Computing Lab and Data Warehousing & BI Project – Milestone III

Niti Mishra, Miquel Torrens, Bálint Ván

We created the database `omsong` for our project. It consists of four interrelated tables that act as data warehouse for the analytics part. The schema of `omsong` database is attached at the end of this document. Provided below are the descriptions of the tables:

1. `song_metadata`: This table contains majority of the information. It includes the details at a song level: aggregates on the technical features such as tempo, loudness, etc., as well as metadata, such as artist, album, year or location. The table has 10,000 observations and its primary key is the `song_id`. At the moment, the table is indexed by `song_id` and also by `artist_id`. If the analytics require extra indexes, we will add them as we go.
2. `artist_relations`: This table contains the relations between artists, that is, which are the most usually related artists to each artist and ranks them according to how frequently they are related. The `artist_relation_id` is the primary key of the table and it is indexed by this field and also by the `song_id` and the `artist_id` fields, which are foreign keys to the table `song_metadata`.
3. `artist_terms`: this table contains the terms (e.g genre) associated to each song referring to the artist. It also features the frequency and the weight of its terms. The `term_artist_id` is the primary key of the table. The table is indexed on the primary key, and the fields `song_id` and `artist_id`, which are foreign keys to the table `song_metadata`.
4. `song_tags`: This table contains the tags associated to each song gathered from Musicbrainz database. It also features the frequency of the terms. The `song_tag_id` is the primary key of the table. The table is indexed on this primary key, the `song_id` and the `artist_id` fields, which are foreign keys to the table `song_metadata`.
5. `words`: it contains the top 5,000 words from the musiXmatch data set. The primary key is the `word_id` variable. The variable `word` contains the word itself.
6. `word_count`: this table has a composite primary key: `words` and `word_id`. These variables reference the tables `song_metadata` and `words` respectively. The variable `word_count` shows how many times the respective word occurred in the respective song.

Note, that the lyrics are represented in a bag-of-words format. Not every word is represented in the data set, only the top 5,000, but there is data about all of these words for each song present in the musiXmatch data set.

These tables will eventually be complemented with some summary tables or functions/procedures on the results of the analysis. This second part will be designed according to the analytics outcomes. This second part of the database will be built once these outcomes are established.

Schema for the `omsong` database