**EE219 Project 3**

**Collaborative Filtering**

**Winter 2018**

**Jianfeng He (005025694)**

**Shouhan Gao (304944056)**

**ZhengXu Xia(104250792)**

**Tairan Zhu(605031908)**

**02/21/2018**

## Introduction and Problem Statement

The increasing importance of the web as a medium for electronic and business transactions has served as a driving force for the development of recommender systems technology. An important catalyst in this regard is the ease with which the web enables users to provide feedback about their likes or dislikes. The basic idea of recommender systems is to utilize these user data to infer customer interests.

The basic models for recommender systems works with two kinds of data: User-Item interactions such as ratings and attribute information about the users and items such as textual profiles or relevant keywords. Models that use first type data are referred to as collaborative filtering methods, whereas models that use second type data are referred to as content based methods. This project is a recommendation system using collaborative filtering methods.

**Solution**

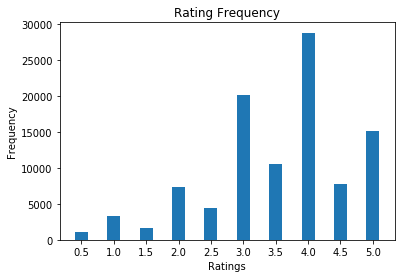
## Problem 1

Available Ratings: 100004

Possible Ratings: 6083286

Sparsity: 0.016439

## Problem 2

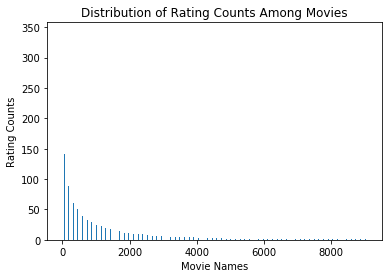


It can be concluded that most of ratings are in interval 3.0 - 5.0.

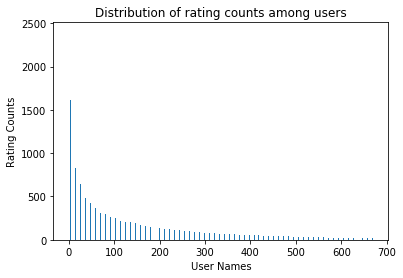
The rating with score 4 appears most frequent.

The rating with score 0.5 appears least frequent.

## Problem 3



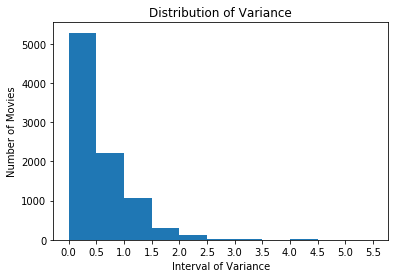
## Problem 4



## Problem 5

From the distribution of rating counts among users, we can observe that many people are most likely to have similar preferences to some kinds of movies, because for some specific sort of movies the rating counts are pretty high. This fact demonstrates that we can rate the movies collaboratively and recommend them to other users according to the same preferences.

## Problem 6

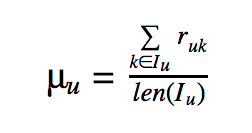


It can be concluded that most of movie have variance in rating between 0.0 - 2.5.

Also, the number of movies that have a variance between 0.0 - 0.5 is largest.

## Problem 7

the formula for μu in terms of Iu and ruk:



## Problem 8

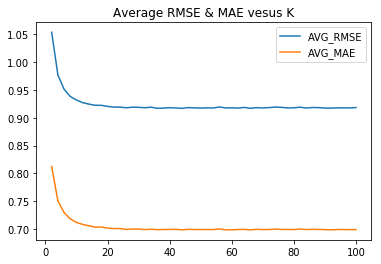
It means the indices of movies that both users have rated. Since Rating Matrix is sparse, it can be null if those two users rated completely different two sets of movies.

## Problem 9

If users' ratings on all items are always at one extreme, then the absolute rates can not reflect the true rates of items. Therefore, relative rates should be used in this case.

