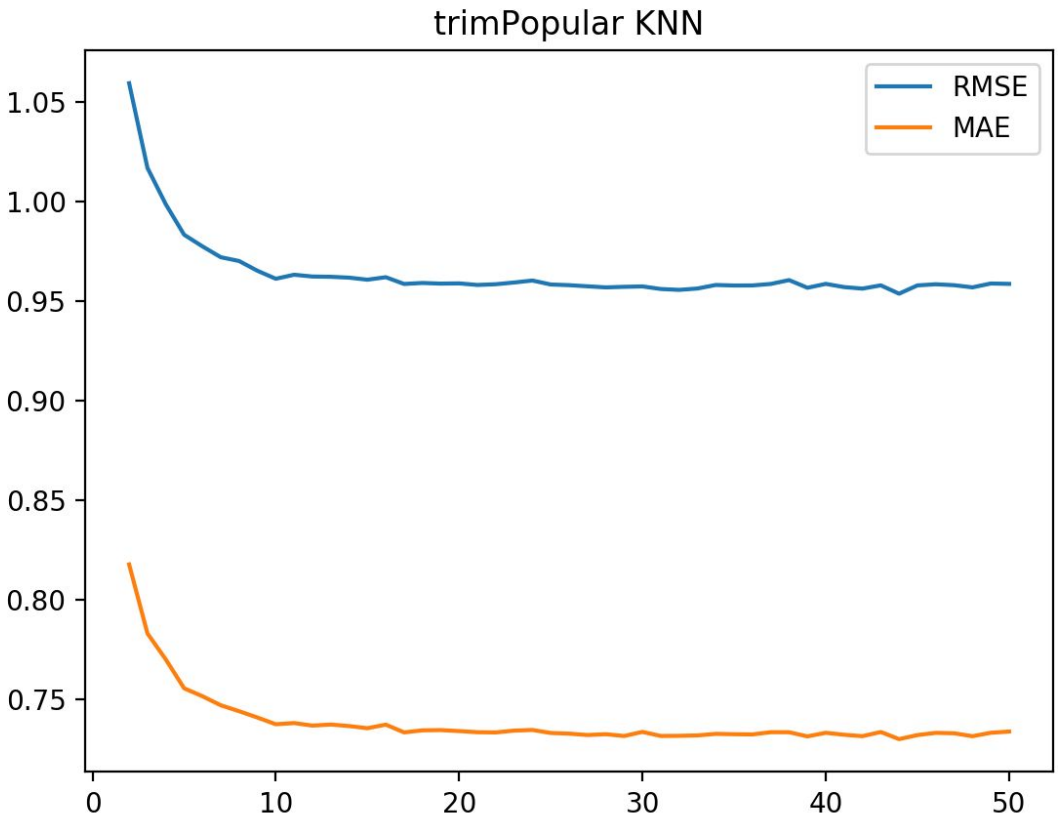
Q10



Steady RMSE = 0.973
Steady MAE = 0.747

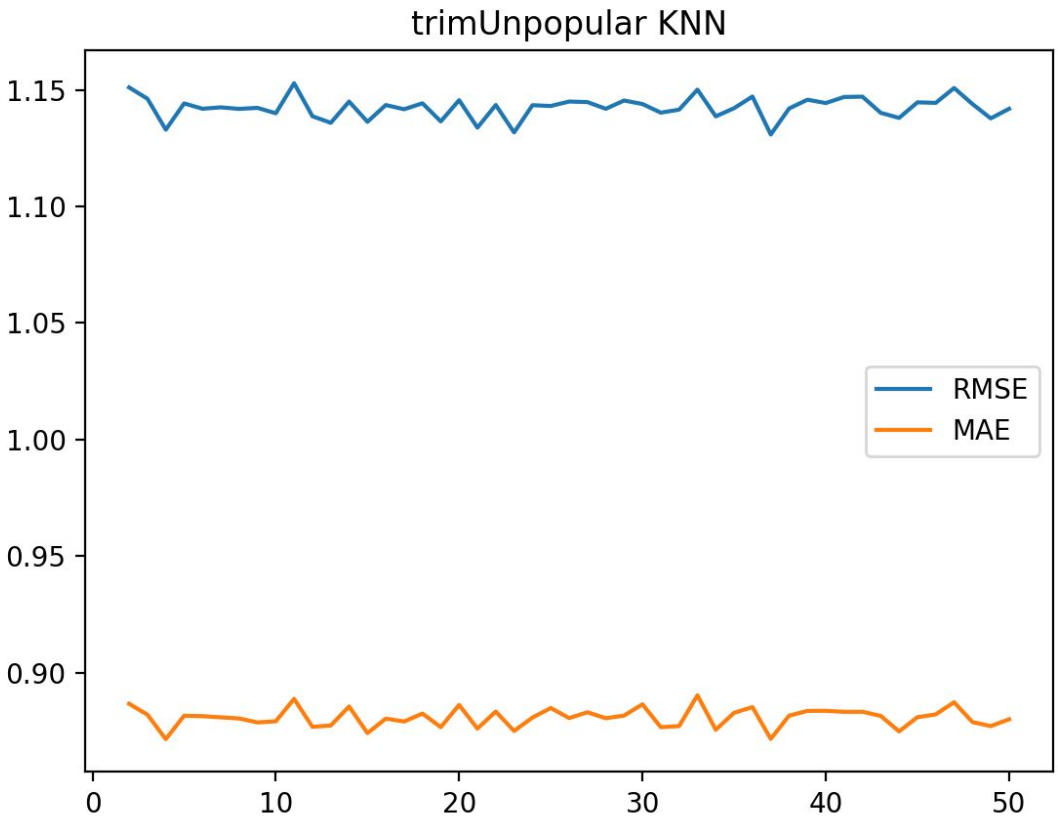
Q11
K = 20

Q12



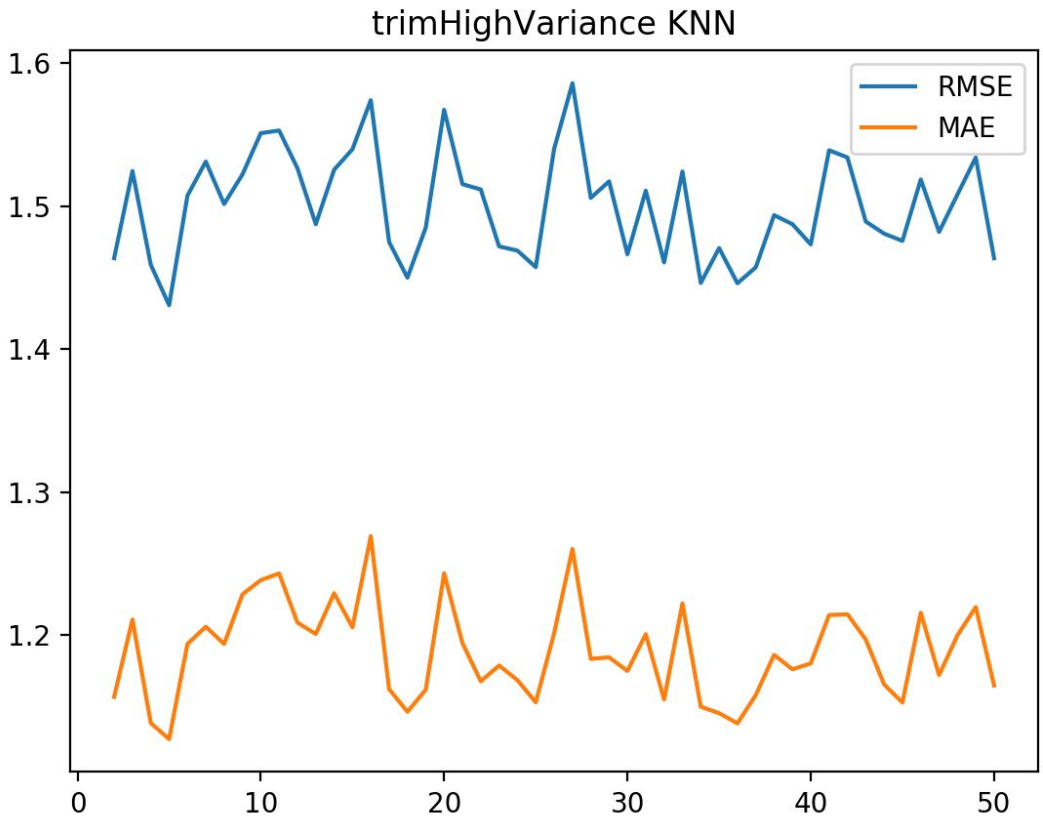
Steady RMSE = 0.960
Steady MAE = 0.746

Q13

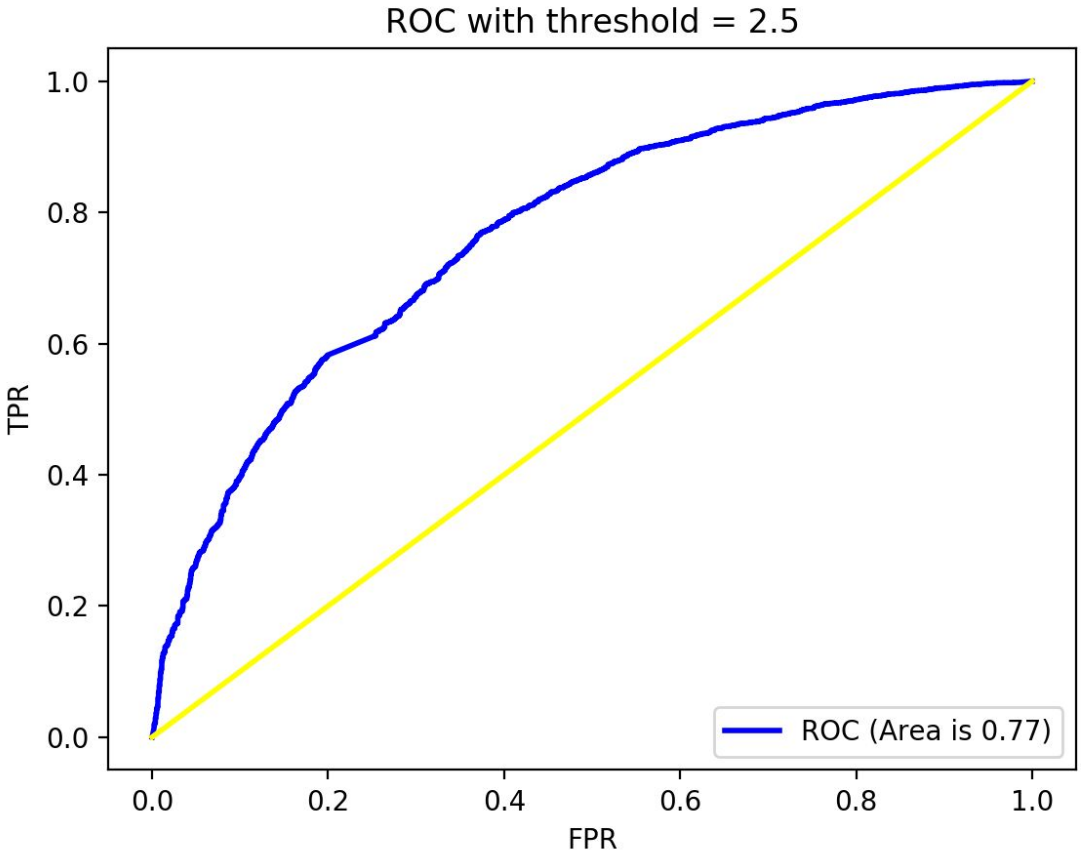


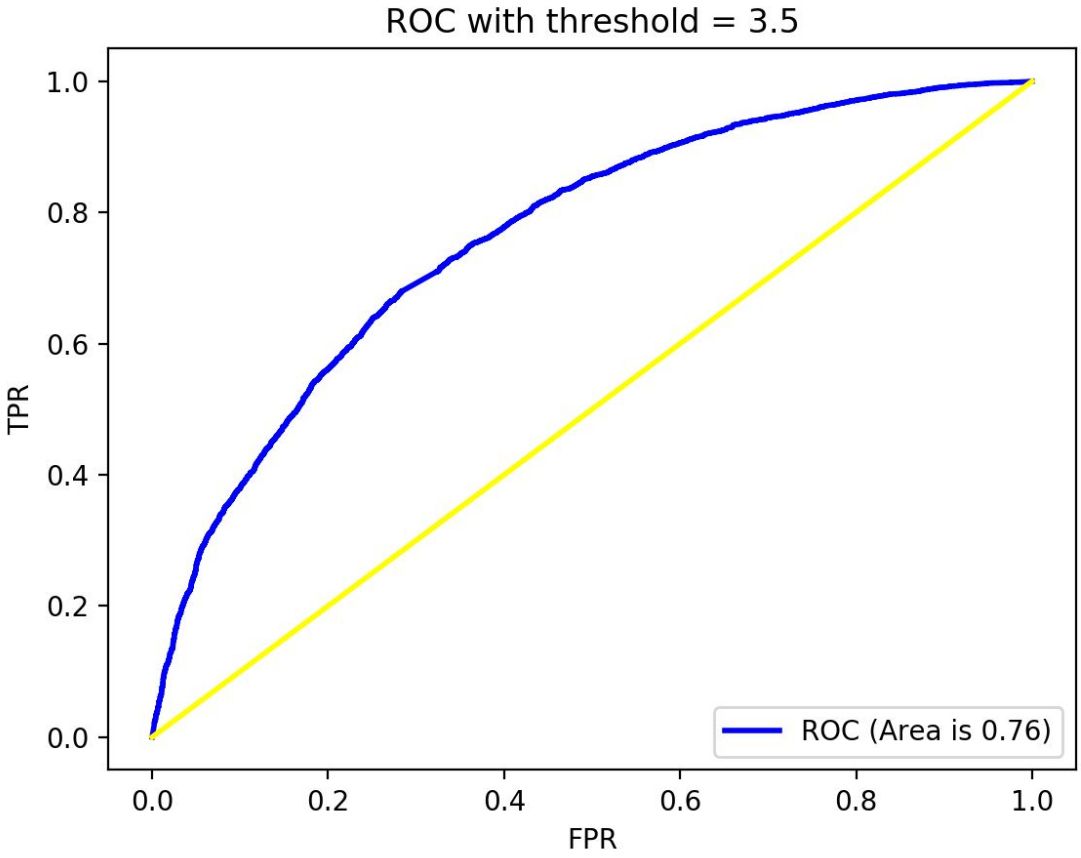
Steady RMSE = 0.875
Steady MAE = 1.14

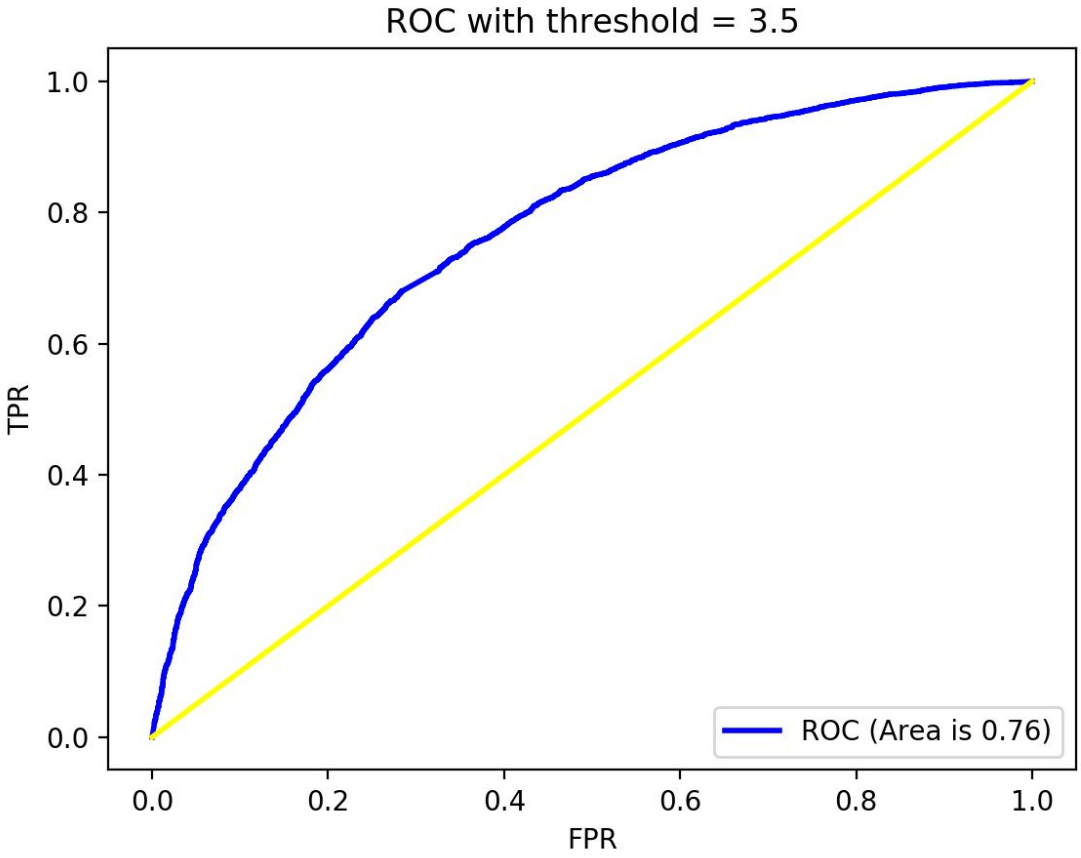
