Data Mining:Assignment -1 Report

SASANK CS21BTECH1019
ABHINAY CS21BTECH11055
KARTHIK CS21BTECH11030
YAGNESH CS21BTECH11003

September 15, 2024

Contents

1 Question 1 **2**
1.1 Top 3 Answerers 2
1.2 Top 3 Tags 2

2 Expert Matrix **2**

3 Question 3 **3**
3.1 Metric of utility matrix 3
3.2 Metric of training and test data 4

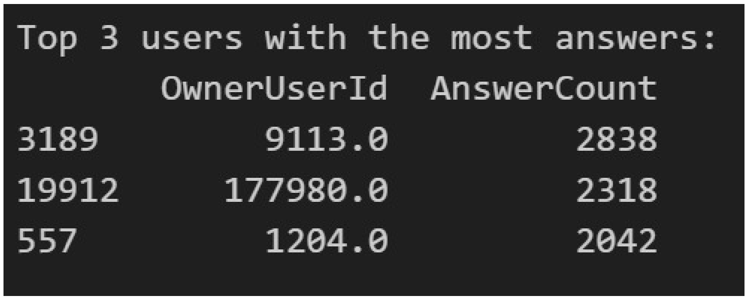
4 Question 4 **4**

5 Question 5 **4**

6 Question 6 **5**
6.1 Collaborative Recommendation 5
6.2 Matrix Factorization Recommendation 5

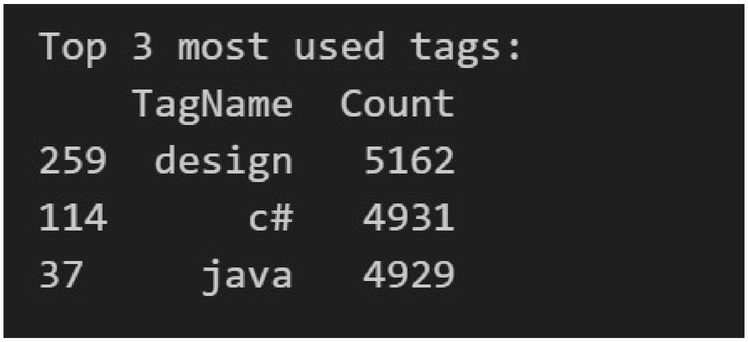
1 Question 1

1.1 Top 3 Answerers



```
Top 3 users with the most answers:
       OwnerUserId  AnswerCount
3189        9113.0         2838
19912     177980.0         2318
557         1204.0         2042
```

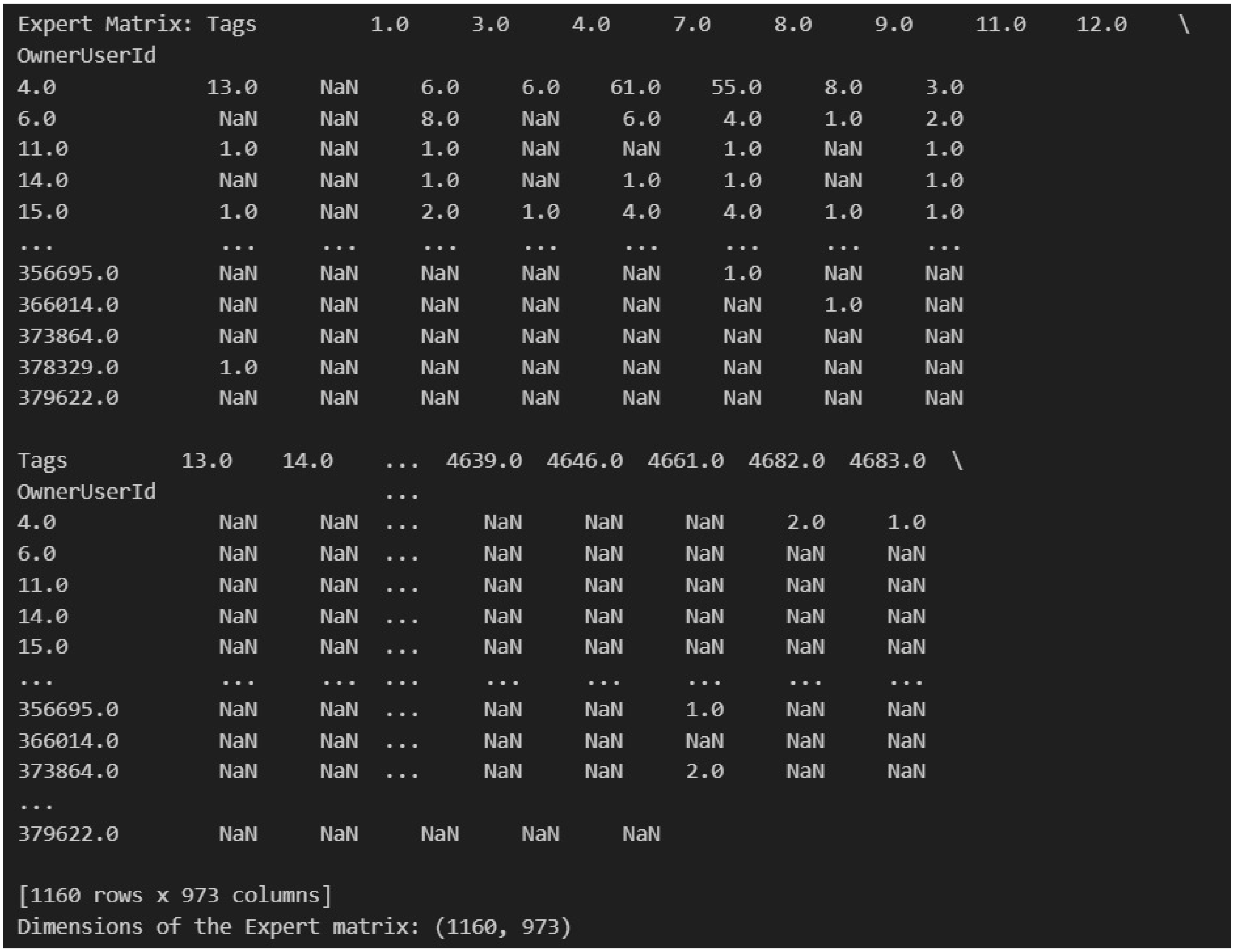
1.2 Top 3 Tags



```
Top 3 most used tags:
    TagName  Count
259  design   5162
114      c#   4931
37     java   4929
```

