Music Recommendation System

Purpose

- Using user-artist matrix to analyze the similarity of each artist and divide them into 20 clusters. Similarity, do this process on users data.
- Make the recommendation according to the similarity of style of artists.

Dataset

- **Last.fm dataset**
- Total Lines: 17,559,530
- Unique Users: 359,347
- Artists with MBID: 186,642
- Artists without MBID: 107,373

- List of features:
- user ID, artist ID, artist Name, play times

Data Gathering and Preparation

- Rename the columns
- Cleaned dataset: remove the missing values
- Turned the complicated ID format to the simple ones

	userID	playerID	playNum
0	0	37425	2137
1	0	152038	1099
2	0	112364	897
3	0	38434	717
4	0	117441	706

Using SQL to do the data exploration

- Count the total play times of a user
- Count the total play times group by artist with a descent order

	COUNT(*)
0	49
1	51
2	46
3	48
4	49
5	50
6	49
7	55
8	48
9	57
10	53
11	52
12	47
13	44
14	45
15	47
16	43
17	50

	playerID	number
0	104608	77254
1	110770	76271
2	127797	66658
3	87685	48930
4	97938	46954
5	63907	45233
6	82630	44444
7	93784	41229
8	153687	39778
9	127760	37271
10	2185	34174
11	102399	33206
12	3341	33001
13	66397	32626
14	64894	32296
15	83018	32072
16	83461	31918
17	52777	31864

Feature Engineering

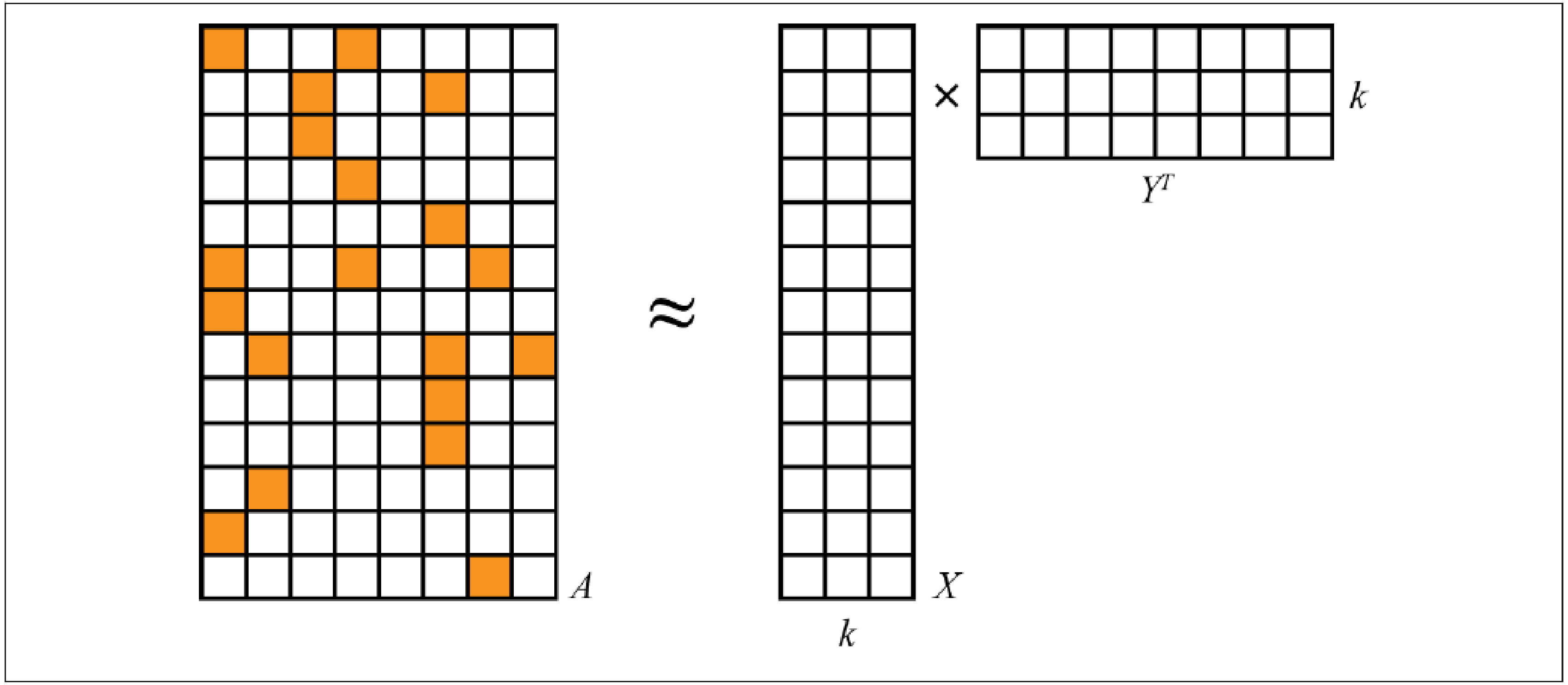
- Turned dataset using sparse matrix (358858,160111)