Q14

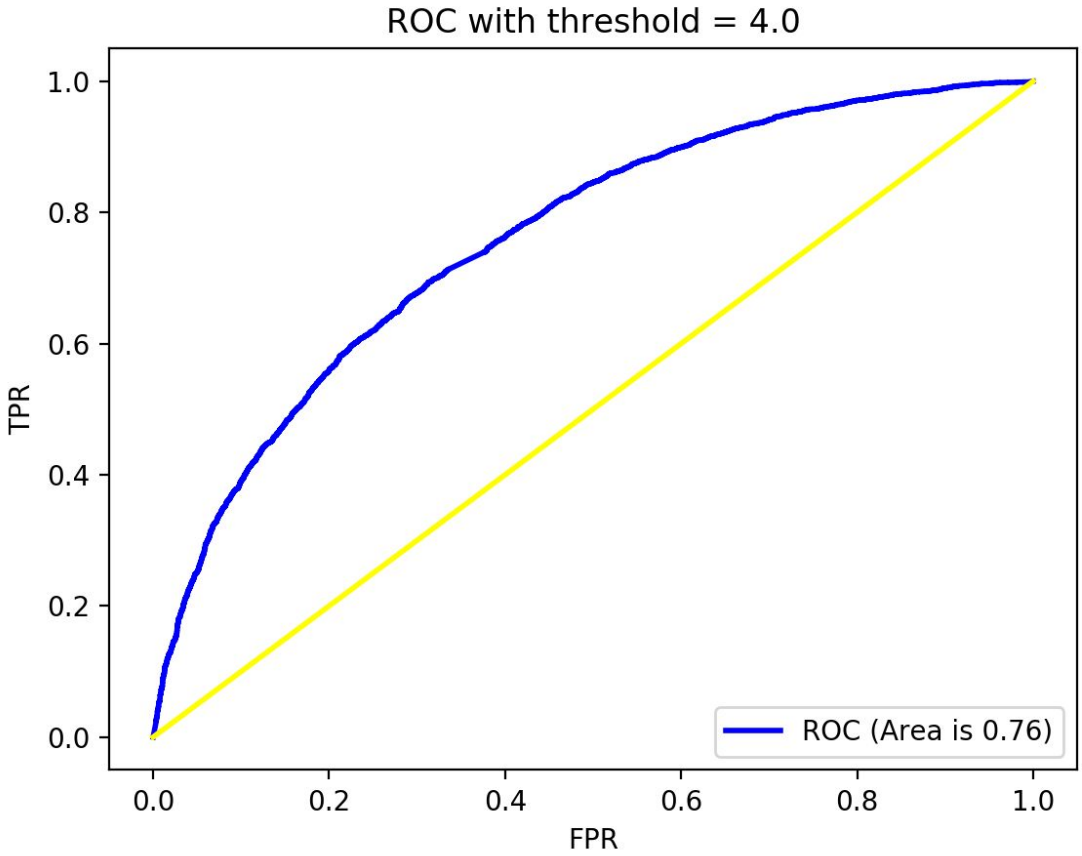


Q15


ROC with threshold = 2.5
TPR
FPR
0.0
0.2
0.4
0.6
0.8
1.0
ROC (Area is 0.77)

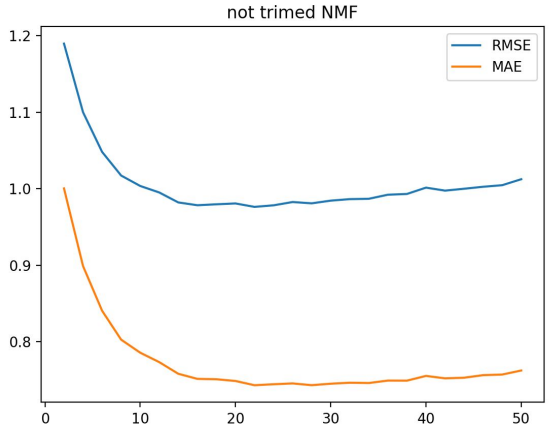

ROC with threshold = 3.5
TPR
FPR
0.0
0.2
0.4
0.6
0.8
1.0
ROC (Area is 0.76)


ROC with threshold = 3.5
TPR
FPR
0.0
0.2
0.4
0.6
0.8
1.0
ROC (Area is 0.76)



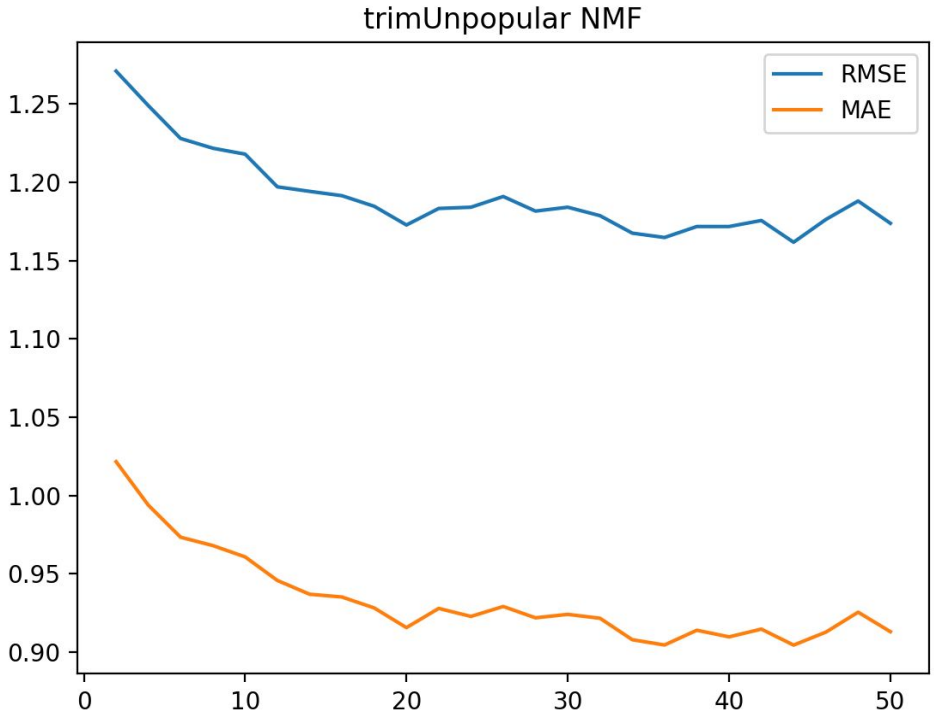
Q16:
The equation is a convex function. When U fixed, the least square is the lowest point in the convex shape.

Q17 - 22:
Steady RMSE = 0.963
Steady MAE = 0.737



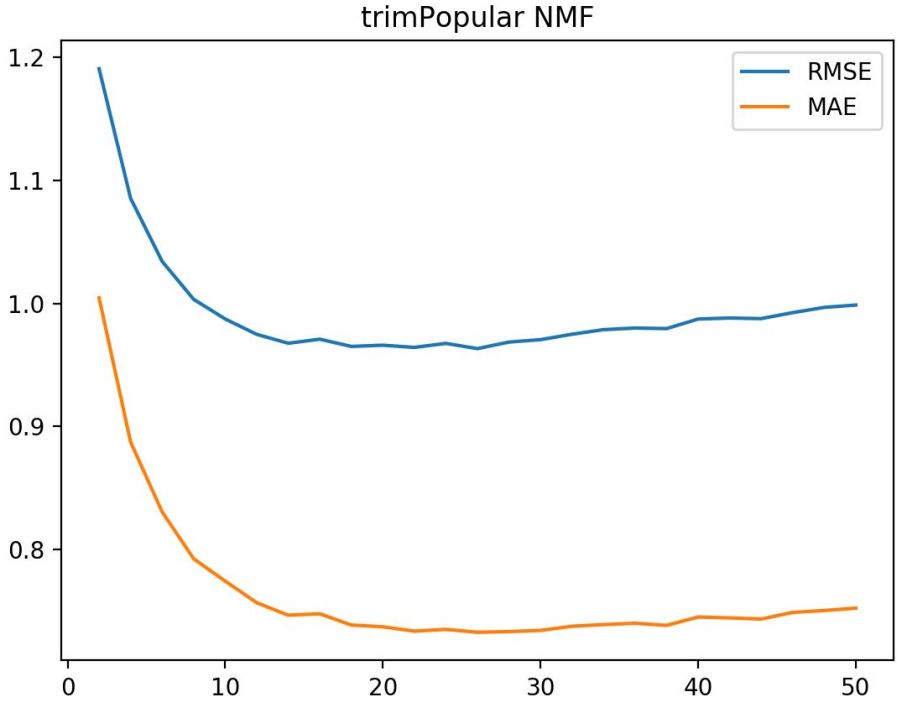
Q18

The optimal k found from plot of Q17 is about k = 20. There are 18 named movie genres and also an empty option in the dataset. The optimal value of k is found to be very close to the number of movie genres.

