1 Data

1.1 Dataset

The dataset [1] used to train and test the machine learning model was obtained from another paper [2] that was working on a similar project. The authors of that paper made their dataset available online. There are 2252 songs in this dataset, 1802 of which had been categorised as the training set while the rest had been categorised as the test set. Each song is in major key and only have a single chord per bar. For each song, all the relevant features had been extracted and placed into a single *CSV* file as shown in Figure.

As can be seen in Figure, the rows each contain information about a single note. Each bar is taken to be a single measure. The columns each represent a different piece of information about that particular note. *time* refers to the time signature, *measure* refers to the measure to which that particular note belongs to, *key_fifths* indicates the number of sharps/flats (e.g. -1 for one flat and 1 for one sharp), *chord_root* is the root of the chord with *chord_type* indicating the type of chord, *note_root* identifies the particular single note of that row, *note_octave* is the octave of that note, and *note_duration* indicated the duration of the note (4.0 for a quarter note).

1.2 Preprocessing of dataset

The dataset has to be preprocessed in order to make things simpler later on.

1. All songs are transposed to C major key. The key of a song determines the notes and the set of chords present in the song. Transposing all songs to a common key will basically normalise the different features of melodies and chords in different songs. The number of chord types present in the dataset will be reduced, which will decrease the number of chord types during the training process. Each song can be shifted to a different key without loss of the song's subjective character [3].

2. The time signatures are all normalised. Different songs have different time signatures. To do so, each *note_duration* is multiplied by the reciprocal of the time signature *time* to give a normalised note duration.

3. Chord types are restricted to C major and C minor chords. All other chord types are converted to their most similar scale tone chords in key C.

4. Some measures in the dataset contain rest notes. These measures are removed from the dataset.

5. Octave information is not required and is removed from the dataset.

References

[1] H. Lim, S. Ryu, and K. Lee, "CSV Leadsheet Database," CSV_Leadsheet_DB. [Online]. Available: http://marg.snu.ac.kr/chord_generation/. [Accessed: 05-Jan-2022].

[2] Hyungui Lim, Seungyeon Ryu and Kyogu Lee. "Chord Generation from Symbolic Melody Using BLSTM Networks", 18th International Society for Music Information Retrieval Conference, 2017

[3] Chen, Ziheng and Jie Qi. "Machine Learning in Automatic Music Chords Generation." (2015).