	Artist1	Artist2	Artist3	Artist4	Artist5	…	Artist160110
User1	play times	play times	play times	play times	play times	…	play times
User2	play times	…	…	…	…	…	…
User3	play times	…	…	…	…	…	…
User4	play times	…	…	…	…	…	…
User5	play times	…	…	…	…	…	…
…	…	…	…	…	…	…	…
User358857	play times	…	…	…	…	…	…

Dimension reduction
LDA

- This is a unsupervised Problem. We used Latent Dirichlet allocation(LDA) to do the dimensions reduction which is clustering. We divided 160k artists into 20 class, according to the taste of users which is the latent variable.

Matrix factorization



$$A = XY^T$$

$$(358858,160111) = (358858,20)*(20,160111)^T$$

```
Topic #0: tom waits, sonic youth, animal collective, pixies, the magnetic fields
Topic #1: nofx, bad religion, misfits, ramones, dropkick murphys
Topic #2: nightwish, sonata arctica, blind guardian, kamelot, apocalyptica
Topic #3: blink-182, fall out boy, my chemical romance, paramore, rise against
Topic #4: the beatles, bob dylan, the rolling stones, johnny cash, u2
Topic #5: miles davis, frank sinatra, johann sebastian bach, norah jones, amy winehouse
Topic #6: opeth, in flames, slayer, katatonia, amon amarth
Topic #7: pink floyd, metallica, iron maiden, ac/dc, queen
Topic #8: tori amos, enya, hans zimmer, enigma, yann tiersen
Topic #9: dir en grey, as i lay dying, bring me the horizon, larc~en~ciel, parkway drive
Topic #10: red hot chili peppers, tool, queens of the stone age, foo fighters, incubus
Topic #11: system of a down, linkin park, rammstein, in flames, koЯn
Topic #12: radiohead, death cab for cutie, arctic monkeys, bloc party, sufjan stevens
Topic #13: kanye west, lil wayne, eminem, 2pac, nas
Topic #14: coldplay, britney spears, madonna, avril lavigne, the killers
Topic #15: boards of canada, aphex twin, daft punk, the prodigy, burial
Topic #16: kent, böhse onkelz, lars winnerbäck, håkan hellström, cmx
Topic #17: radiohead, nine inch nails, muse, placebo, björk
Topic #18: explosions in the sky, mogwai, god is an astronaut, 65daysofstatic, converge
Topic #19: the cure, depeche mode, the smiths, morrissey, joy division
```