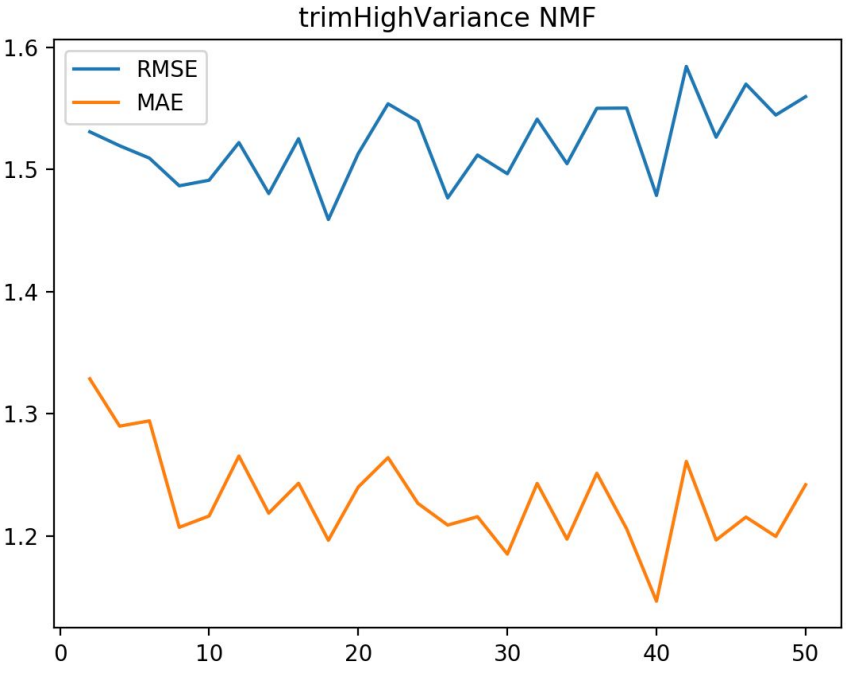


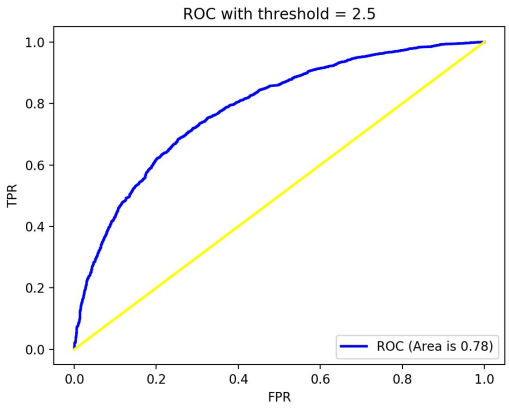
Steady RMSE = 0.966

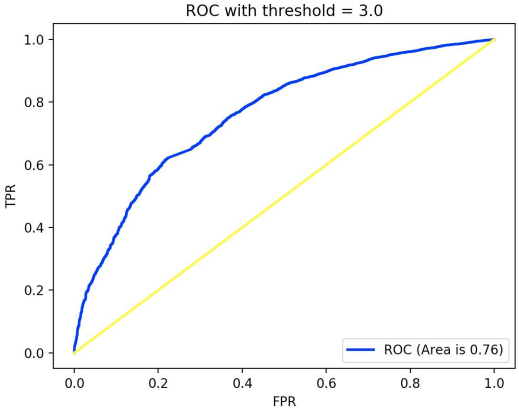
Steady MAE = 0.733

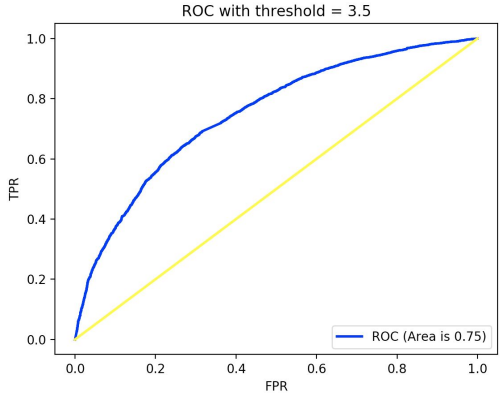


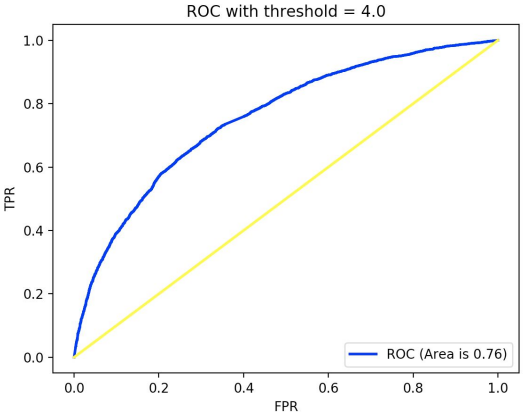
Steady RMSE = 0.973
Steady MAE = 0.747







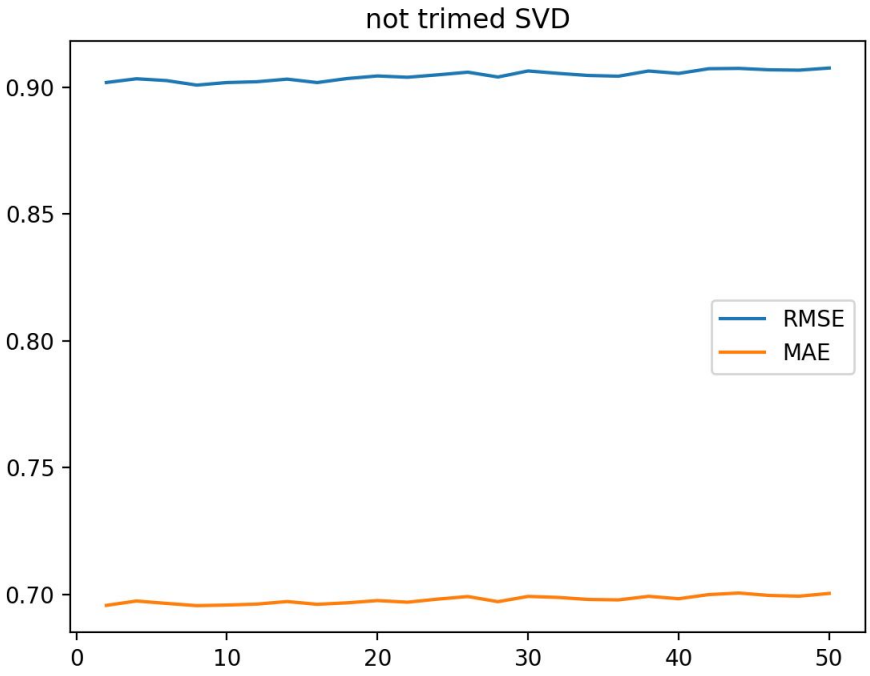


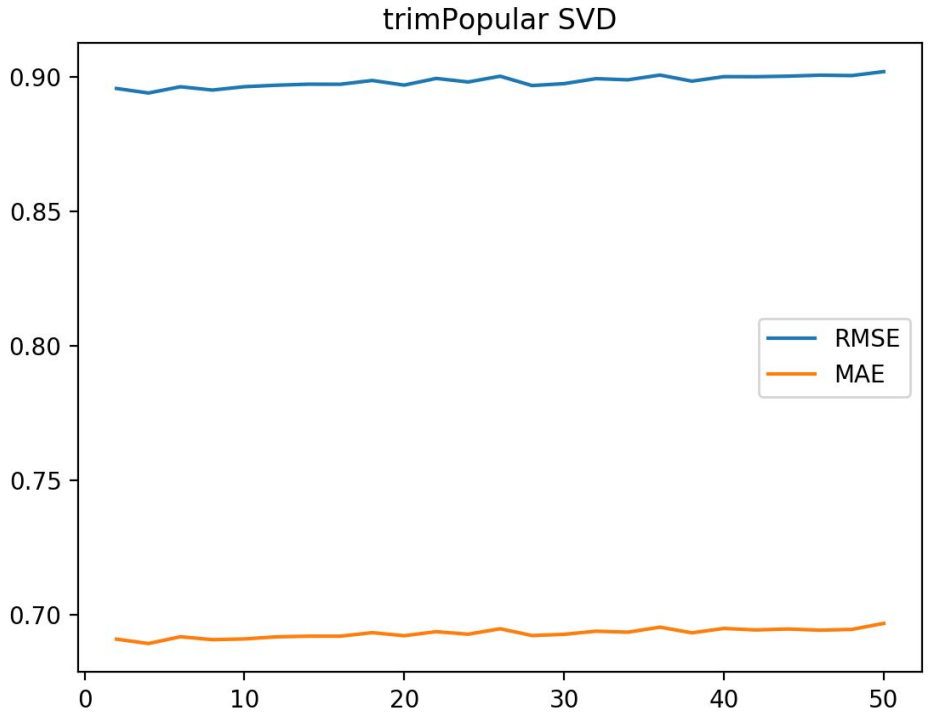


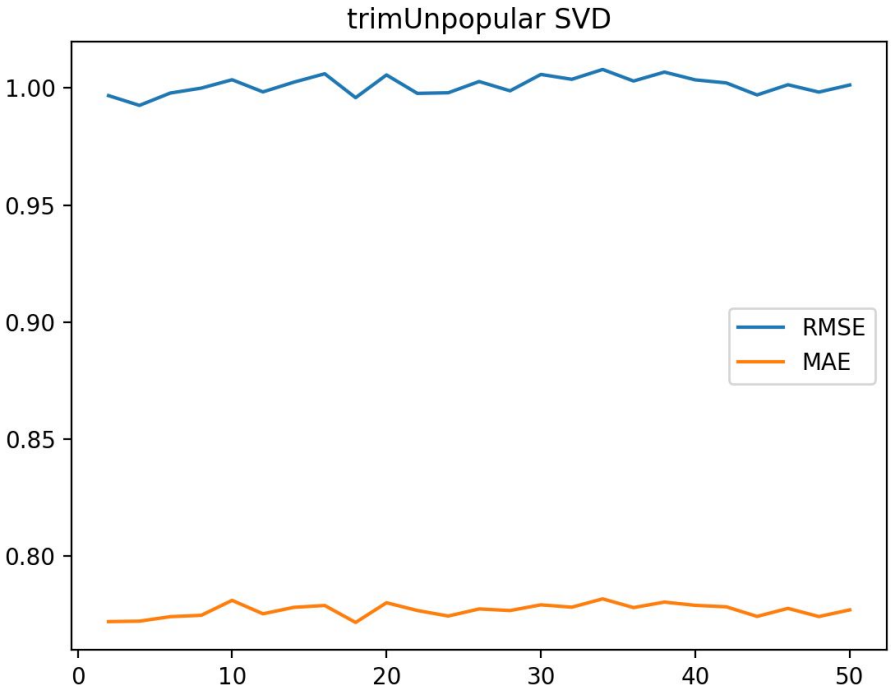
Q23

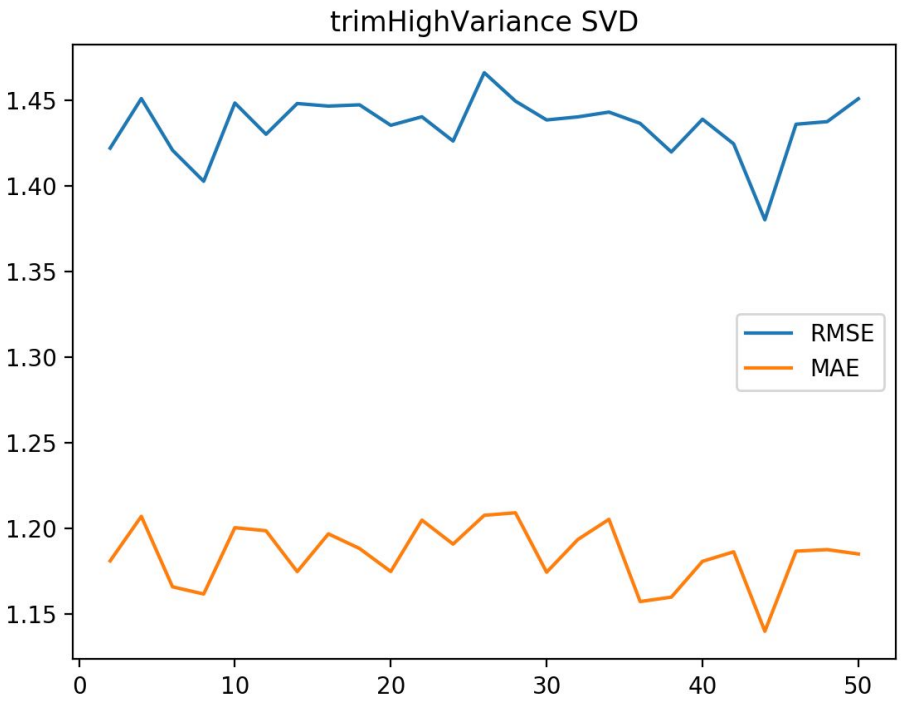
From one result of factoring the full rating matrix R into U and V, which are user latent matrix and movie latent matrix respectively, we sorted V by its first column and looked the top 10 movies. Resulted movieId and genres are shown: [[3685, 'Comedy|Drama|Romance'], [4316, 'Drama'], [2967, 'Drama|Thriller'], [3444, 'Action'], [3284, 'Comedy|Mystery|Romance'], [3926, 'Adventure|Sci-Fi'], [6591, 'Drama']]. As can be seen, these movies are mainly in the genre of comedy and drama. Sorting the second column of another resulted V yielded these results: [[851, 'Drama'], [3339, 'War'], [4136, 'Drama'], [4332, 'Crime|Drama|Thriller'], [7394, 'Action|Adventure|Comedy'], [704, 'Action|Adventure'], [2146, 