HW 5 (Using the sum of all the ratings to determine which are the three best and which are the three worst)

In [2]:
```
import numpy
file_name_test='testTrack_hierarchy.txt'
file_name_train='trainIdx2_matrix.txt'
output_file='output1.txt'

fTest= open(file_name_test, 'r')
fTrain=open(file_name_train, 'r')
Trainline= fTrain.readline( )
fOut = open(output_file, 'w')

trackID_vec=[0]*6
albumID_vec=[0]*6
artistID_vec=[0]*6
lastUserID=-1

user_rating_inTrain=numpy.zeros(shape=(6,3))

for line in fTest:
    arr_test=line.strip( ).split('|')
    userID= arr_test[0]
    trackID= arr_test[1]
    albumID= arr_test[2]
    artistID=arr_test[3]

    if userID!= lastUserID:
        ii=0
        user_rating_inTrain=numpy.zeros(shape=(6,3))

    trackID_vec[ii]=trackID
    albumID_vec[ii]=albumID
    artistID_vec[ii]=artistID
    ii=ii+1
    lastUserID=userID


    if ii==6 :
        while (Trainline):
            arr_train = Trainline.strip( ).split('|')
            trainUserID=arr_train[0]
            trainItemID=arr_train[1]
            trainRating=arr_train[2]
            Trainline=fTrain.readline( )
```

```
                    if trainUserID< userID:
                        continue
                    if trainUserID== userID:
                        for nn in range(0, 6):
                            if trainItemID==albumID_vec[nn]:
                                user_rating_inTrain[nn, 0]=trainRating
                            if trainItemID==artistID_vec[nn]:
                                user_rating_inTrain[nn, 1]=trainRating
                    if trainUserID> userID:
                        for nn in range(0, 6):
                            outStr=str(userID) + '|' + str(trackID_vec[nn])+ '|' + str(
                            user_rating_inTrain[nn,0]) + '|' + str(user_rating_inTrai
                            n[nn, 1])
                            fOut.write(outStr + '\n')
                        break

        fTest.close( )
        fTrain.close( )
```

In [15]:
```
import pandas as pd

output=pd.read_csv('output1.txt',sep='|',header=None)

output.columns=['User ID','Track ID','Album Rating','Artist Rating']
output['Sum']=output[['Album Rating','Artist Rating']].sum(axis=1)

train_df=pd.read_csv('trainIdx2_matrix.txt', sep='|',header=None)
train_df.columns=['trainUserID','trainItemID','trainRating']

output_cp=output.copy( )
output_cp['Predictor']=0

output_cp['Predictor'][output_cp[0:6]['Sum'].nlargest(3).index]=1

start=0
for i in range(20000):
    output_cp['Predictor'][output_cp[start:6*(i+1)]
['Sum'].nlargest(3).index]=1
    start=6*(i+1)

output_cp_ans= output_cp[['User ID', 'Track ID','Predictor']]
output_cp_ans['Track ID']=output_cp_ans['User ID'].astype(str) +'_'+ output_cp_ans['Track ID']
.astype(str)
output_cp_ans.drop(columns={'User ID'},inplace=True)
output_cp_ans.to_csv('sum.csv', index=False)
```