## Problem 10

The KNN-filter was used to predict the ratings of all movies in the dataset. The plot of different average RMSE & MAE for different number of neighbors (k) is shown below:



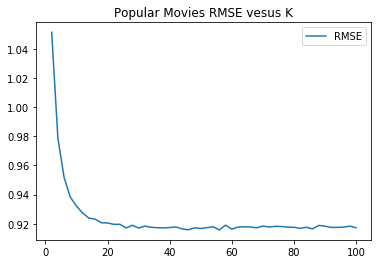
## Problem 11

It can be concluded that the slopes of both AVG\_RMSE and AVG\_MAE approach to zero around k=25.

The minimum k is about 25, where AVG\_RMSE is about 0.925 and AVG\_MAE is about 0.7.

## Problem 12

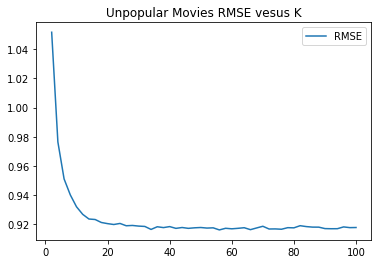
The performance of prediction on popular movies (more than 2 ratings) :



It can be concluded that the slopes of RMSE approach to zero around k=30.

The minimum k is about 30, where RMSE is about 0.92.

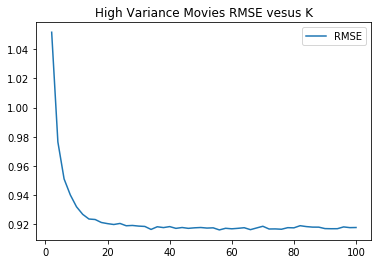
## Problem 13



It can be concluded that the slopes of RMSE approach to zero around k=35.

The minimum k is about 35, where RMSE is about 0.92.

## Problem 14

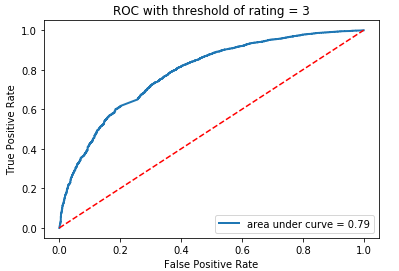
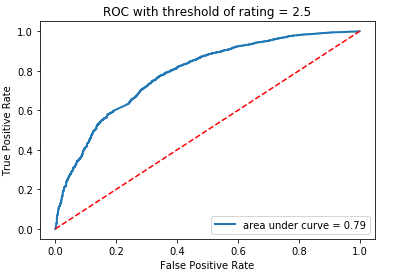


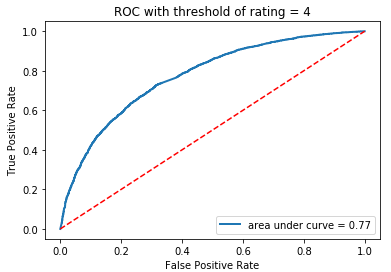
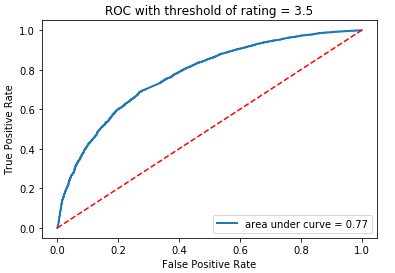
It can be concluded that the slopes of RMSE approach to zero around k=35.

The minimum k is about 35, where RMSE is about 0.92.

## Problem 15

Below are ROC curves for the k-NN collaborative filter for threshold values [2.5, 3, 3.5, 4]. The area reported are [0.79, 0.79, 0.77, 0.77] respectively.





## Problem 16

Yes. Equation 5 is a convex.

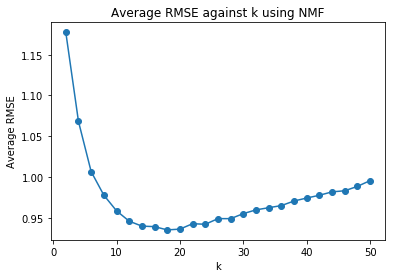
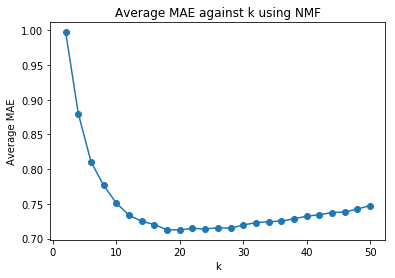
Let R' be a vector where contains all non-zero entry of rating matrix R with N \* 1 where N is the number of non-zero entries in rating matrix R.

