Dataset:

The dataset you will work with is a sample of a challenge dataset on Kaggle for the WSDM'17 conference. The challenge is "WSDM - KKBox's Music Recommendation Challenge" and its main purpose is to recommend music to users by predicting their chances of listening to a song repetitively. But, our mandatory questions are not about this prediction.

The sample dataset includes 1036 users and 923 songs that they have listened to. The dataset files have one entry per line. The users are listed in a file named "`members.csv`" that includes user id, location (city), age, gender, registration method, registration time, and expiration date columns. Here is the list of columns for this file:

- msno: user id
- city
- bd: age. Note: this column has outlier values, please use your judgement.
- gender
- registered_via: registration method
- registration_init_time: format %Y%m%d
- expiration_date: format %Y%m%d

The songs users had listened to are listed in "train_dataset.csv". This file includes user id, song id, source system tab, source screen name, source type, and target.

- msno: user id
- song_id: song id
- source_system_tab: the name of the tab where the event was triggered. System tabs are used to categorize KKBOX mobile apps functions. For example, tab my library contains functions to manipulate the local storage, and tab search contains functions relating to search.
- source_screen_name: name of the layout a user sees.
- source_type: an entry point a user first plays music on mobile apps. An entry point could be album, online-playlist, song .. etc.
- target: this is the target variable. target=2 means there are recurring listening event(s) triggered within a month after the user's very first observable listening event, and target=1 non-recurrent listening.

1. Build matrix *A* using `train_dataset.csv`. Find the similarities between users based on user listening profiles *A*. Which two users are the most similar to each other? Which two users are the most different from each other?

2. Find the most similar user to user "+FllhLa9X3CCwNeQLg1bgpVxHyfRTaKeVJrCmdjWzY0=".

3. Build matrix *B.* Find the similarities between songs based on song profiles *B*. Which two songs are the most similar to each other?

4. Build matrix *F* using the data in `members.csv`. Using matrix-matrix products (of *F* and *A*), build a matrix *D* for "city listening profiles", in which $D[i, j]$ represents how often song *j* is played in city *i*. Which city has the most number of recurrent songs played? Which two cities are the most similar to each other, in terms of listening to songs?

5. Build matrix *G* using `members.csv`. Using matrix-matrix products, build a matrix *D* for "gender profiles", in which $D[i, j]$ represents how often (recurrently) song *j* is played by males or females. Which song is the most played by females? Which is the most played by males?

6. Similar to lab of Chapter 8, Assume that we want to predict if a user is male or female based on the songs they listen to. Suppose that a classifier function in the form of $C(y)$

(defined below) can classify users into these two classes. Here, y is one user's listening profile and w is a D-vector that is consisted of classifier coordinates (coefficients) that separate the two classes.

$$C(y) = \begin{cases} 1\ (female) & if\ h(y) \geq 0 \\ -1\ (male) & if\ h(y) < 0 \end{cases}$$
$$h(y) = w.y$$

a. Write a procedure to create a vector *b*, whose domain is the set of user ids, and $b[i] = 1$ if the user is "female" and $b[i] = -1$ if the user is male. ($b[i] = 0$ if the gender is not indicated.)
b. Write another procedure fraction_wrong(A, b, w) with the following spec:
    i. input: A matrix *A* whose rows are user listening profiles, a vector b whose entries are +1 and −1, and a vector *w* that has coordinates
    ii. output: The fraction of of row labels *i* of *A* such that the sign of $(row\ i\ of\ A) \cdot w$ (or $A[i,:] \cdot w$) differs from sign of $b[i]$.
c. Use file `sample_coordinates.csv` (it contains song coefficients and song ids) to create a vector w, whose domain is the set of song ids, and $w[i]$ = the coefficient that is listed in front of song *i* in `sample_coordinates.csv`.
d. Calculate and report fraction_wrong(A, b, w) for this dataset assuming matrix *A* as user listening profiles, and vectors *b* and *w*, as created above. What does this number represent?