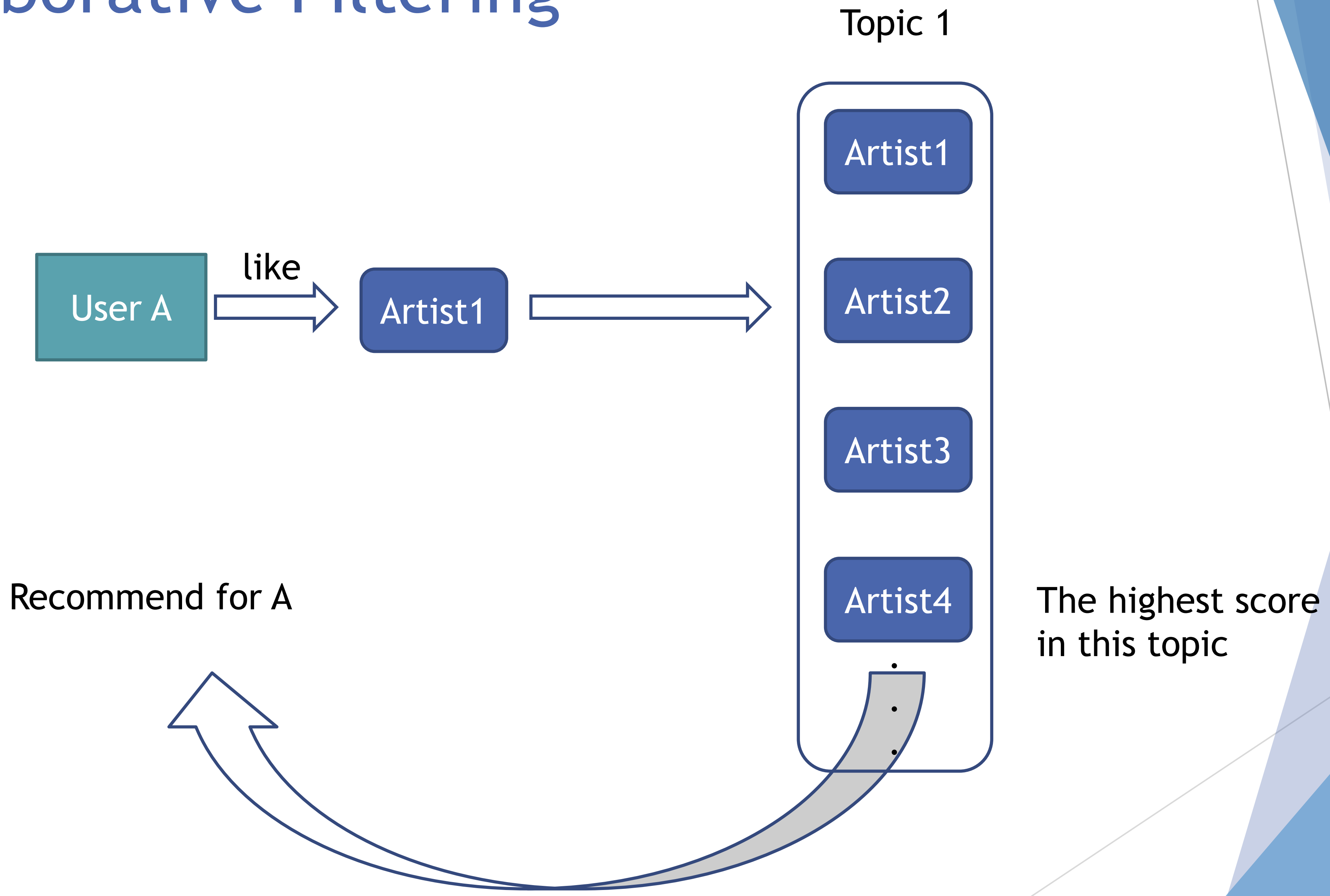
Stability

Because out dataset has 17000K rows, so we take 100K row of the, and run LDA 3 times, and check the topics of them, we can see that LDA is not stable.

```
Topic #0: bad religion, nofx, misfits, rancid, the clash
Topic #1: coldplay, john mayer, moby, keane, hans zimmer
Topic #2: metallica, koЯn, system of a down, die Ärzte, ac/dc
Topic #3: muse, red hot chili peppers, system of a down, pink floyd, nirvana
Topic #4: nine inch nails, queens of the stone age, radiohead, the cure, dep
Topic #5: radiohead, sigur rós, bright eyes, boards of canada, broken social
Topic #6: the beatles, jack johnson, led zeppelin, oasis, the rolling stones
Topic #7: Последние Танки в Париже, sonata arctica, bob marley, within tempt
Topic #8: garbage, the cardigans, lady gaga, metric, thomas dybdahl
Topic #9: lil wayne, daft punk, new order, radiohead, ryan adams
Topic #10: iron maiden, slayer, café tacuba, opeth, megadeth
Topic #11: bob dylan, tom waits, elliott smith, ray lamontagne, johnny cash
Topic #12: kanye west, atb, 50 cent, jay-z, the game
Topic #13: diary of dreams, clock dva, apoptygma berzerk, skinny puppy, kmfd
Topic #14: tori amos, björk, nouvelle vague, kent, thievery corporation
Topic #15: britney spears, rihanna, o.s.t.r., 植松伸夫, beyoncé
Topic #16: madonna, amy winehouse, kylie minogue, michael jackson, superfly
Topic #17: rise against, linkin park, fall out boy, blink-182, paramore
Topic #18: in flames, as i lay dying, linkin park, placebo, all shall perish
Topic #19: 梶浦由記, caetano veloso, mitsumune shinkichi, illya kuryaki and
```

```
Topic #0: stars, the most serene republic, broken social scene, have heart, bright
Topic #1: madonna, britney spears, lady gaga, kylie minogue, michael jackson
Topic #2: new order, the cure, nine inch nails, depeche mode, joy division
Topic #3: sublime, jack johnson, nouvelle vague, 植松伸夫, amy winehouse
Topic #4: depeche mode, muse, daft punk, the prodigy, coldplay
Topic #5: kanye west, john mayer, rihanna, kent, timbaland
Topic #6: the beatles, pink floyd, led zeppelin, oasis, radiohead
Topic #7: garbage, bad religion, the cardigans, thomas dybdahl, metric
Topic #8: 梶浦由記, the pillows, dir en grey, mitsumune shinkichi, くるり
Topic #9: as i lay dying, the devil wears prada, coheed and cambria, all shall peri
Topic #10: boards of canada, radiohead, aphex twin, tom waits, pj harvey
Topic #11: ryan adams, modest mouse, the magnetic fields, belle and sebastian, pave
Topic #12: bob dylan, the beatles, ac/dc, the rolling stones, led zeppelin
Topic #13: anathema, death, o.s.t.r., ulver, napalm death
Topic #14: linkin park, placebo, death cab for cutie, the offspring, muse
Topic #15: Последние Танки в Париже, pain of salvation, fresno, symphony x, johann
Topic #16: rise against, fall out boy, the killers, arctic monkeys, die Ärzte
Topic #17: lil wayne, café tacuba, kanye west, j dilla, atmosphere
Topic #18: metallica, in flames, iron maiden, system of a down, koЯn
Topic #19: radiohead, nine inch nails, sigur rós, the cure, sufjan stevens
```

