Tony Nguyen

Dr. Gina Sprint

CPSC 222 01

13 October 2022

Project Part 3

For this assignment's project, I choose to use Spotify's takeout data. This has been the one that I am most hesitant to work with, probably due to my prejudice to JSON at first – a completely new file format for me. However, as I have progressed through the semester, I feel more comfortable working with it now. One nice thing about Spotify's takeout is that it provides lots of different data tables compared to my Netflix and YouTube takeout (probably because I do not have a channel, so I did not have its insights like Gina's.)

Among the table Spotify provided, I chose to work with StreamHistory0.json because of the number of instances recorded. This table contains information on the streaming history of my account, just like the name of it, for one year, from September 21, 2021, to September 21, 2022. I plan to combine this table with Playlist1.json, which contains a list of songs that have appeared in "it's me" – my offline playlist – since I started using Spotify in 2017. Through this combination, I want to study my listening genre or most frequently played artist since I added every song I like to "it's me" to listen to over time.

In StreamingHistory0.json, an instance is a song, and its universe of instance is a list of songs streamed within one year since the day I downloaded the takeout; for attributes, it is be "endTime" – timestamp of when the song ended streaming (numeric interval,) "artistName" – name of the artist (categorical nominal,) "trackName" – the name of the song (categorical nominal,) and "msPlayed" – the duration played of each track (numeric ratio-scaled.) This table

has a key of “endTime” since there can be only one song played at a time. If there is a class that can be used for supervised learning in this table, it is probably “artistName” as we can know how many times I have played a song by any artist. My prediction for this will be Taylor Swift (Midnights album is coming!) and Ha Anh Tuan (one of my favorite Vietnamese singers.)

And for the Playlist1.json, its instance is also songs I have played, and the universe of instance is a list of songs I have added to “it’s me.” Playlist1.json’s attributes has “track” – information about the song, “episode” – if it is a podcast-like type (numeric ratio-scaled,) “localTrack” – if a song is imported from my machine (numeric ratio-scaled.) Under “track,” there is a sub-list of attributes that includes “trackName” and “artistName,” which is the same as StreamingHistory0.json, “albumName” – name of the album that the track is in (categorical nominal,) and “trackUri” – a unique key from Spotify database (categorical nominal.) Since “trackUri” is unique for every song, it is also the table's key. Besides, If there is an attribute that can be used for further investigation, it is “addedDate.” I find this one interesting since I can explore genre per time period, as Playlist1.json has a more extended instance range than StreamingHistory0.json. I will have a general idea of how my genre favorite has “evolved” over the last five years. Finally, when working with both tables, I will use “trackName” as the shared key to refer to each instance after joining them together for further analysis.