Let V' be a vector which contains [V11 ... Vnk] with the size nk\*1.

Since U is fixed, U' can be factored as a vector which contains the coefficients of V' with size N\*nk.

Then, the problem becomes min||R'-U'V'||^2 and this is a least square problem and R' and U' are known and V' is unknown.

## Problem 17



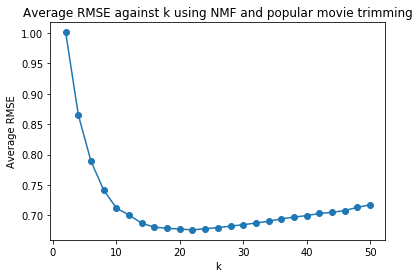
Minimum average MAE is 0.712406724893 and optimal k is 20.  
Minimum average RMSE is 0.935021873173 and optimal k is 18.

Both figures above have U shape plot of average error when number of latent factors becomes larger. The reason of large error when k is too small or too large is that latent factors fail to correctly reflect the internal relationship between two groups of data (in this case users and movies). When k reaches a value where minimum error happens, the number of latent factors can match the number of internal relationship of two groups of data.

## Problem 18

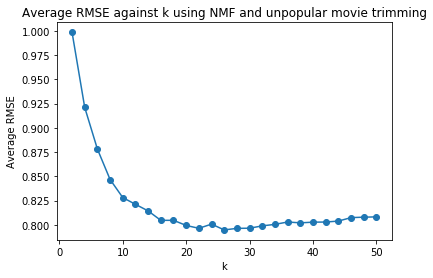
When k = 18, minimum average MAE is 0.713850751067.  
When k = 18, minimum average RMSE is 0.935598311477.  
The number of genres is 18. If including no genre, the number is 19. The optimal number of latent factors k, 18, is equal to the true number of genres.

## Problem 19



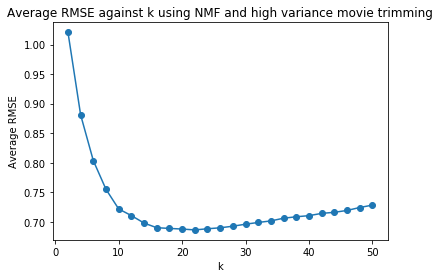
Minimum average rmse is 0.676283784209.

## Problem 20



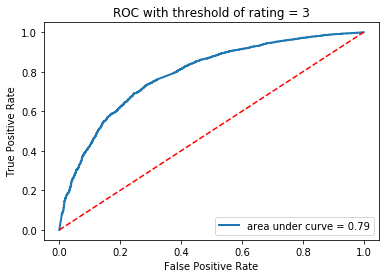
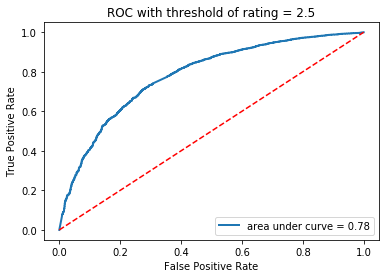
Minimum average rmse is 0.794827623088. After removing popular movies’ ratings, it is clear to see that the prediction error becomes larger. This is because the unpopular movies tend to be unpopular in both train set and test set. The prediction of ratings of unpopular rating lacks information from ground truth rating matrix.

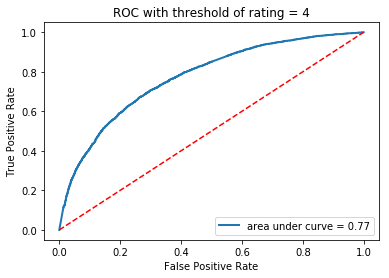
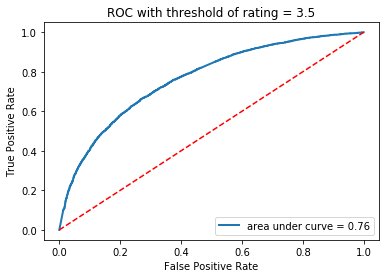
## Problem 21



Minimum average rmse is 0.686406804386

## Problem 22





It is clear to see that the area under ROC curve becomes smaller when preference threshold is too small or too large.

