Capstone 1: song classification of spotify genres – Data Wrangling

Summary

The data wrangling for my project consisted of two major steps:

1. *Scrape https://spotifycharts.com/regional to pull the top 200 most popular songs from the past two years*
2. *Using the scrapped list generated from step 1), leverage the Spotify api to pull further details from the top 200 songs list*

Detailed steps

Step 1) - Scraping Spotify Charts

Data Collection:
Data for the top 200 songs was collected by web scraping Spotify Charts. Using the Python library *requests,* a short script was utilized to download the daily csv files from the past two years and write to a local folder. Error handling was added to capture any dates where the csv did not download correctly. The date of these files were written to a specific folder for review.

Data Validation Issues:
Csv files for 4 dates were unable to be downloaded due to an error on Spotify's website. No data was provided for these 4 dates. Due to the top ranked songs not changing significantly day to day, the missing data was omitted as it is immaterial for the overall classification model to function properly.

Finalizing Data:
Once downloaded locally, another short script was utilized to amalgamate and clean the csv files by using a helper function to read the csv data to a list and using the *Pandas ".Append"* method to amalgamate to one dataframe.

Once summarized in one dataframe, a mask was used to remove all the headers from the csv files as well as any inconsistencies.

The final cleaned dataframe was written to a csv for analysis.

Step 2) – Extracting song genre and features from Spotify api

Data Collection:
Using the cleaned csv from step 1, a unique list of artists was extracted by loading the csv into a dataframe and creating a series of approximately ~700 unique artists.

The module *Spotipy* (an open source library) was then used to make API calls to pull the ~700 artist's unique Spotify artist id.

Using the 700 unique Spotify ids, another API call was made to pull all related artists to the ~700 artist's from the step 1) csv file. The result was a new artist list totaling ~4.6k artists. (Important to note – the resulting ~4.6k was only the name of the artist. A subsequent api pull was required to pull the formal spotify ids for these artists.)

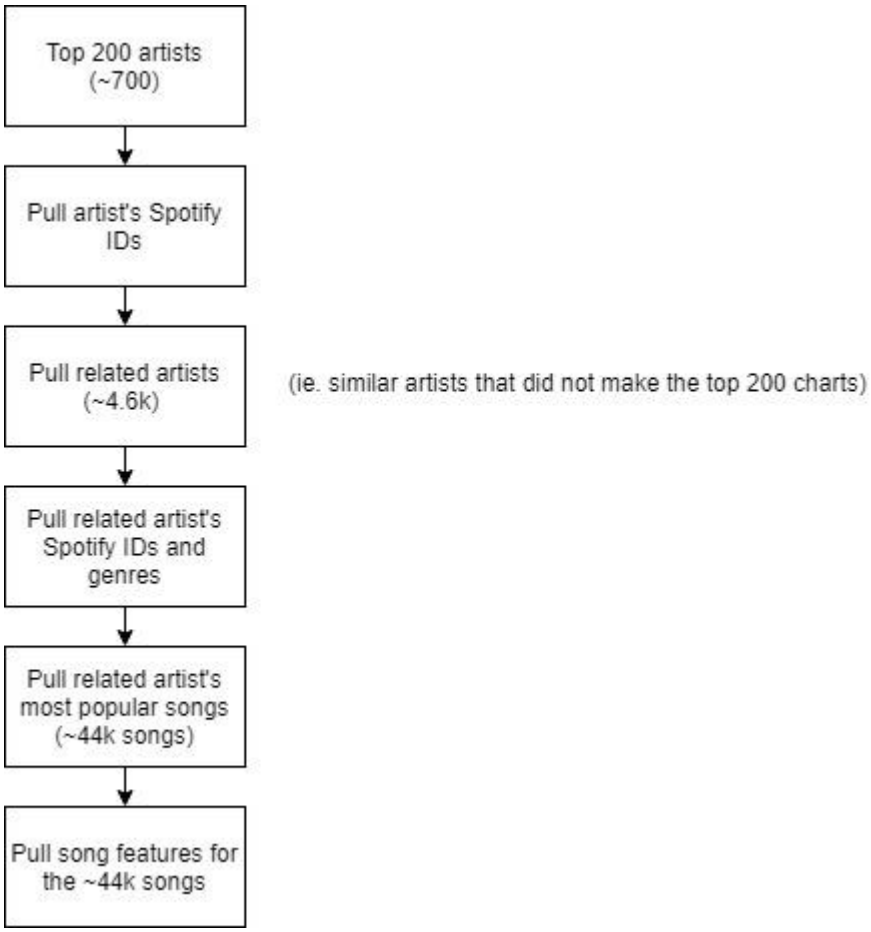
Using the ~4.6k artist list, another API call was made to pull ~4.6k unique Spotify artist ids as well as the artist's genre.

With the ~4.6k unique Spotify artist ids, an API call was made to pull the ~4.6k artists most popular songs. The result from this pull was a giant song list of approximately 44k songs

The final step was one final API call to pull the song features for each of the 44k songs.

Two datasets were created from the above process:

- A ~4.6k list of artists as well as their corresponding genres
- A ~44k list of the songs as well as their features (danceability, energy, key etc.)



Data Validation Issues:

There were instances in the above steps where artist or song name caused an API call to not return a valid response. These instances were due to characters not recognized by normal utf-8 encoding (for example, songs with Chinese characters). Due to the small amount of these instances, all invalid songs or artists were removed from the final dataset. With a ~44k song sample size, these invalid songs or artists are immaterial to the overall project.

Finalizing Data:

The final step step in the data wrangling is was cleaning the final datasets. This included removing any artists without a genre on Spotifiy's system. Two pickle files housing dataframes were used to house the final artist/genre list as well as the song list housing the song features.