Collaborative Filtering



Problem Statement and Usefuleness

- Problem: given the artists a user like, how can we recommend him the other artists he would also like.
- We used LDA to divided artists into 20 classes means 20 styles.
- We will recommend him the artist(s) who is(are) the most representative in the same style.

Extra interesting ideas

- Input user features on artist into a neural network and get a output with n dimensions which is latent features represent the style of the user. Do the same thing to artist and get a output represent the style of artist. The inner product of these latent features is the play times. This is the process similar to LDA, but is non-linear. We hope this kind of non-linear model could explore more deep relationship between users and artists

- Derived features Z_m are obtained by applying the *activation function* σ to linear combinations of the inputs:

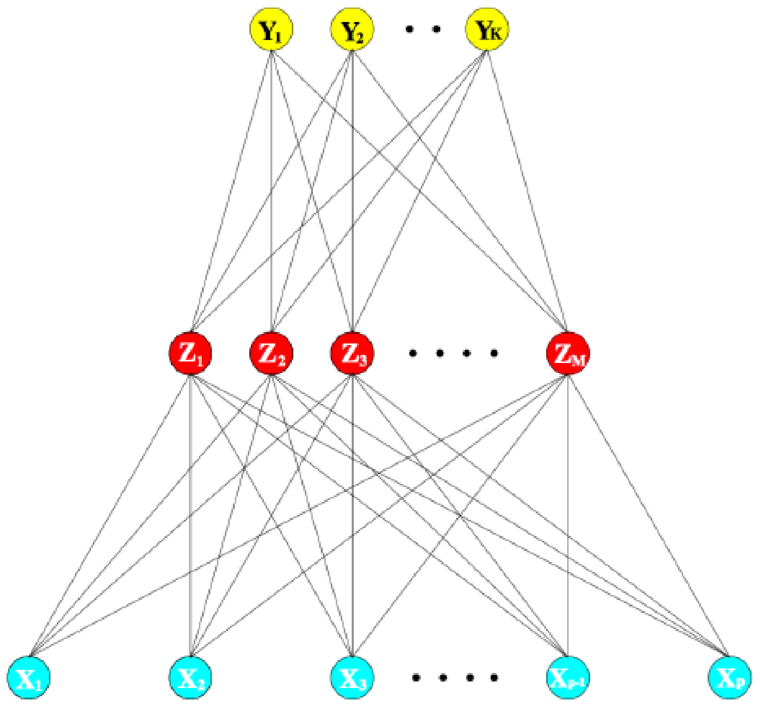
$$Z_m = \sigma(\alpha_{0m} + \alpha_m^T X), m = 1, \ldots, M.$$

- The target Y_k (or T_k in the figure) is modeled as a function of linear combinations of the Z_m:

$$T_k = \beta_{0k} + \beta_k^T Z, \quad k = 1, \ldots, K.$$

- For K-class classification, we use the *softmax* function

$$g_k(T) = \frac{e^{T_k}}{\sum_{l=1}^{K} e^{T_l}}$$



Schematic of a single hidden layer, feed-forward neural network