'Drama|Romance']], which are movies mainly in the drama and action/adventure genres. Each column of the matrix V is a latent factor for movies, meaning that it represent a feature about movies that lead to the ratings received by the movies. A user that dislikes drama movies may generally give drama movies a low rating. If a population dislikes a certain movie genre, all movies in that genre can be receiving low ratings. In other words, movies' genres affect their rating, so do latent factors. One latent factor could represent how closely a movie belongs to a genre, or the opposite.

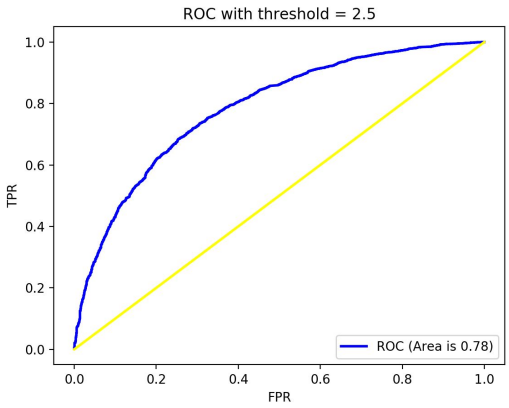
Q24-29
Q25:
Best k for MAE:40
Best k for RMSE: 5
Minimum average RMSE: 0.887
Minimum average MAE: 0.674

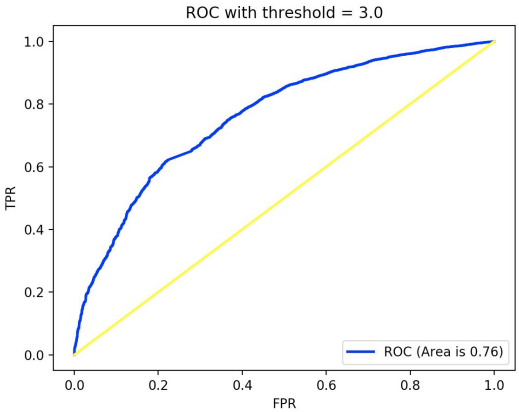

not trimed SVD
0.90
0.85
0.80
0.75
0.70
0
10
20
30
40
50
RMSE
MAE

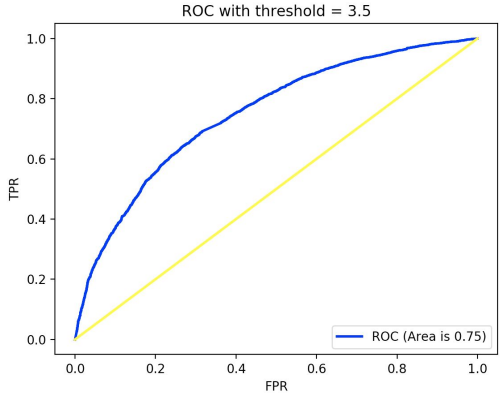

trimPopular SVD
0.90
0.85
0.80
0.75
0.70
0
10
20
30
40
50
RMSE
MAE

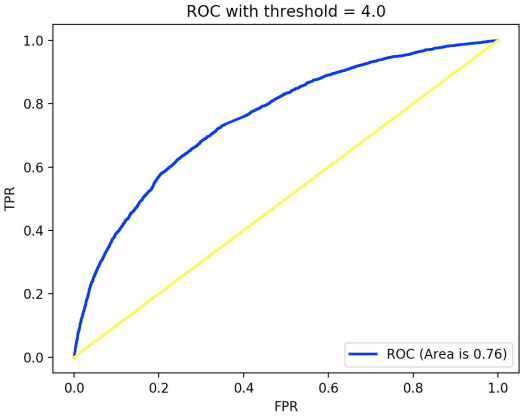

trimUnpopular SVD
1.00
0.95
0.90
0.85
0.80
0
10
20
30
40
50
RMSE
MAE


trimHighVariance SVD
1.45
1.40
1.35
1.30
1.25
1.20
1.15
0
10
20
30
40
50
RMSE
MAE