2 Expert Matrix

- The dimensions of Expert Matrix : $(1160, 973)$
- Total Number of tags with count >= 20 is 974. But One tag was never present in any of the questions answered by any qualified answerer. This tag is of no use to recommend a question to any expert user. So we have removed that column from the Expert matrix



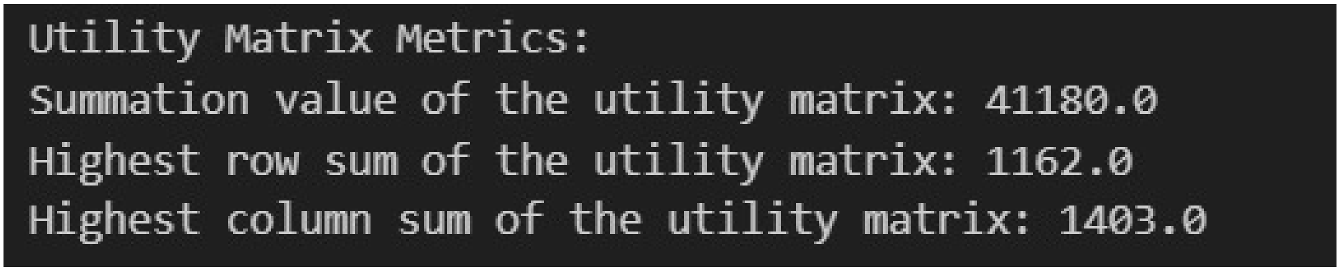
```
Expert Matrix: Tags         1.0     3.0     4.0     7.0     8.0     9.0     11.0    12.0    \
OwnerUserId
4.0            13.0     NaN     6.0     6.0    61.0    55.0     8.0     3.0
6.0             NaN     NaN     8.0     NaN     6.0     4.0     1.0     2.0
11.0            1.0     NaN     1.0     NaN     NaN     1.0     NaN     1.0
14.0            NaN     NaN     1.0     NaN     1.0     1.0     NaN     1.0
15.0            1.0     NaN     2.0     1.0     4.0     4.0     1.0     1.0
...             ...     ...     ...     ...     ...     ...     ...     ...
356695.0        NaN     NaN     NaN     NaN     NaN     1.0     NaN     NaN
366014.0        NaN     NaN     NaN     NaN     NaN     NaN     1.0     NaN
373864.0        NaN     NaN     NaN     NaN     NaN     NaN     NaN     NaN
378329.0        1.0     NaN     NaN     NaN     NaN     NaN     NaN     NaN
379622.0        NaN     NaN     NaN     NaN     NaN     NaN     NaN     NaN

Tags         13.0    14.0    ...  4639.0  4646.0  4661.0  4682.0  4683.0  \
OwnerUserId                  ...
4.0             NaN     NaN  ...     NaN     NaN     NaN     2.0     1.0
6.0             NaN     NaN  ...     NaN     NaN     NaN     NaN     NaN
11.0            NaN     NaN  ...     NaN     NaN     NaN     NaN     NaN
14.0            NaN     NaN  ...     NaN     NaN     NaN     NaN     NaN
15.0            NaN     NaN  ...     NaN     NaN     NaN     NaN     NaN
...             ...     ...  ...     ...     ...     ...     ...     ...
356695.0        NaN     NaN  ...     NaN     NaN     1.0     NaN     NaN
366014.0        NaN     NaN  ...     NaN     NaN     NaN     NaN     NaN
373864.0        NaN     NaN  ...     NaN     NaN     2.0     NaN     NaN
...
379622.0        NaN     NaN     NaN     NaN     NaN

[1160 rows x 973 columns]
Dimensions of the Expert matrix: (1160, 973)
```

