Collaborative-Filter Recommender System

DSGA-1004 Final Project Checkpoint Report

Drishti Singh
New York University
NY USA
ds6730@nyu.edu

Pranjal Srivastava
New York University
NY USA
ps4379@nyu.edu

Vaishnavi Rajput
New York University
NY USA
vr2229@nyu.edu

ABSTRACT

In this project checkpoint, a baseline popularity model was implemented and evaluated using RankingMetrics on the provided ListenBrainz dataset.

1 PRE-PROCESSING

It was decided to use interaction data as a starting point for the project, which included 3 columns - user_id, recording_msid, and timestamp. To ensure that the data was clean, organized and less complex for further analysis, various pre-processing techniques were employed. Firstly, the interaction dataframe was sorted based on the user_id column. Next, the data was repartitioned into 10 partitions and duplicate records were removed. Following this, the dataset was split into a training set and a validation samples in a ratio of 60:40. The training dataset also included 60% of the validation data to ensure that all users were seen during the training process. Both sizes of the provided training dataset (small and complete) were used to evaluate the baseline popularity model explained in the next section.

2 BASELINE POPULARITY MODEL

In the next phase of the analysis, the pre-processed data was utilized to develop a baseline popularity model that recommends popular songs based on user interactions. The model was developed using two distinct approaches:

(i) Approach 1: Calculate the average number of listens for each track per user.

(ii) Approach 2: Calculate the raw count of listens per track on the entire dataset.

The code for the project can be found at the following GitHub link: https://github.com/nyu-big-data/final-project-group-29.

3 EVALUATION METRICS

To test the performance of the baseline popularity model, ranking technique was used on the validation and test data splits to determine the top 100 songs listened by each user which acted as ground truth. These top 100 songs were compared to the prediction of the baseline model and accuracy was evaluated using RankingMetrics. Precisely, Mean Average Precision (MAP) and Normalized Discounted Cumulative Gain (NDCG) were used for the calculation and the following results were obtained:

Dataset	MAP@100	NDCG@100
Small Validation	4.901960784313726e-07	8.02964105594774e-06
Validation	5.3670527366601884e-08	2.5602350374031623e-06
Test	1.5528946403147546e-05	5.832928007163257e-05

Table 1: Results of evaluation on popularity baseline model with average listen (Approach 1)

Dataset	MAP@100	NDCG@100
Small Validation	0.0002762836633979165	0.002881124712965931
Validation	0.0004567715459866738 6	0.004339677005888783 5
Test	9.285713485266364e-05	0.0012220825266727297

Table 2: Results of evaluation on popularity baseline model with total listen (Approach 2)

Table 1 and Table 2 show the results of the evaluation using both the approaches. The accuracy of both the metrics were considerably low for approach 1 in comparison to approach 2.

4 FUTURE WORK

The next phase of the project involves including bias and damping factor in approach 1 for the baseline model and check if the prediction accuracy increases. Additionally, the next phase of the project will involve implementing the ALS model and developing extensions.