## Problem 23

Column 0  
Drama|Mystery|Romance  
Action|Comedy|Crime|Fantasy  
Comedy|Documentary  
Drama|War  
Children|Comedy  
Action|Adventure|Sci-Fi|War|IMAX  
Thriller  
Comedy|Western  
Action|Adventure|Sci-Fi|IMAX  
Drama  
Column 1  
Comedy|Documentary  
Adventure|Animation  
Musical  
Comedy|Romance  
Drama|Fantasy|Horror  
Comedy  
Drama|Sci-Fi  
Drama  
Adventure|Drama|Sci-Fi  
Drama|Romance  
  
  
  
Column 2  
Adventure|Drama|Fantasy|Romance  
Action|Adventure|Drama|Thriller  
Action|War  
Action|Crime|Thriller  
Comedy  
Comedy|Drama|Romance  
Comedy  
Comedy|Mystery|Thriller  
Drama  
Comedy

Column 3  
Drama  
Comedy|Crime  
Comedy|Drama  
Documentary|Drama  
Action  
Drama|Thriller  
Adventure|Children  
Drama|Romance  
Comedy  
Drama

Column 4  
Comedy|Crime|Mystery|Thriller  
Action|Comedy  
Documentary  
Action|Adventure|Sci-Fi|Thriller  
Comedy|Drama  
Drama|Romance  
Children|Comedy|Musical|Romance  
Comedy  
Documentary  
Comedy|Romance  
  
Column 5  
Action|Crime|Drama|Mystery|Thriller  
Comedy|Crime|Drama  
Horror  
Action|Sci-Fi|Thriller|IMAX  
Horror|Sci-Fi  
Comedy|Drama  
Action|Adventure|Fantasy|IMAX  
Action|Adventure|Sci-Fi|IMAX  
Action|Adventure|Comedy  
Drama|Romance|Thriller

Column 6  
Crime|Horror|Mystery|Thriller  
Adventure|Comedy|Thriller  
Adventure|Children  
Action|Sci-Fi  
Drama|Mystery|Thriller  
Horror  
Comedy|Drama|Romance  
Crime|Drama  
Comedy|Horror  
Horror

Column 7  
Action  
Drama  
Animation  
Thriller  
Drama|Sci-Fi  
Horror|Sci-Fi  
Comedy|Horror|Sci-Fi  
Horror  
Comedy|Musical  
Sci-Fi  
  
Column 8  
Comedy  
Action|Comedy  
Animation|Musical  
Action|Adventure|Fantasy  
Crime|Drama  
Action|Sci-Fi  
Adventure|Fantasy  
Drama|Horror|Thriller  
Drama  
Drama|Romance

Column 9  
Action|Drama|Thriller  
Comedy  
Crime|Drama|Mystery|Thriller  
Drama|Horror|Mystery  
Documentary|War  
Comedy  
Action|Comedy  
Adventure|Comedy|Fantasy|Sci-Fi  
Comedy|Drama  
Action|Comedy  
Column 10  
Drama  
Drama  
Drama  
Adventure|Children|Drama|Fantasy|IMAX  
Drama  
Crime|Drama  
Action|Adventure|Sci-Fi  
Drama  
Action|Comedy|Crime  
Comedy|Crime|Drama|Mystery|Romance  
  
Column 11  
Documentary  
Children|Comedy  
Drama  
Action|Drama|Thriller  
Drama|Mystery|Sci-Fi  
Horror  
Horror|Thriller  
Comedy|Romance  
Comedy  
Documentary

Column 12  
Drama|Romance  
Crime|Drama|Thriller  
Comedy|Crime|Musical  
Comedy|Horror|Sci-Fi  
Comedy|Romance  
Adventure|Drama|Sci-Fi  
Horror  
Comedy|Drama  
Crime|Drama|Mystery|Thriller  
Horror|Thriller  
  
  
Column 13  
Comedy  
Drama  
Action|Comedy|Drama  
Comedy  
Fantasy|Horror  
Children|Comedy  
Action|Sci-Fi  
Drama|Musical  
Comedy|Drama|Romance  
Drama  
  
  
  
Column 14  
Adventure|Drama|War|Western  
Musical  
Action|Sci-Fi  
Adventure|Children|Drama  
Comedy  
Horror  
Children|Comedy|Fantasy  
Comedy|Romance  
Horror|Thriller  
Comedy

