Clustering of Last.fm Music Artists

Jordan Rosenblum, Justin Law & Erin Grand

Data Science Institute, Columbia University, New York, NY 10027

May 3, 2015

ABSTRACT

This is the abstract.

This is the body.

Music Recommendation Dataset

Sparsity of $8.08199453451 \times 10^{-5}$ need updating

The matrix was too big for our computers to handle the matrix factorization using the whole set of users and songs, so we subset the data.

The full matrix was originally 163206 songs x 110000 users.

Subset songs by how many times the song was listened to. i.e the popularity of the song = new number

Subset users by how many songs they've listened to (how many popular song?) = new number

SONG PLAY COUNT
mean = 28
max = 35432
std = 215.826789

USER PLAY COUNT
mean = 42
max = 1305
min = 5
std = 53.31547

29 - users 34 - songs

How to split into training and test? Random selection by user or element? ELEMENT. Need matrixes to be the same size to compare them.

Interesting plots: - Cumulative Distribution of number of songs with a given play count