

# IT-562 Recommendation Systems and Engines

## Assignment-5

### Collaborative Filtering: Estimating SVD through Stochastic Gradient Descent

**Name:** Keya Desai

**ID:** 201501012

**Group Name:** Worthless Without Coffee

#### 1. Optimum parameters obtained for GoodBooks dataset:

Using GridSearchCV, we can obtain the best set of parameters. Given a dict of parameters, this class exhaustively tries all the combinations of parameters and reports the best parameters for any accuracy measure (averaged over the different splits).

RMSE = 0.886961380671

Parameters:

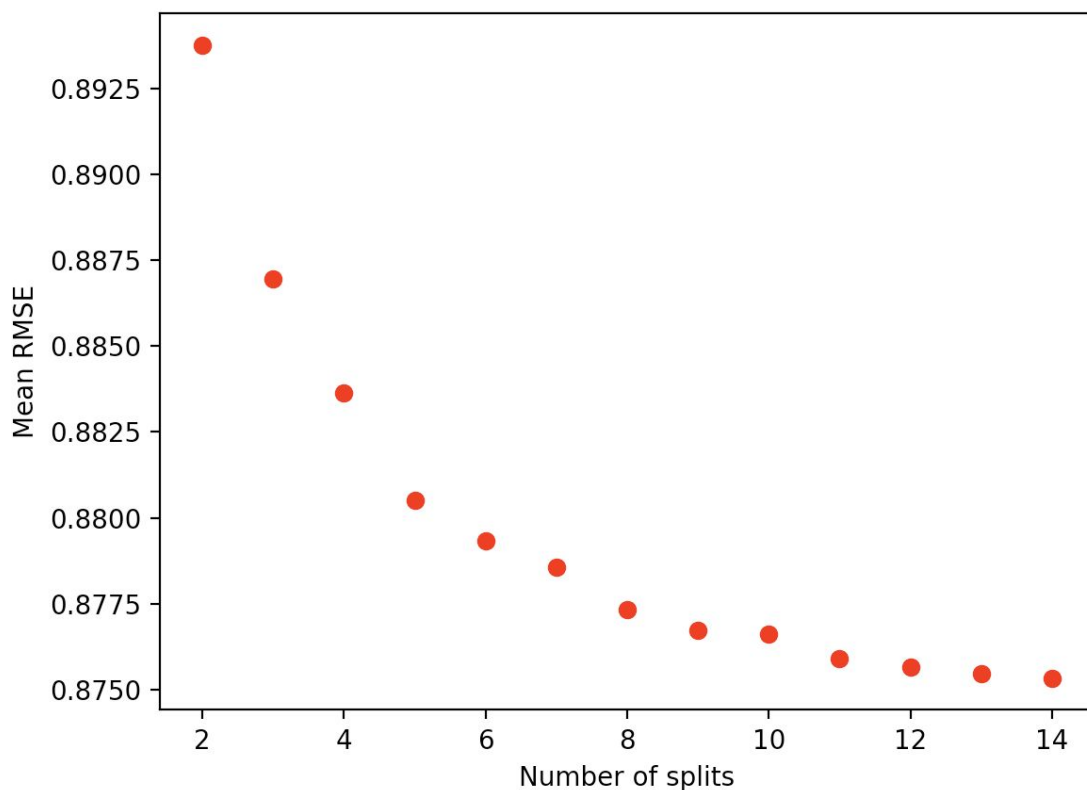
{'n\_epochs': 10, 'lr\_all': 0.005, 'reg\_all': 0.02}

## 2. Changing the number of splits:

### 1. RMSE accuracy:

Learning Rate = 0.005

Epochs = 10



RMSE Values:

[0.8937528736071223, 0.88694801840558579, 0.88363353707226033, 0.88050485821037339, 0.87934449654045721, 0.87857425711353088, 0.87732663688246304, 0.87672474716644566, 0.87662690917829378, 0.87590407333195064, 0.87567070829652749, 0.87547187888627565, 0.87534035786231634]