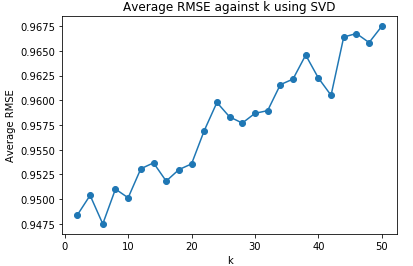
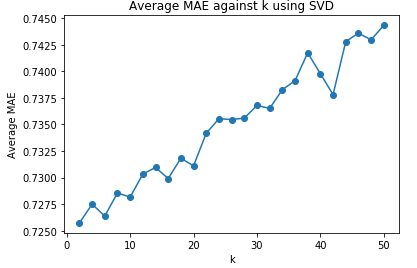
Column 15  
Action|Adventure|Animation|Crime|Fantasy  
Comedy|Romance  
Action|Comedy  
Animation|Comedy|Musical  
Comedy|Crime  
Horror|Mystery|Thriller  
Drama  
Comedy  
Comedy|Drama  
Comedy|Romance  
  
  
  
Column 16  
Horror|Sci-Fi|Thriller  
Drama  
Comedy|Drama|Romance  
Comedy  
Action|Adventure|Comedy|Fantasy  
Comedy  
Documentary  
Adventure|Children|Fantasy  
Comedy  
Adventure|Animation|Children|Sci-Fi|IMAX  
  
  
Column 17  
Comedy|Thriller  
Adventure|Animation|Children|Fantasy|Sci-Fi  
Drama  
Comedy  
Animation|Children|Musical  
Adventure|Drama  
Drama|Horror|Mystery|Thriller  
Comedy|Fantasy|Romance  
Drama|Romance  
Comedy|Western

Column 18  
Comedy|Drama  
Horror  
Comedy|Romance  
Action|Drama|Thriller  
Crime|Drama|Thriller  
Comedy|Drama|Musical  
Comedy|Drama  
Horror|Sci-Fi  
Comedy  
Comedy|Drama  
  
Column 19  
Action  
Comedy  
Comedy|Fantasy|Musical|Romance  
Comedy  
Horror|Sci-Fi  
Comedy|Romance  
Action|Adventure|Animation|Crime|Fantasy  
Adventure|Sci-Fi  
Adventure|Children|Fantasy  
Drama|Mystery|Romance

Viewing the genres of top 10 movies after sorting columns of item latent matrix V, we found that top 10 movies within a column tend to share the same set of genres. Since one movie can have multiple genres, even though they might not have one exactly the same genre, there still are internal relationships among them in terms of genres.

## Problem 24

MF with bias collaborative lter to predict the ratings of the movies in the MovieLens dataset

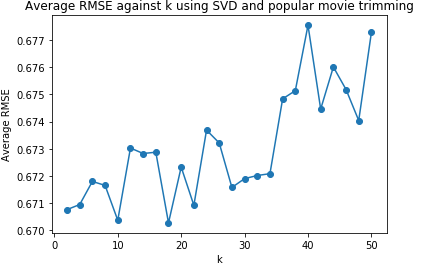


## Problem 25

When k = 6, it gives the minimum average MAE which is 0.727182512535.  
When k = 6, it gives the minimum average RMSE which is 0.948641321929.

## Problem 26

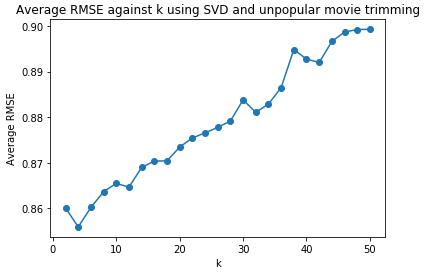
a MF with bias collaborative filter to predict the ratings of the movies in the popular movie trimmed test set.



Minimum average rmse is 0.670272206435 at around k = 18.

## Problem 27

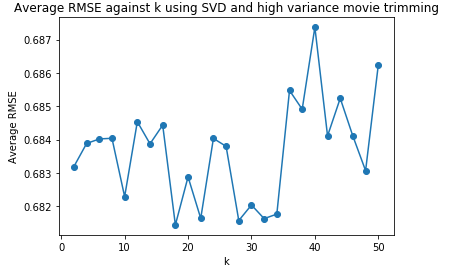
a MF with bias collaborative filter to predict the ratings of the movies in the unpopular movie trimmed test set.