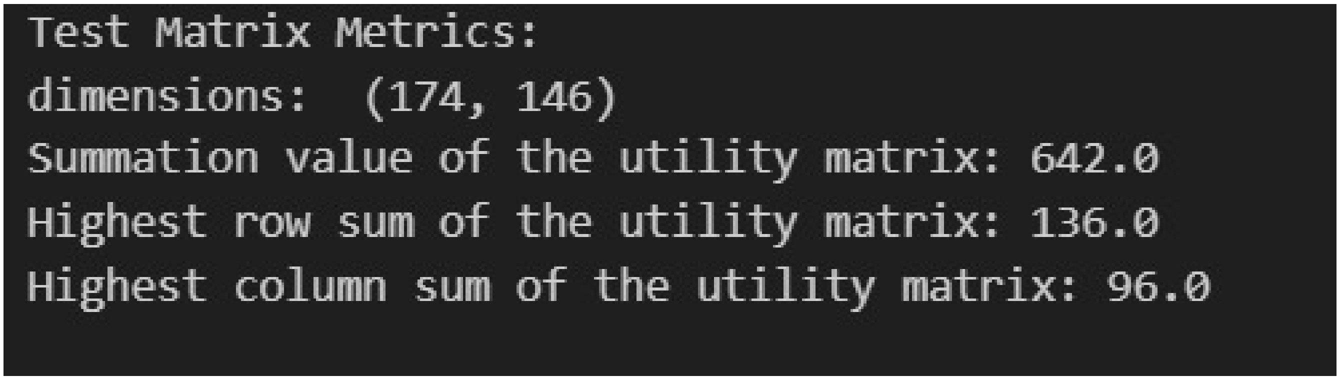
3 Question 3

3.1 Metric of utility matrix



```
Utility Matrix Metrics:
Summation value of the utility matrix: 41180.0
Highest row sum of the utility matrix: 1162.0
Highest column sum of the utility matrix: 1403.0
```

3.2 Metric of training and test data



```
Test Matrix Metrics:
dimensions:  (174, 146)
Summation value of the utility matrix: 642.0
Highest row sum of the utility matrix: 136.0
Highest column sum of the utility matrix: 96.0
```

4 Question 4

Method	Rating Prediction Function	Metric	N		
			N=2	N=3	N=5
Item-Item	Simple average	RMSE	0.8368	0.8068	0.7667
	Weighted average	RMSE	0.8369	0.8066	0.7681
User-User	Simple average	RMSE	0.7006	0.6903	0.6769
	Weighted average	RMSE	1.0257	0.7457	0.6830

- The RMSE decreases as the number of neighbors (N) increases. This suggests that the predictions become more accurate when more neighbors are considered.
- In this dataset the user-user similarities are higher than tag-tag similarities. Hence we get better performance for the user-user method.

5 Question 5

Method	Metric	K=2	K=5	K=10
Without Regularisation	RMSE	0.7190	0.7008	0.6798
With Regularisation	RMSE ($\lambda_1 = 0.001, \lambda_2 = 0.003$)	0.7196	0.6937	0.6784
	RMSE ($\lambda_1 = 0.05, \lambda_2 = 0.05$)	0.7241	0.6885	0.6899
	RMSE ($\lambda_1 = 0.50, \lambda_2 = 0.75$)	0.8515	0.8513	0.8514

- The RMSE value decreases as we increase the no.of Latent factors(K). This is because when we have more latent factors hence information can be gathered.
- The model acheives best performance when $\lambda_1 = 0.001, \lambda_2 = 0.003$ and $K = 10$
- We observe that for high values $\lambda_1 = 0.50, \lambda_2 = 0.75$, the RMSE is greater than without regularization. Therefore choosing optimal hyperparameters ($\lambda 1, \lambda 2$) is important to enchance the model's performance.

6 Question 6

6.1 Collaborative Recommendation

Algorithm	Method	RMSE for N=2	RMSE for N=3	RMSE for N=5
Item-Item	Our method	0.8368	0.8066	0.7667
	Surprise	0.7669	0.7276	0.6942
User-User	our method	0.7007	0.6904	0.6770
	Surprise	0.6711	0.6439	0.6278

- In the Surprise Library, the KNNBaseline method uses an extra baseline rating along with ratings of similar items.
- But in our method we predict the rating only using the ratings of similar items. Hence the Surprise Library method performs better than our method.

6.2 Matrix Factorization Recommendation

Method	RMSE for K=2	RMSE for K=5	RMSE for K=10
Our method	0.7190	0.6885	0.6784
Surprise	0.7275	0.7281	0.7129

The follwing Table gives the best Hyper-parameters obtained for our method and Surprise Library method

Method	Hyperparameters for K=2	Hyperparameters for K=5	Hyperparameters for K=10
Our method	Without Regularisation	$\lambda 1 = 0.05, \lambda 2 = 0.05$	$\lambda 1 = 0.001, \lambda 2 = 0.003$
Surprise	$\lambda 1 = 0.001, \lambda 2 = 0.003$	Without Regularisation	Without Regularisation

- We observe that best performance is obtained without regularisation (or) with very low values of regularisation constants in both Our method and Surprise Library method.