## 2. Time:

Epochs = 10

Learning rate = 0.005



Time( in s):

[6.396276,  
10.733293999999999,14.552478999999998,19.028800999999994  
,24.821213999999998, 28.939272000000003, 33.877045999999999,  
38.961564999999998 ,45.191922999999974, 49.74811999999997  
54.278856000000002, 60.066863000000001, 67.533452000000001]

### Observation:

- Accuracy increases i.e. RMSE decreases as the number of splits/folds increase.
- As the number of splits/folds increases the time taken to run SGD increases almost linearly.
- After 3 splits, the decrease in RMSE error is not significant but the time taken is significantly more. Hence due to the tradeoff between time and accuracy we have used 3 splits in all further steps.

### 3. Changing the number of Users

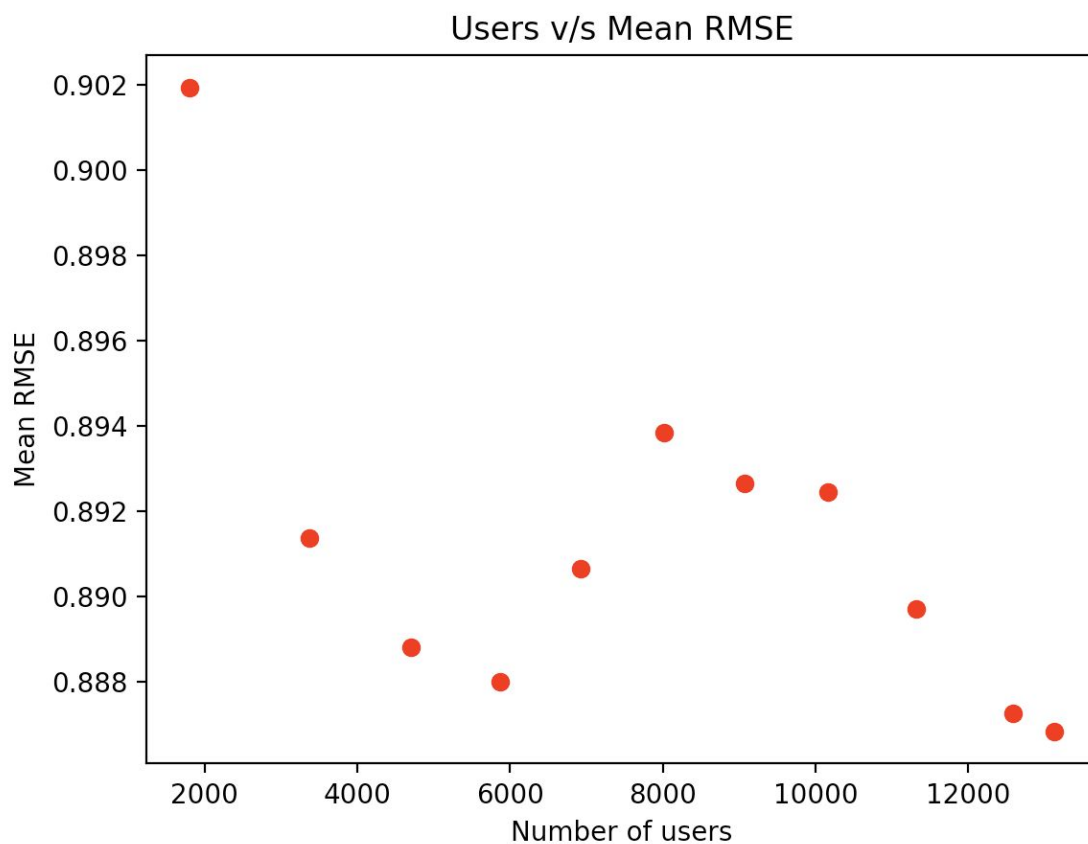
Total Users: 13123

Increasing the number of users by adding next 1L rows from user-rating matrix at a time, we get the following set of number of users:

[1806, 3374, 4698, 5876, 6923, 8012, 9065, 10161, 11318, 12586, 13123]

#### 1. RMSE

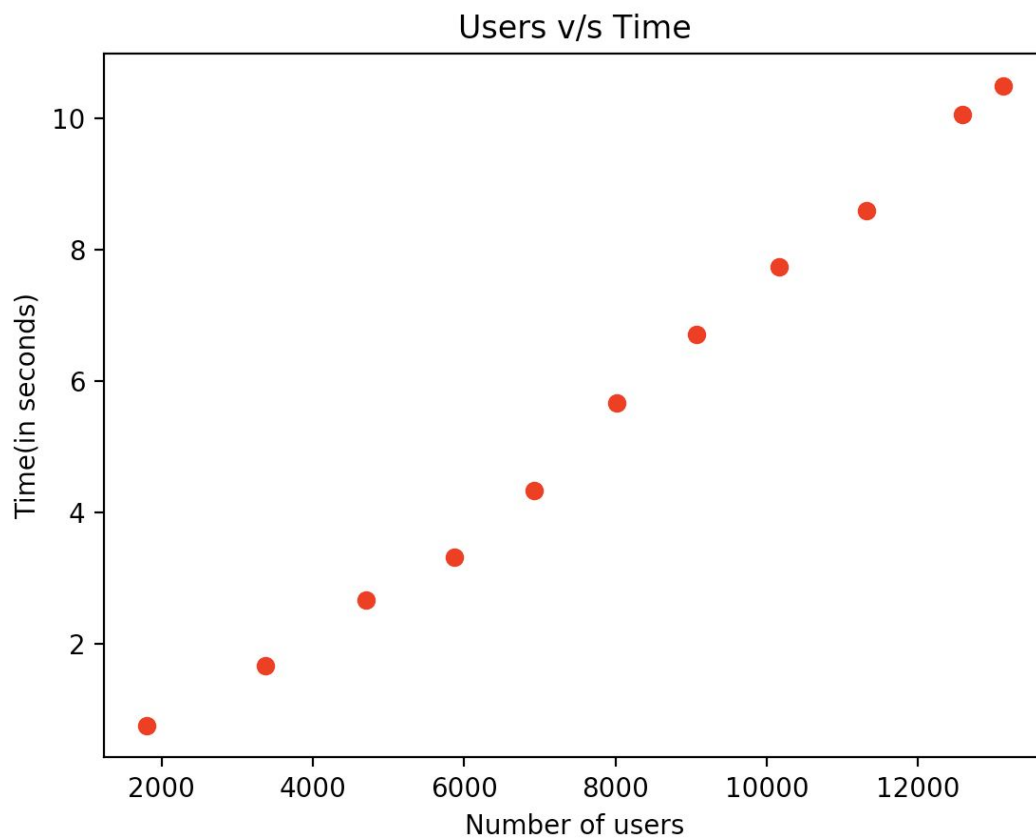
---



RMSE Values:

[0.90194789636887374, 0.8913601628063188, 0.88880639995218058,  
0.88798858998567776, 0.89064997057943696, 0.89384676154995646,  
0.89264289479855075, 0.89244414467015087, 0.88969756915044862,  
0.88724054275037079, 0.88682390608117745]