Q30

normal naive filter:
rmse: 0.963
mae: 0.750

Q31

trimPopular naive filter:
rmse: 0.953
mae: 0.745

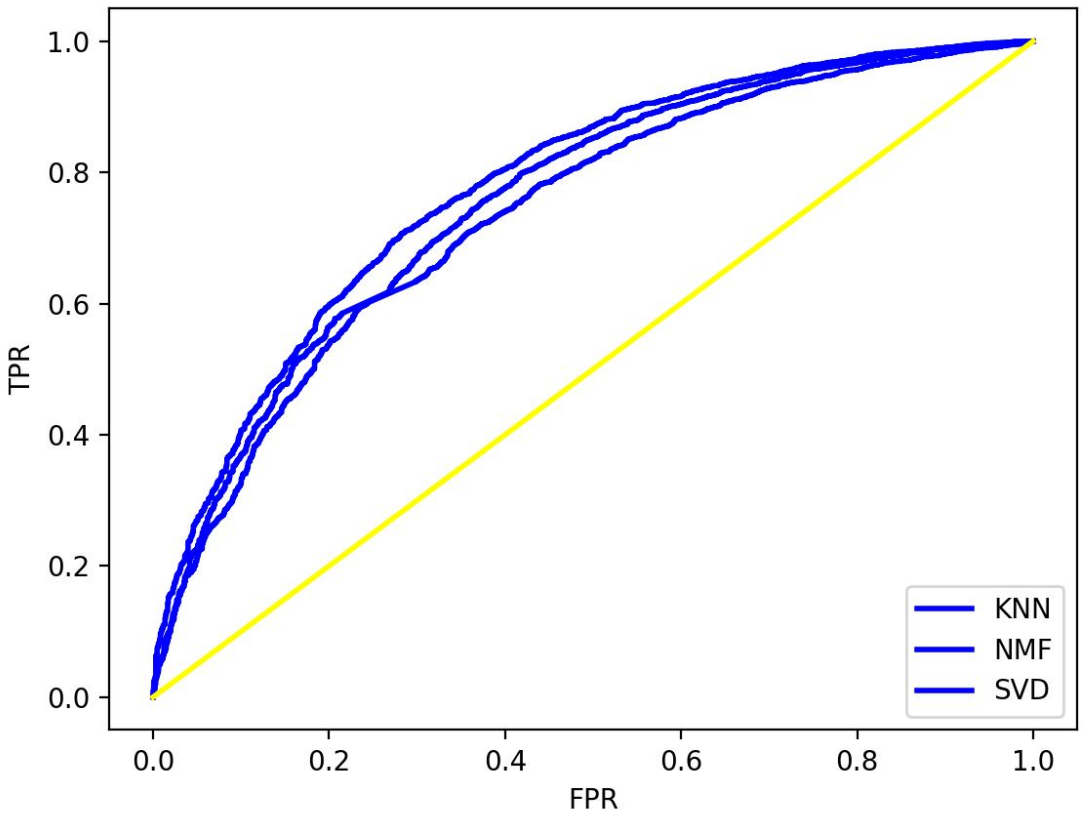
Q32

trimUnpopular naive filter:
rmse: 0.997
mae: 0.769

Q33

trimHighVariance naive filter:
rmse: 1.480
mae: 1.185

Q34:



Q35:

Precision: in the prediction how many are correct

Recall: how many of the ground truth are labels correctly

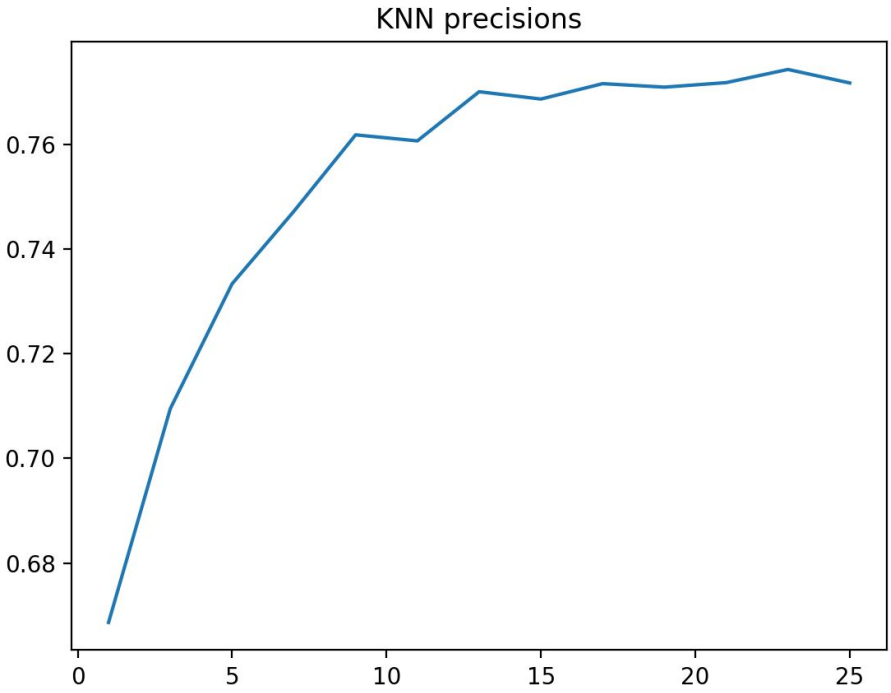
The formula of both are represented as below

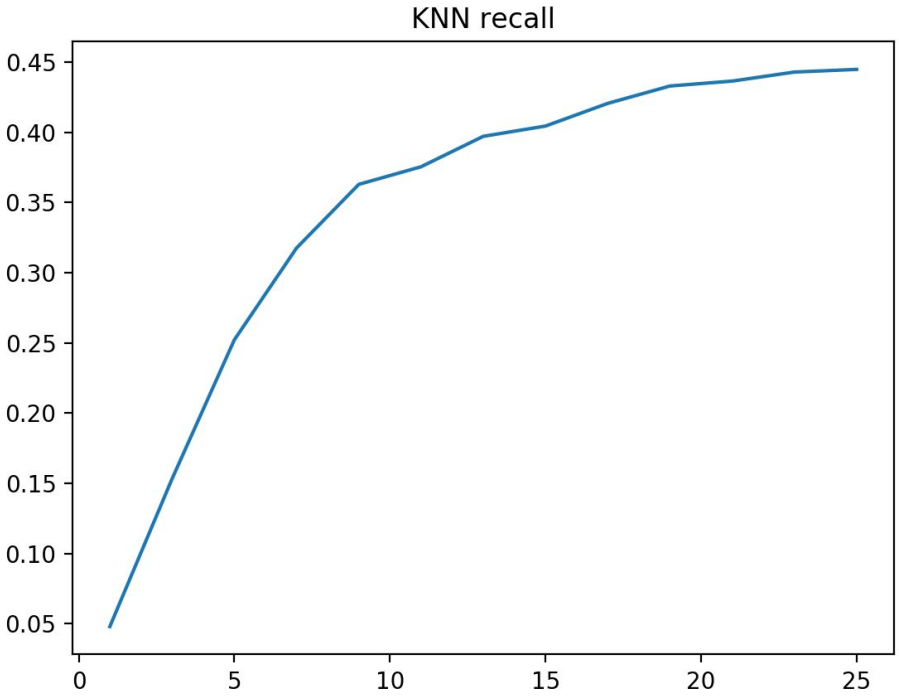
$$\text{Precision} = \frac{tp}{tp + fp}$$