Minimum average rmse is 0.855906892249 at around k = 4.

## Problem 28

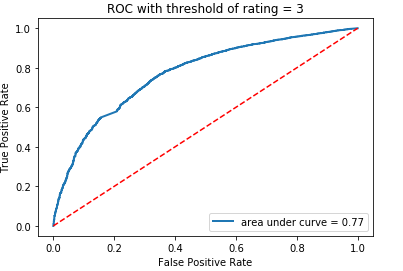
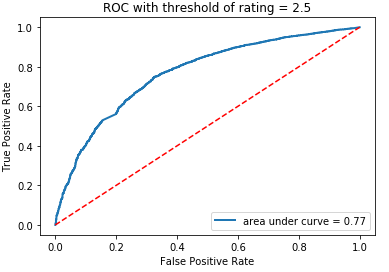
a MF with bias collaborative filter to predict the ratings of the movies in the high variance movie trimmed test set.

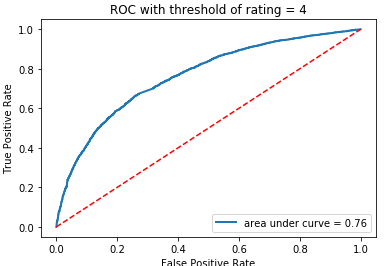
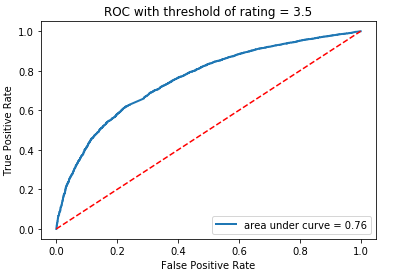


Minimum average rmse is 0.681440108235 at around k = 18.

## Problem 29

Below are ROC curves for theMF with bias collaborative filter for threshold values [2.5, 3, 3.5, 4]. The area reported are [0.77, 0.77, 0.76, 0.76] respectively.





**Problem 30 – 33:**

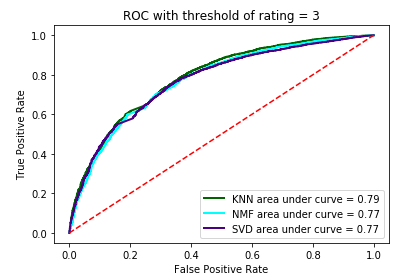
From problem 30 to 33, RMSE is calculated based on different dataset of movies. RMSE is 0.922 for the entire movie data. The value of RMSE is similar to the entire movie dataset for popular movies and high variance movie. However, the RMSE is very small comparing to the other three values.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | MovieLens | Popular Movies | Unpopular Movies | High Variance Movies |
| RMSE | 0.922 | 0.916 | 0.563 | 0.916 |

Table: RMSE for different dataset of movies

**Problem 34:**

ROC curves (threshold = 3) for the k-NN, NNMF, and MF with bias based collaborative filters.

****

According to the figure shown above, the three kinds of filters demonstrate an almost same performance at threshold = 3, because they have almost the same ROC curve and the area. But we should notice that, KNN collaborative filter has about 2% larger area than the other two filters. Thus, we can say that KNN collaborative filter has better performance than the other two.

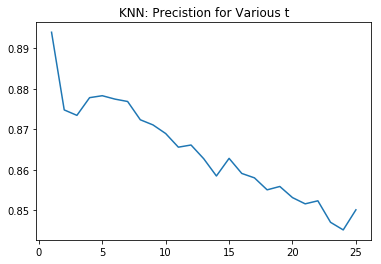
**Problem 35:**

Precision is the fraction of relevant elements over retrieved elements. For instance, precision in this case is the fraction of the recommended movies intersected with movies liked by target user over recommended movies

Recall is the fraction of relevant elements over total amount of elements. For instance, precision in this case is the fraction of the recommended movies intersected with movies liked by target user over all the movies liked by target user

**Problem 36:**

As the recommended items t increases, the precision for KNN filtering method decreases. In contrast, the recall increases as t increases. The plot of precision vs recall decreases as t increases. They are almost monotonic curves.

