Jou Lee
Fall IS578
Intro to Digital Humanities

Data Biography/DH Project Review

The COVID-19 pandemic has changed the behavior of music listeners around the globe. Researches from different parts of the world examined the swift from various viewpoints and methods. Inspired by previous works, this project hopes to create a comprehensive dataset to analyze the impact of the pandemic on U.S. music listeners' behavior. The selected works below were considered relevant to the project's purpose and implementation.

The LFM-1b Dataset

The purpose of the dataset was to forecast the music behavior of listeners. Therefore, the dataset could apply to music information retrieval and recommendation system. The data were collected on a single platform, Last.FM. This dataset selected over 120,000 users' data with 1 billion songs from its 250 genre tags. The period of the users' data was fetched from January 2013 to August 2014. To predict users' listening behavior, the dataset designed metadata for songs and users but focused more on the users' metadata.

The Million Song Dataset

The Million Song Dataset (MSD) is a freely-available dataset to train machine learning algorithms for music recommendation systems. Developed by The Echo Nest and LabROSA of Columbia University, the MSD has become the most widely used dataset

in music information retrieval (MIR) research. The dataset collected the metadata[1] of a million popular music from 1922 to 2011. Many of its metadata collaborated with partners, e.g., “tags” with MusicBrainz, as well as its additional datasets. The MSD partnered with Second Hand Song to release the “SecondHandSong dataset[2]” for cover songs and musiXmatch to collect lyrics in “musiXmatch dataset[3]” which both linked to the MSD metadata. For its comprehensive data sources and metadata, MSD has become a popular dataset for MIR research in the academia and industry.

The Dataset for Spotify Streaming and the COVID-19 Pandemic.

This dataset is most relevant to my project goal and data collecting strategy. This dataset coped with “emotion-focused strategies" and compared the music behavior of listeners during 2019 to the first lockdown in 2020. It collected Spotify’s daily top 200 charts and their audio features from the DACH countries. DACH is an acronym for Germany (D), Austria (A), and Switzerland (CH), where German is spoken by the majority of people. The metadata consisted of the general information of the songs, country name, and the audio features generated from Spotify’s “Organize Your Music”.

The datasets above are widely used in music analysis and recommendation fields. Although none of them are directly the same as my final project, their dataset purposes and data collecting strategy are similar and inspiring to some parts of it. The LFM-1b Dataset creates metadata for song and user from Last.FM platform to analyze behavior and make the prediction. This single platform contributed to research on the

[1] An Example Track Description. http://millionsongdataset.com/pages/example-track-description/
[2] SecondHandSongs dataset, the official list of cover songs within the Million Song Dataset, available at: http://labrosa.ee.columbia.edu/millionsong/secondhand.
[3] musiXmatch dataset, the official lyrics collection for the Million Song Dataset, available at: http://labrosa.ee.columbia.edu/millionsong/musixmatch.

music behavior of the pandemic. The Million Song Dataset is known for its collaboration with many platforms and partners. It inspired my project to include data from other platforms, e.g., social media, and perform text analysis to observe popular word choices and trends. The Dataset for Spotify Streaming and the COVID-19 Pandemic is the most relevant one to my project. It includes metadata of audio features which is generated by Spotify Organize Your Music. However, my final project would extend the time range and work on COVID-19 behaviors in the U.S. Also, the songs would be collected from the Billboard chart instead of the playlists on the Spotify platform.

REFERENCE

Bertin-Mahieux, T., Ellis, D. P.W., Whitman, B., & Lamere P. (2011).The Million Song Dataset. *Proceedings of the 12th International Society for Music Information Retrieval Conference.*

Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infectious Diseases, 20*(5), 533-534. doi:10.1016/S1473-3099(20)30120-1.

Kalustian, K., & Ruth, N. (2021). “Evacuate the Dancefloor”: Exploring and Classifying Spotify Music Listening Before and During the COVID-19 Pandemic in DACH Countries. *Jahrbuch Musikpsychologie*, 30, 1-26. http://dx.doi.org/10.23668/psycharchives.5020.

Liu, M., Zangerle, E., Hu, X., Melchiorre, A., & Schedl, M. (2020). Pandemics, music, and collective sentiment: Evidence from the outbreak of covid-19. *Proceedings of the 21st International Society for Music Information Retrieval Conference. https://archives.ismir.net/ismir2020/paper/000362.pdf.*

Million Song Dataset. (2011, February 8). Million Song Dataset. Retrieved October 31, 2021, from http://millionsongdataset.com/.

Schedl, M. (2016). The LFM-1b dataset for music retrieval and recommendation. *In Proceedings of the 2016 ACM on international conference on multimedia retrieval*.

103-110. *ACM* Digital Library. https://dl.acm.org/doi/abs/10.1145/2911996.2912004.