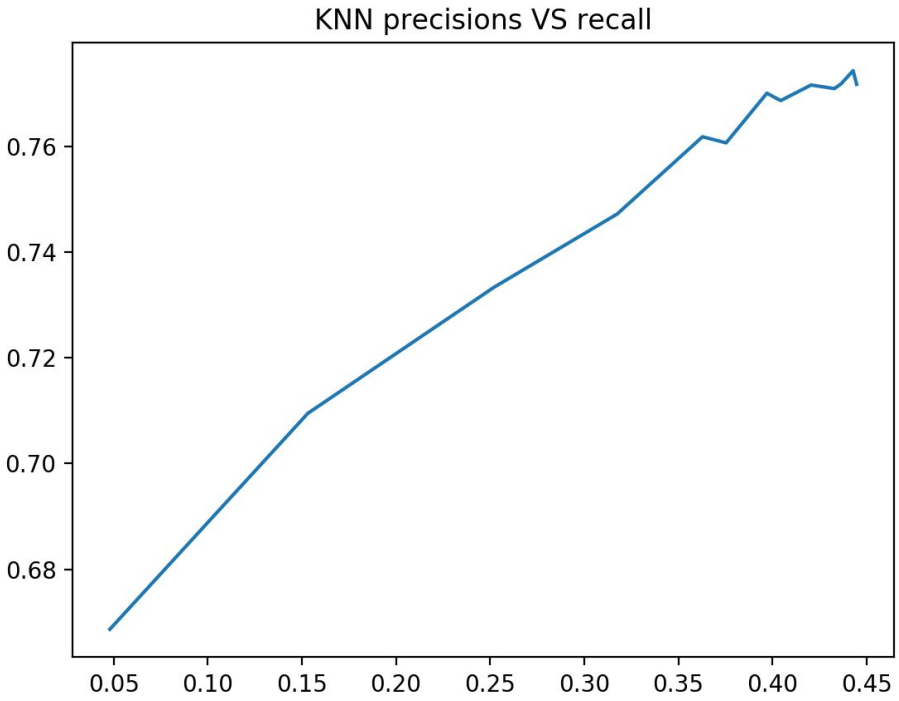
$$\text{Recall} = \frac{tp}{tp + fn}$$

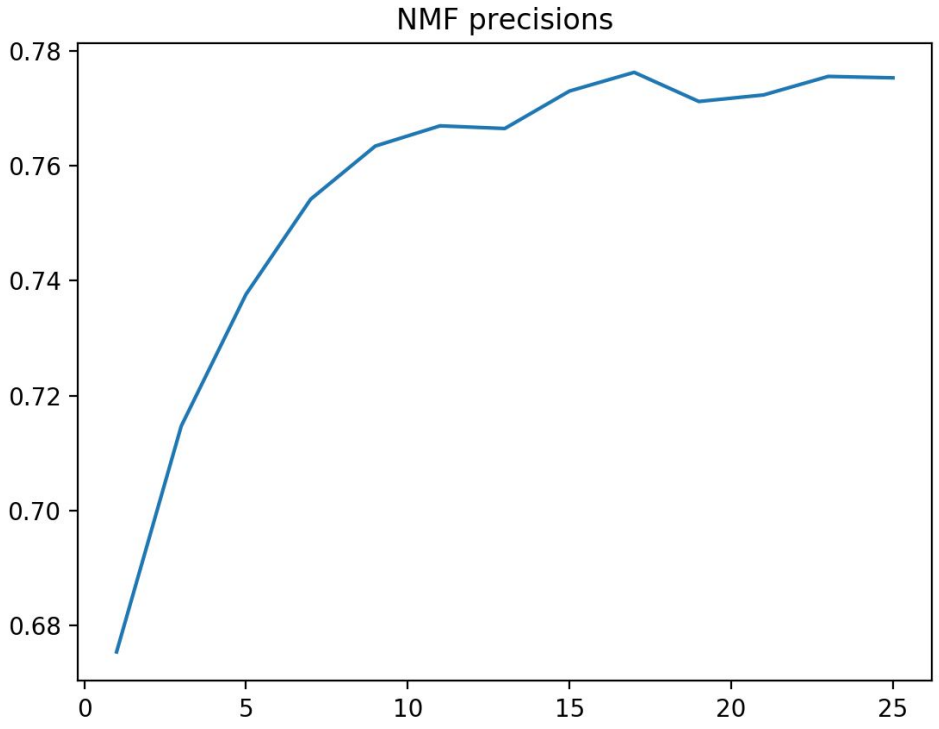
Q36-39:

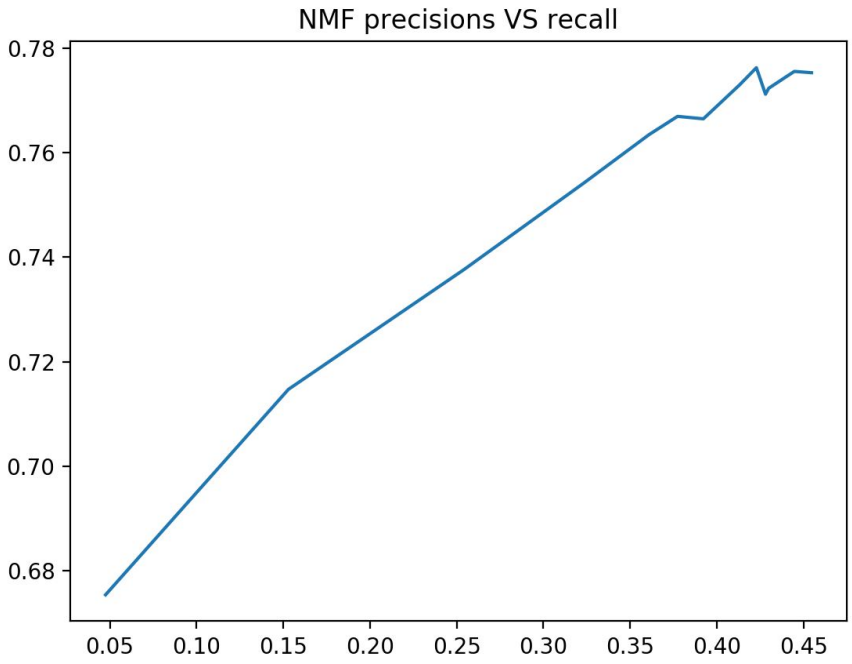
All of the shapes are increasing when t increase.

