Data Wrangling Steps

This document describes the data wrangling steps that were performed in each of the following Capstone Project files:

- train.csv
- members.csv
- songs.csv
- song_extra_info.csv

train.csv

- 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=1` means there are recurring listening event(s) triggered within a month after the user’s very first observable listening event, `target=0` otherwise .

Here is the list of steps that were followed:

- Identify ‘Unknown’ values as NaN
- Transform source_screen_name values to lower letters
- Transform source_system_tab, source_screen_name and source_type columns into categorical variables
- Drop rows containing missing source_system_tab and source_type values, since they represent less than 5% of the data.

Details can be found in data_wrangling_train.ipynb

members.csv

User information.

- msno
- 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`

Here is the list of steps that were followed:

- Transform registration_init_time and expiration_date columns into datetime variables
- Transform city, gender and registered_via columns into categorical variables
- Remove rows where expiration_date is before 2004, since KKbox was launched in 2004.
- Age values less than 0 and greater than 120 are converted into missing values (NaN)

Details can be found in data_wrangling_members.ipynb

songs.csv

The songs. Note that data is in unicode.

- song_id
- song_length: in ms
- genre_ids: genre category. Some songs have multiple genres and they are separated by `|`
- artist_name
- composer
- lyricist
- language

Here is the list of steps that were followed:

- Replace one missing language value for -1
- Format language values removing decimals and convert language into categorical variable.
- Express song_length in minutes instead of milliseconds
- Capitalize first letter of each word and remove unnecessary spaces (leading, ending and multiple spaces) in the artist_name, composer and lyricist columns.

Notes:

- Language values equal -1 sometimes refer to songs with only melody and other times to songs were the language has not been categorized. Therefore, no further steps have been taking to transform those values.
- Song_length outliers were kept since they correspond to playlists

Details can be found in data_wrangling_songs.ipynb

song_extra_info.csv

- song_id

- song name - the name of the song.
- isrc - International Standard Recording Code, theoretically can be used as an identity of a song. However, what worth to note is, ISRCs generated from providers have not been officially verified; therefore the information in ISRC, such as country code and reference year, can be misleading/incorrect. Multiple songs could share one ISRC since a single recording could be re-published several times.

Here is the list of steps that were followed:

- Capitalize first letter of each word and remove unnecessary spaces in the name column.

Details can be found in data_wrangling_song_extra_info.ipynb