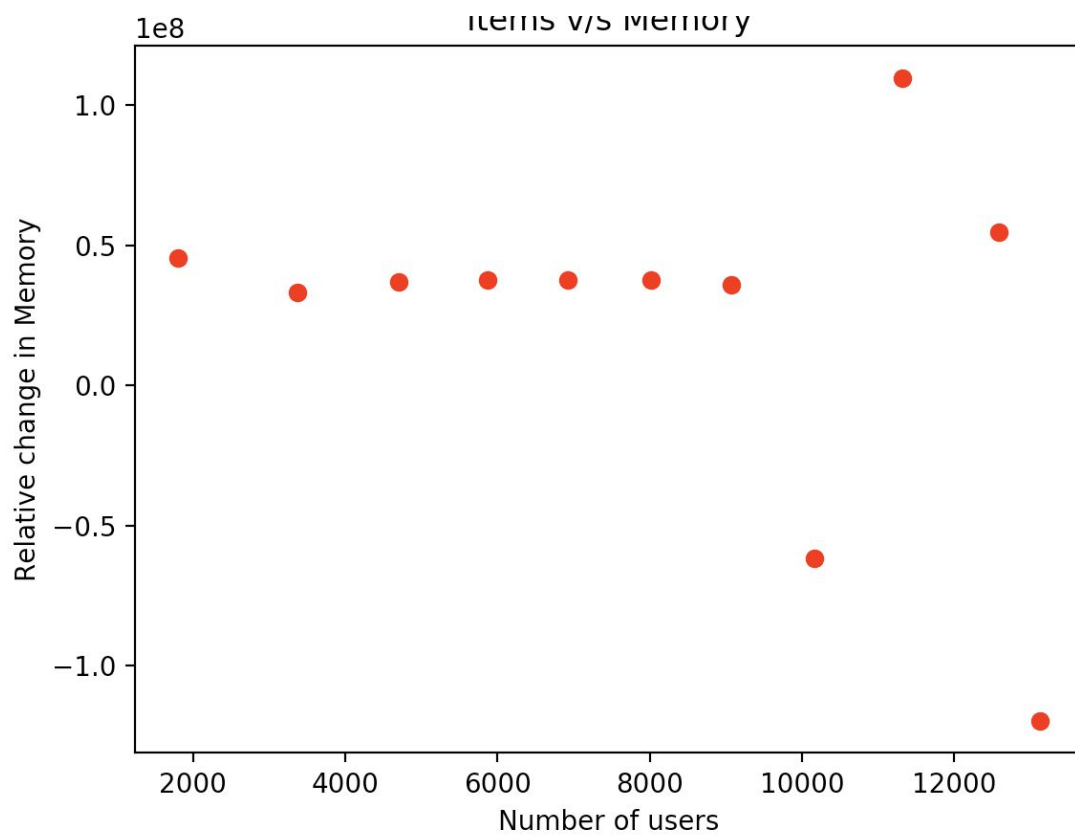
## 2. Time

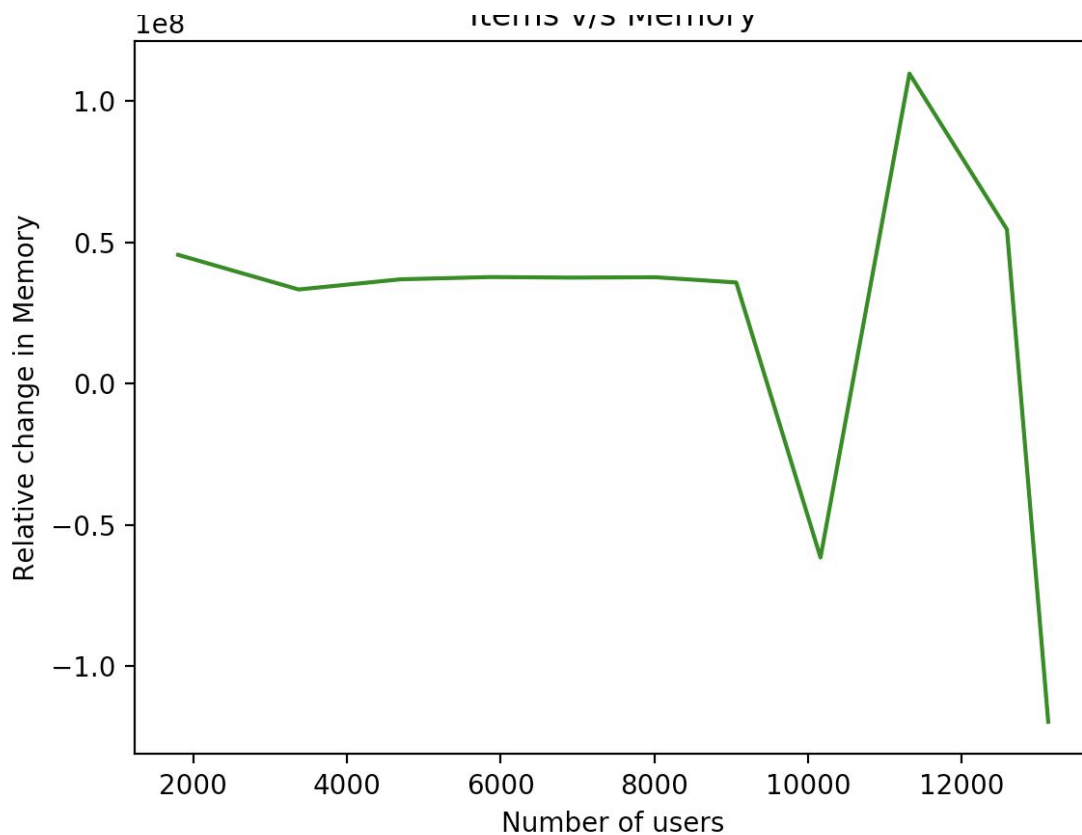


Time Values:

[0.7564719999999998, 1.672616, 2.670185, 3.3192300000000001,  
4.330952, 5.663919, 6.7118390000000005, 7.750394,  
8.6007670000000005, 10.0610709999999991, 10.501936999999998]

### 3. Memory





Memory Values:

[226791424, 267415552, 224571392, 342929408, 313212928, 275554304]

Observations:

- RMSE value is decreasing on giving more number of user ratings except for a few exceptions. Hence the accuracy is increasing.
- Again, the time is almost linearly increasing on increasing the size of the dataset.



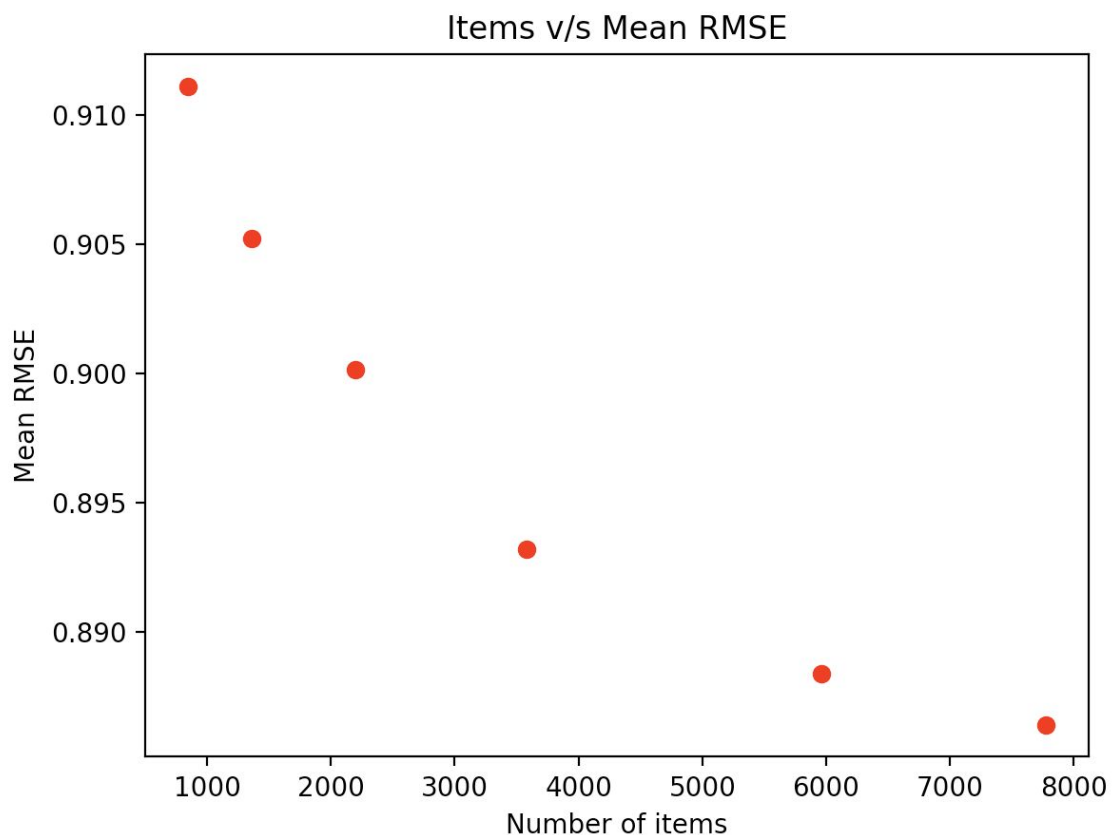
#### 4. Changing the number of items:

Total Items: 7774

Increasing the number of items by adding next 1L rows (starting from 6L upto the end) from user-rating matrix which is sorted in ascending order according to item no. at a time, we get the following set of number of items:

[846, 1360, 2201, 3584, 5964, 7774]

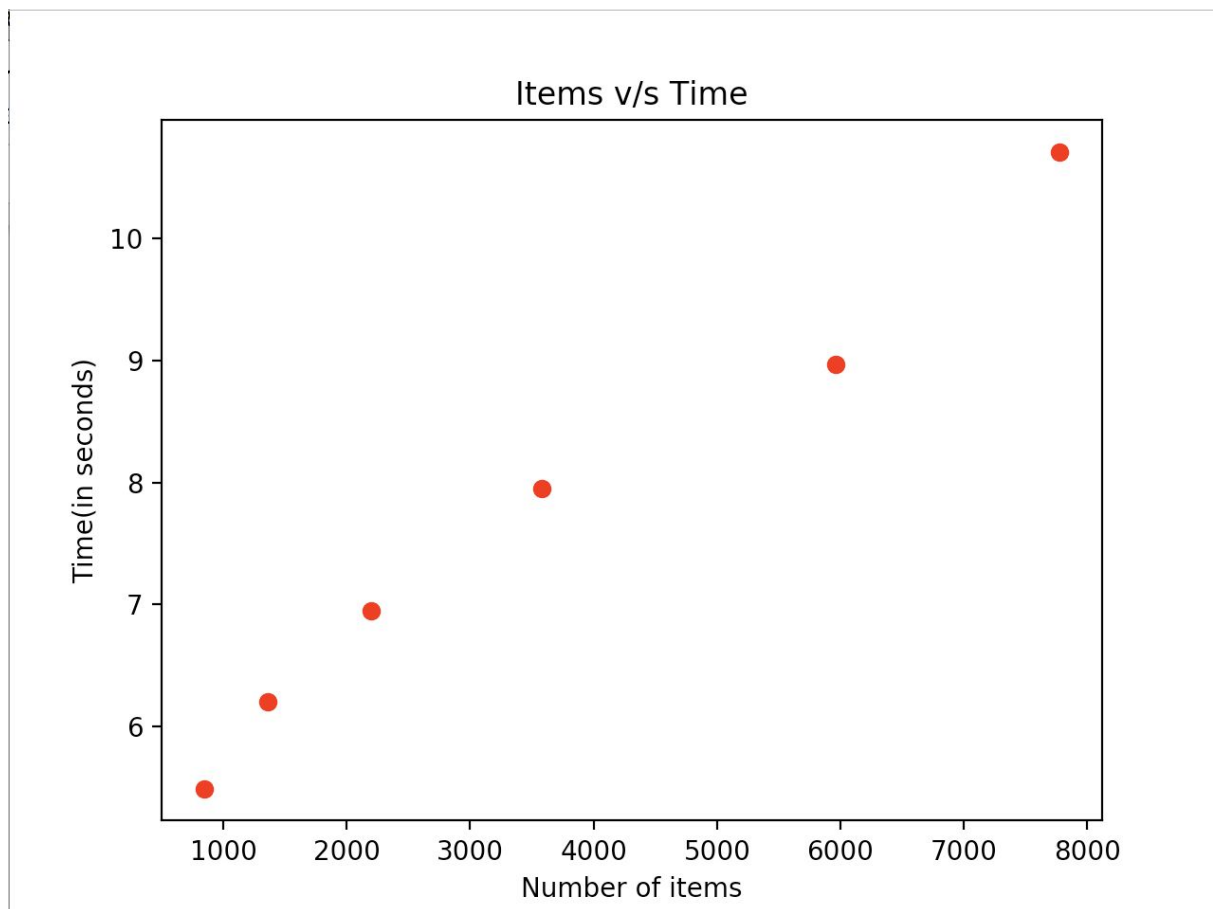
##### 1. RMSE:



RMSE Values:

[0.91112207163097203, 0.90522555689103479, 0.90015257008217942,  
0.89319559712390684, 0.88836846403409264, 0.88639420560383175]

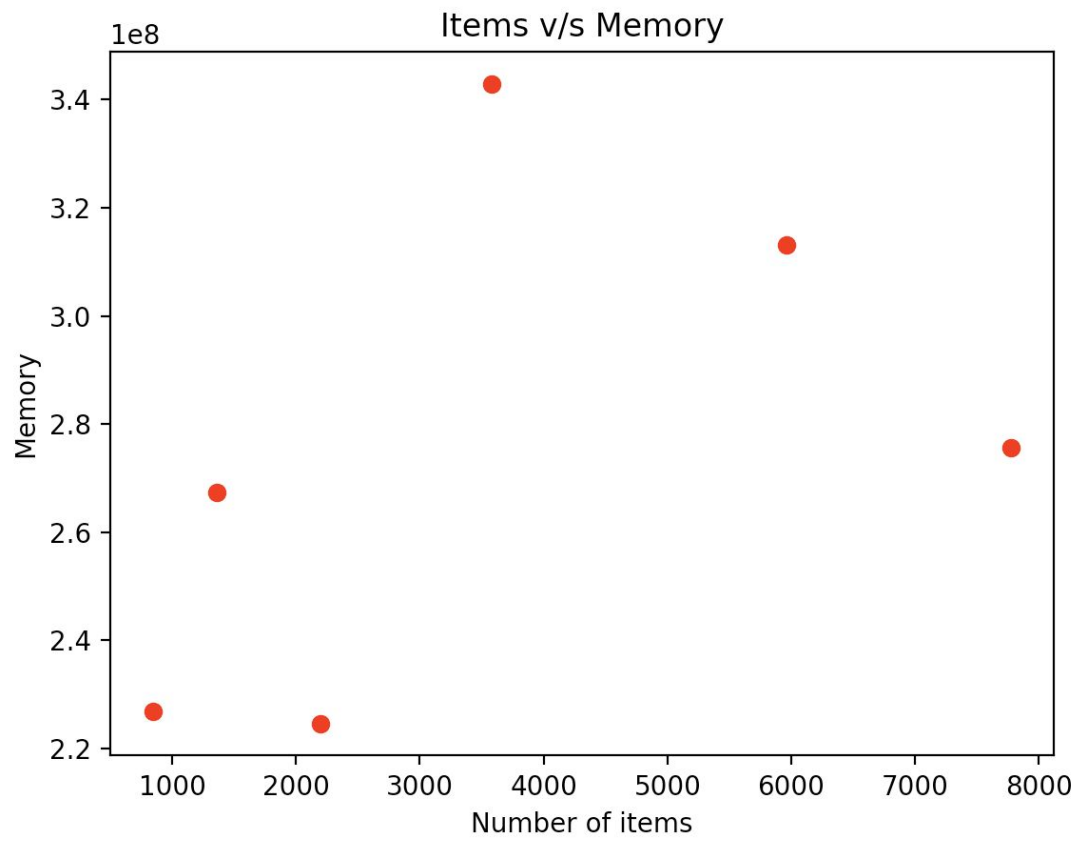
## 2. Time:

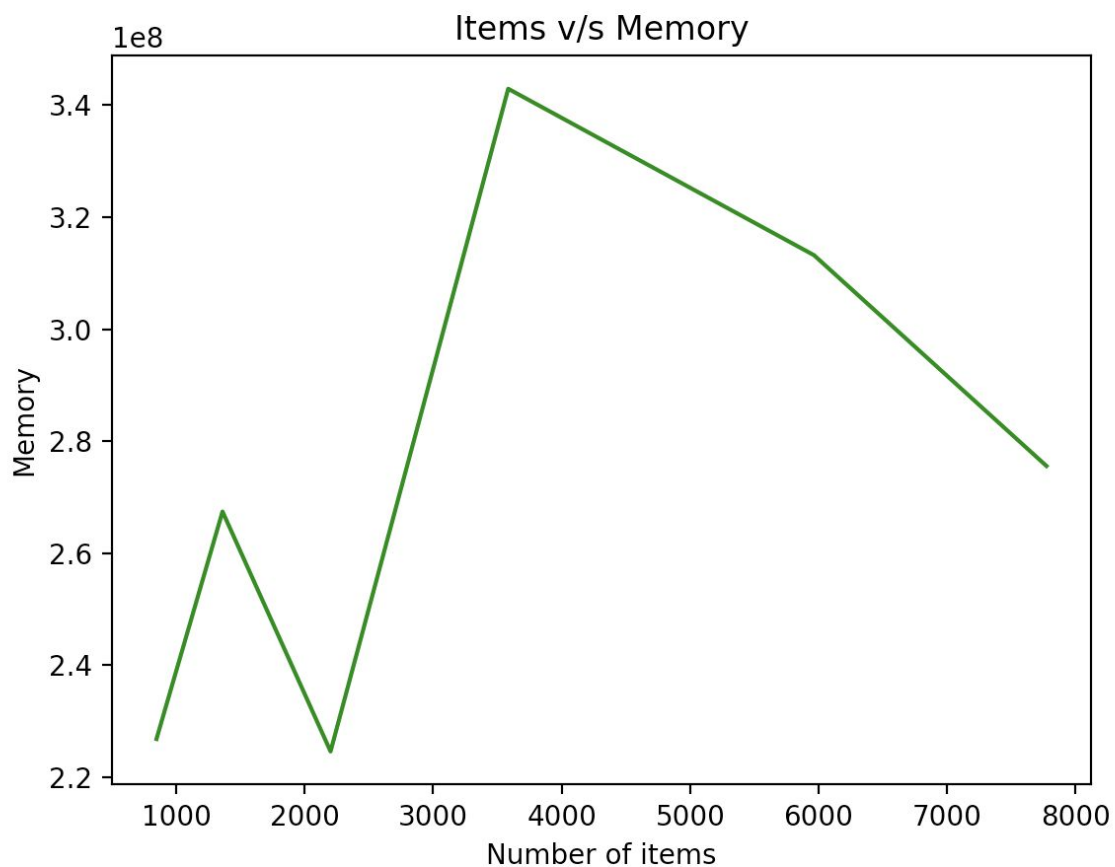


Time Values:

[5.491403999999999, 6.204091, 6.9489730000000002,  
7.947951999999997, 8.966673, 10.708690999999995]

Memory:





#### Observations:

- Unlike in the case of increasing the number of users, accuracy is constantly increasing on changing the number of items without any exceptions.
- Time is again increasing almost linearly on increasing the size of the dataset.
- Theoretically, memory consumption should increase on increasing the size of the dataset(i.e the number of rows) but based on the above shown results, no generalization can be made.