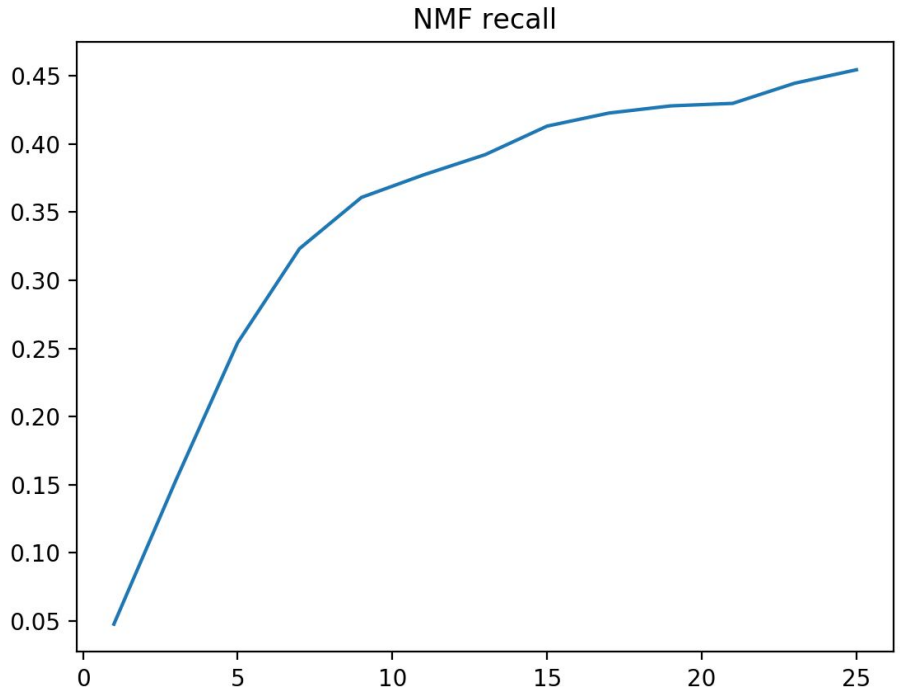

KNN precisions
0.76
0.74
0.72
0.70
0.68
0
5
10
15
20
25

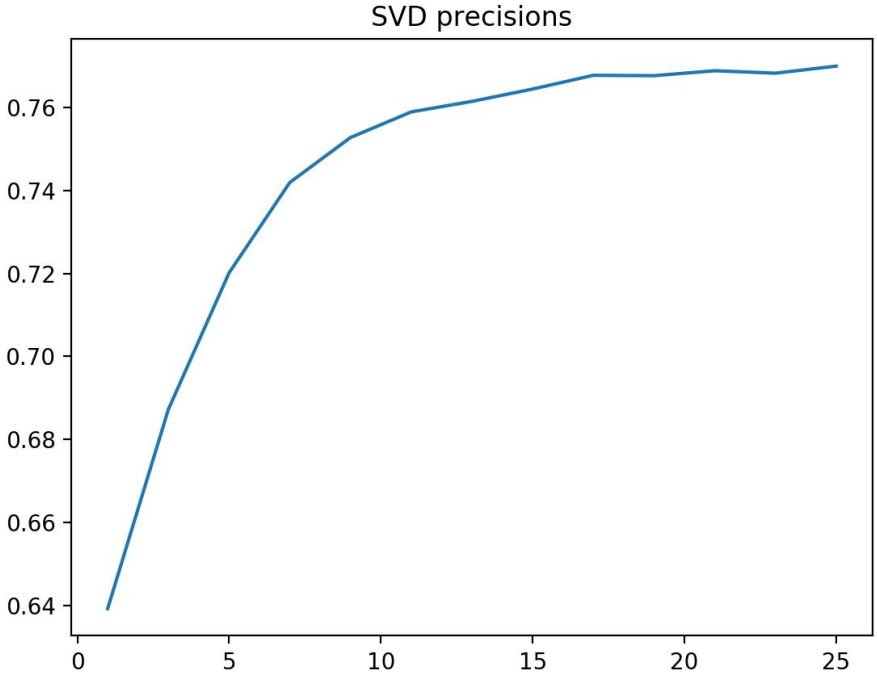

KNN recall
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0
5
10
15
20
25

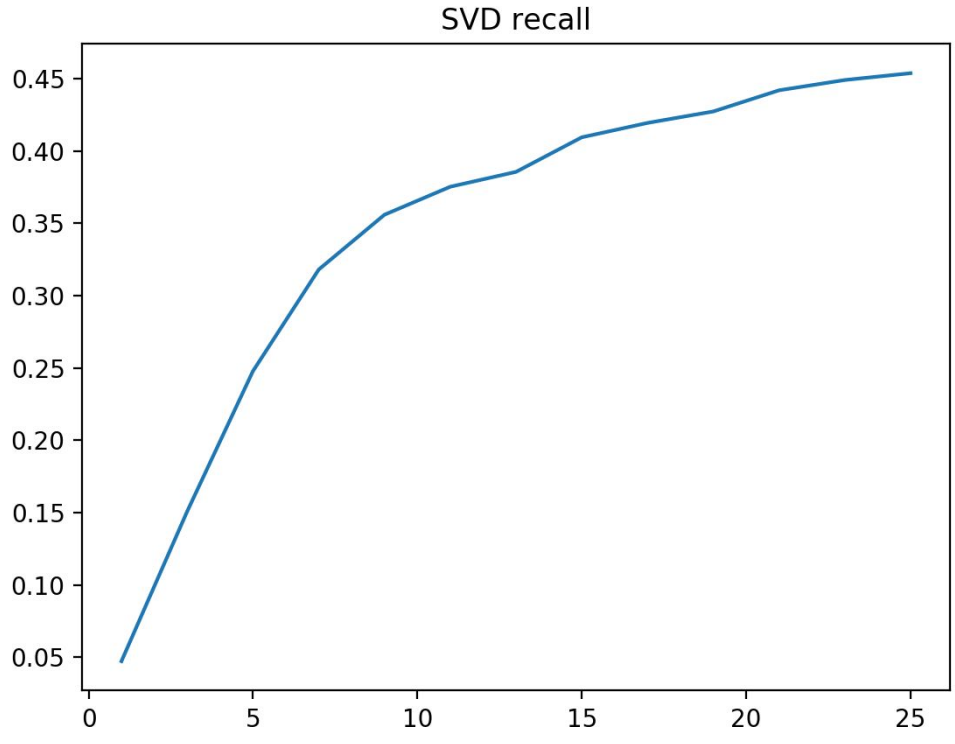

KNN precisions VS recall
0.76
0.74
0.72
0.70
0.68
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45

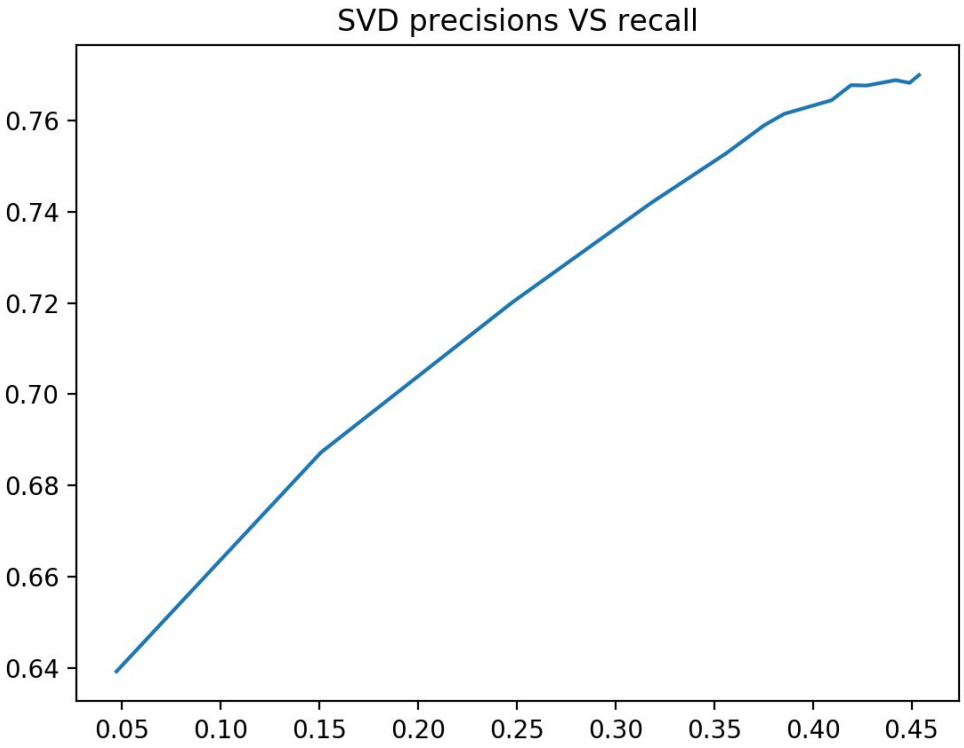

NMF precisions
0.78
0.76
0.74
0.72
0.70
0.68
0
5
10
15
20
25

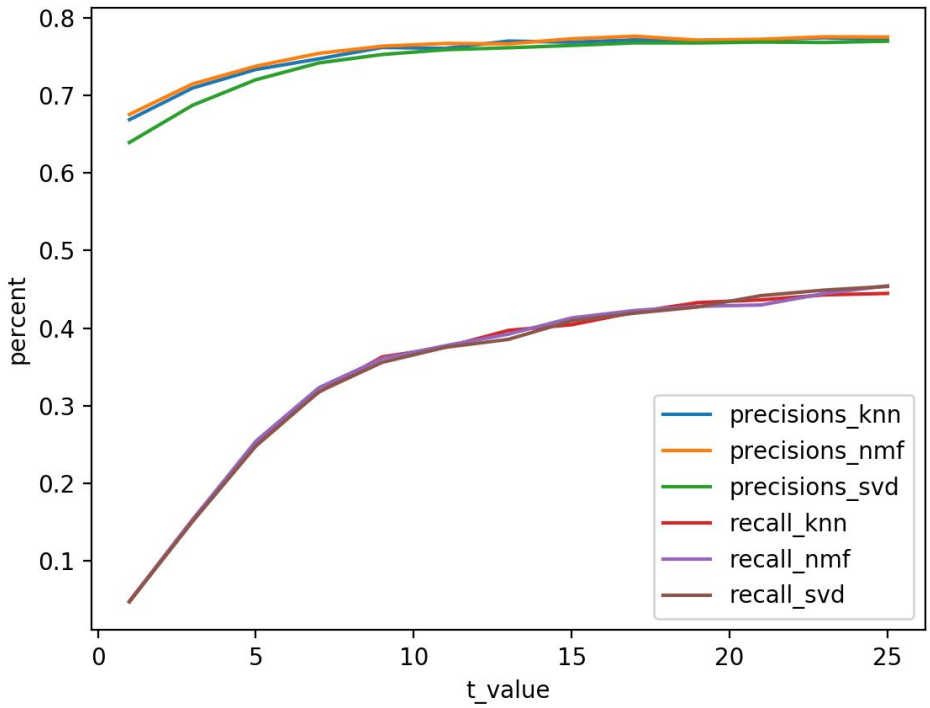

NMF precisions VS recall
0.78
0.76
0.74
0.72
0.70
0.68
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45


NMF recall
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0
5
10
15
20
25


SVD precisions
0.76
0.74
0.72
0.70
0.68
0.66
0.64
0
5
10
15
20
25


SVD recall
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0
5
10
15
20
25


SVD precisions VS recall
0.76
0.74
0.72
0.70
0.68
0.66
0.64
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45


percent
t_value
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
5
10
15
20
25
precisions_knn
precisions_nmf
precisions_svd
recall_knn
recall_nmf
recall_svd