Tweets, Language and Music Predictor

Soham Desai - Springboard Data Science - Capstone 2 - February 2020

Why?

Music has always been an interest of mine. Personally it helps me be me. I can listen to it to relax, to focus, to workout and so much more. Besides what it can do at a personal level, music has the ability to connect people in ways they may or may not now. Thanks to the ever growing use of social media and technology, these connections are formed even more frequently. For my second capstone, I'm interested in seeing if twitter users with similar music interests can be identified by their tweets. To do so the first step needed to be taken was acquiring the data.

Approach

1. Data Acquisition
    a. FollowerWonk
    b. TWINT
2. Data Wrangling
    a. Advanced NLP - (Sentence Similarity)
    b. LangDetect
3. Exploratory Data Analysis
    a. Predictive Features - Words, Emojis
    b. Word Counts
4. Machine Learning
    a. Various Models with Various Vectorizers

Client?

- Market Researchers
  - People who are trying to focus on music related segments can use this to find specific individuals related to a musical genre
  - This can be edited and tweaked to fit any category
    - Movies
    - Sports
    - Cars

Data Acquisition

- Using FollowerWonk I found 200 individuals per category of music
    - Hip Hop, Country, EDM, Jazz, Metal Rock
- Using TWINT I then scraped 2000 tweets per individual per genre accumulating a dataset of 40,000 tweets

Data Wrangling

- Some tweets were not in English.
  - Using LangDetect I dropped those tweets out of my dataset
- I did not want any tweets related to music
  - Using Advanced NLP (sentence similarity) I determined that if a tweet was >50% in relation to the words “music” “song” “artist” then I would drop that tweet from data set

- I was left with 6600 tweets after cleaning my dataset.