Music Recommender

# Objective

You’ve taken a job with your favorite music streaming service and they’ve asked you to come up with a way to make recommendations based on a user’s play count. Your boss challenges you to build an SVD recommender system using a subset of the Million Song Dataset.

## Getting the Data

You are provided with two files, the first file, “kaggle\_visible\_evaluation\_triplets.txt”, contains user, song, play counts. The second file “Unique\_tracks.txt” contains details about the song (track, artist, song). Both datasets are in different formats, I’ve provided you with a sample notebook to import the files.

Creating a Dataset of You and a Peer.

Create a dataset of you and a peer’s favorite songs with a play count of 5 or 10. For example here are 5 of my favorite songs.

|  |  |  |
| --- | --- | --- |
| user\_id | song\_id | play\_count |
| MIKE A | SOOHYZC12AC3DF92D3 | 10 |
| MIKE A | SOSGBFG12A6D4FA22F | 10 |
| MIKE A | SOXKWED12AF72A9043 | 10 |
| MIKE A | SOQXNFV12A670215CC | 10 |
| MIKE A | SOZVVMJ12A58A7C548 | 10 |

Append this dataset to your user,song, playcount data.

## Preprocessing

We are going to use the play count column as a surrogate for an explicit rating, play counts range from 1 to 923. My recommendation is to BIN the data so it’s either 1 – 5 or 1 through 10 but that’s up to you. Our assumption is that if someone listened to a song only once then they didn’t love it but if they listened to something say 5 or more times, they probably enjoyed it.

Merge the two datasets together to simplify your analysis!

## Exploratory Data Analysis

Do some basic analysis of the users. What users listen the most, how many play counts, how many unique songs etc. make some charts and tables to support your anlaysis.

Do some basic analysis of the artists and songs. What are the most popular songs and artists, how many unique users have listened to them?

# Train & Evaluate Recommenders

Create a baseline recommendation, see BaselineOnly() method of surprise. Next train a SVD, document the hyperparameters chosen. Did you do any hyper parameter tuning? Compare the BaselineI() performance to that of the SVD. Even if Baseline() RMSE/MAE is lower than the SVD why would having a SVD be preferable to a baseline only model?

* Document performance RMSE/MAE
* Document any hyper parameter tuning you’ve performed.

## Answer the following:

1. For a random sample of 5 users with 10 or more song plays make 5 recommendations of songs they have not listened to with your SVD.
2. Recommendation systems should provide “relevant” recommendations expanding the user’s pool of options (songs in our case) what would you do to improve your recommendation to expand a user’s relevant song recommendation pool?
3. What are your top 10 recommendations for a net new user? That is a user with no user/song play count? essentially the cold start problem.
4. What are your top 10 recommendations for you and your peer?
   1. Do the recommendations make sense? What could you do to improve them? Do you think an Item based KNN might better?