Milestone 1 Report

Current State of the Project

Milestone 1 goal: Collect at least ~100 samples from the Million Song Dataset for each planned genre.

- Dataset collection done
  - 8 Genres with at least 99 samples each
  - The genres are Rock, Jazz, Folk, Blues, Latin, Electronic, Rap, and Country
- Could not download the entire Million Song Dataset (MSD), due to its giant size of over 200 GB. Instead, we worked with the Million Song Subset, a subset of 10_000 songs from the MSD, which occupies a more manageable ~2 GB.
- Two approaches tried (the second one being more successful):
  - Compiled a list of songs for the desired genre by scraping the web, and search for those songs in the Million Song Subset. For each genre, this yielded on the order of 1 to 10 samples within the Million Song Subset, and 100s when searching the entire MSD.
  - Used the MAGD genre annotations provided as a companion to the MSD to filter for songs of the desired genres from the Million Song Subset. For each genre, this yielded on the order of 100s of songs within the Million Song Subset, and 1000s when searching the entire MSD.

Adjustments to Proposal

- Could not download the entire Million Song Dataset (MSD), due to its size in the 100s of gigabytes. Instead, we used the Million Song Subset, a subset of 10_000 songs from the MSD.
- Matthew's two genres (Classical / Metal) changed to Electronic and Latin due to lack of samples for those genres in the Million Song Subset.
  - Classical only had 6 samples for example
  - Replaced with Latin and Electronic which had 257 and 228 samples respectively

Current Challenges / Bottlenecks

- Some genres, like rock (pop rock), are over-represented in the Million Song Subset. Others, like classical, are almost non-existent.
  - We are working around under-represented genres by replacing them with better represented genres.
- If we had access to the entire MSD, lack of data for specific genres would be less of an issue.
  - Will try using Google Cloud credits to download and use the entire MSD from Milestone 2 and onwards.

Team Member Contributions

- Carroll, Quinn
  - Assembled lists of track IDs in .h5 format for the rap and country song genres in the Million Song Subset (10000 songs from Million Song Dataset). These samples were identified by genre by the MSD All-music Genre Dataset (MAGD). These are submitted on GitHub as Rap Subset (236 samples) and Country Subset (175 samples). I used similar python scripts and methodology as created by Matthew Poon in the AssembleDataset_LatinElectronic folder.
- Jung, Cassiel

    - All the works are in the folder AssembleDataset_FolkBlues. Used data from the site "http://www.ifs.tuwien.ac.at/mir/msd/download.html#groundtruth" which classifies track id of music based on its genres. Basically, my work was to separate out track ids of folk and blues music. Wrote the code to read through trackIDs.txt line by line until there is no more line to read then copy track id to appropriate text file if the line contains either "Folk" or "Blues". Successfully filter around 3500 track ids for each genre.
- Poon, Matthew
    - Wrote two python scripts to
        - Filter MAGD genre annotations to only contain track ids for a given genre
        - Take the filtered MAGD genre annotations and apply it to the Million Song Subset to filter out all tracks matching the desired genre(s).
    - Created genre subsets from the Million Song Subset for the following genres:
        - Classical (Not enough samples, replaced with Electronic)
        - Metal (Not enough samples, replaced with Reggae and later, Latin)
        - Electronic (228 samples)
        - Latin (257 samples)
        - Reggae (120 samples, dropped in favor of Latin for more samples)
    - The python scripts and a text file detailing steps taken / instructions to use the scripts can be found in the AssembleDataset_LatinElectronic directory in the project repository
- Sai Subramanian, Aswin
    - Built lists of track IDs for rock and jazz songs in the Million Song Subset, a sample of 10_000 songs from the Million Song Dataset (MSD), and collected the corresponding data files into the folders "Rock Subset" and "Jazz Subset."
        - Work can be found in AssembleDataset_RockJazz.ipynb.
            1. Downloaded the Million Song Subset, a sample of 10_000 songs from the Million Song Dataset (MSD).
            2. Downloaded and learned to work with the unique_tracks.txt reverse-index file provided for the MSD (and Million Song Subset)
            3. Wrote code to discern which of the 1_000_000 entries in unique_tracks.txt are included in the Million Song Subset.
            4. Filtered out repeated titles among the songs that were included in the Million Song Subset.
            5. Learned to access info from hdf5 data files of the MSD using hdf5_getters python module (source and citation for hdf5_getters.py in AssembleDataset_RockJazz.ipynb).
            6. Compiled a list of rock songs and a list of jazz songs by scraping 3 webpages using pandas.
            7. Found 700-800 rock track IDs and 700-800 jazz track IDs. However, most of these were from outside the Million Song Subset.
            8. Used the genre annotation file found at http://www.ifs.tuwien.ac.at/mir/msd/partitions/msd-MAGD-genreAssignment.cls to find about 1000 rock ("Pop_Rock") track IDs, and about 100 Jazz track IDs inside the Million Song Subset.

9. Wrote the trackids found to rock_trackids.txt and jazz_trackids.txt, which can be found in the folders “Rock Subset,” and “Jazz Subset,” respectively.
10. Collected the files corresponding to the rock and jazz track IDs found, and put them into the folders “Rock Subset,” and “Jazz Subset,” respectively.