****

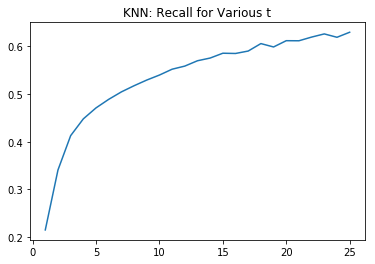
****

Figure: Precision for KNN Figure: Recall for KNN

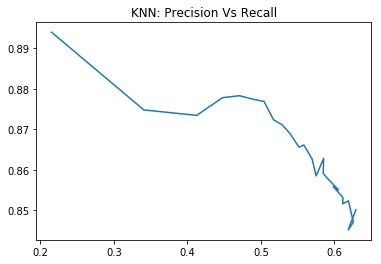
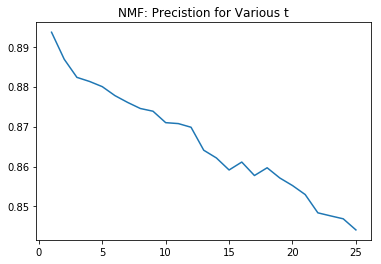
****

Figure: Precision Vs Recall for KNN

**Problem 37**

As the recommended items t increases, the precision for NMF filtering method decreases. In contrast, the recall increases as t increases. The plot of precision vs recall decreases as t increases. They are almost monotonic curves.

****

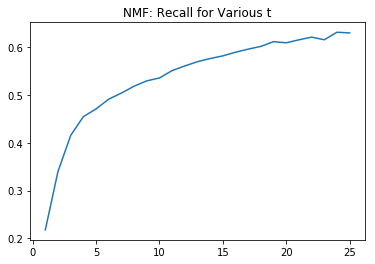
****

Figure: Precision for NMF Figure: Recall for NMF

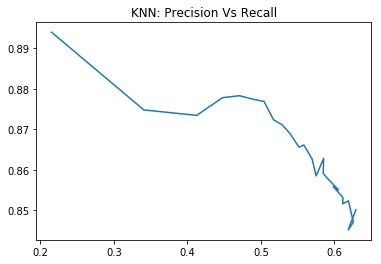
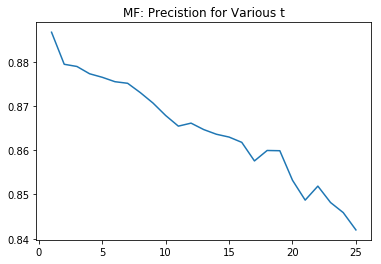
****

Figure: Precision Vs Recall for NMF

**Problem 38**

As the recommended items t increases, the precision for MF filtering method decreases. In contrast, the recall increases as t increases. The plot of precision vs recall decreases as t increases. They are almost monotonic curves.

****

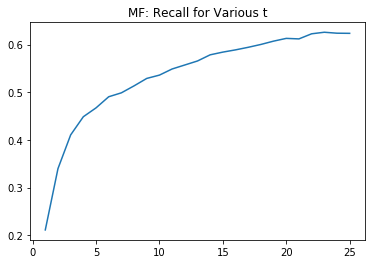
****

Figure: Precision for MF Figure: Recall for MF

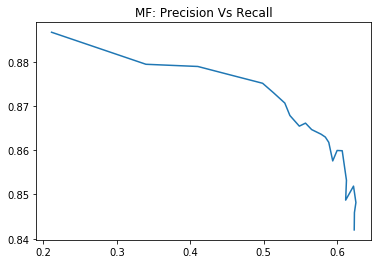
****

Figure: Precision Vs Recall for MF

**Problem 39:**

The precision for NMF is slightly greater than that of the other two filtering predictions. In addition, the shape of prediction vs recall for MF is smoother than the shape of KNN.

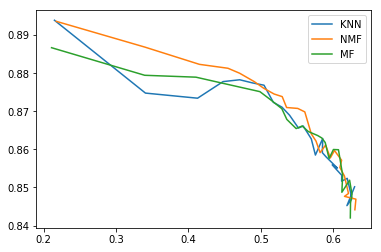


Figure: Precision Vs Recall among KNN, NMF, and MF