DMML 2019 PROJECT
Group TESLA

Yassin Hediger, Marija Krsteska, Samuel Lew, Martynas Savickas

In this project we will mainly analyze song lyrics using text analysis. This document will give a brief overview of the dataset we are going to use, the ideas that we plan to explore, the machine learning techniques we plan to apply, and the usefulness of our project.

Dataset

We are going to create our own dataset using Spotipy. Spotipy is a Python library for the Spotify Web API, that enables full access to all of the music data provided by the Spotify platform.[1] The Spotify Web API returns all response data as a JSON object. For the lyrics, we will integrate Spotify and Genius API to crawl song lyrics.[2]

We will start with getting all Spotify categories, then category's playlists and tracks for each playlist. We will choose a few categories, and playlists from each of the chosen categories. Spotify categories are the ones that are shown under 'Browse' tab, 'Genres & Moods'.[3] We will use the songs from these playlists to create the datasets. Our main dataset will contain 'uri', 'id', 'title', 'lyrics', 'genre', 'artist', 'year of release', 'popularity' for each track. Additionally, we will have a dataset containing the audio features of each track, 'spotify uri', 'spotify id', 'duration_ms', 'key', 'mode', 'time_signature', 'acousticness', 'danceability', 'energy', 'instrumentalness', 'liveness', 'loudness', 'speechiness', 'valence', 'tempo'.[4]

We plan to include data from Google Trends using pytrends, an unofficial API for Google Trends.[5] We plan to get the worldwide popularity of each song in the past 12 months. So, this data will consist of a song title, together with the month when the song had reached maximum popularity.

Ideas

First, we want to predict the category (genre, mood) of a newly released song based on its lyrics. This is a classification problem.

Secondly, we want to predict the popularity of a newly released song based on its lyrics and the additional features. This is a regression problem. Additionally, using data from Google Trends, we will try to map each song's popularity to a particular month of the year in which the song is most likely to be popular. This is a classification problem.

Another idea is to classify the songs based on artist's gender, analyze the main differences between genders and predict whether a song is from a male or female artist.

At the end, we will try to implement a recommender system using the lyrics of the song together with the additional features. The system can recommend songs to a user, based on a song he/she already likes.

Impact

We assume that our project can give some useful insights to users. The users can predict the popularity of a song. Also, they can predict what is the best time of the year to release the song in order to gain maximum popularity. The recommender system can help users find similar songs based on their taste.

[1] Spotipy library - https://spotipy.readthedocs.io/en/latest/

[2] Integrate Spotify and Genius API - https://dev.to/willamesoares/how-to-integrate-spotify-and-genius-api-to-easily-crawl-song-lyrics-with-python-4o62

[3] Get category's playlists - https://developer.spotify.com/documentation/web-api/reference/browse/get-categorys-playlists/

[4] Get audio features - https://developer.spotify.com/documentation/web-api/reference/tracks/get-audio-features/

[5] Pytrends - https://github.com/GeneralMills/pytrends