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 Twitter Intelligence Tool (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.

Exploratory Data Analysis

- Predictive Features - Words
  - I examined the best and worst predictive words per genre.

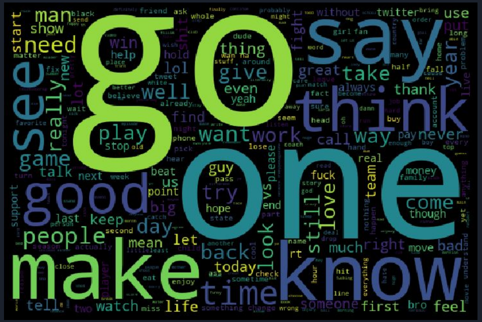
Genre	Top Words	Worst Words
Country	f**k, stream, business	birthday, tour, amazing
EDM	trump, important, plus	india, rn, lbs
Hip Hop	meet, amazing, flight	bro, homie, italy
Jazz	favorite, yeah, tho	student, trump, constitution
Metal Rock	half, student, teach	metal, hospital, sort

Exploratory Data Analysis

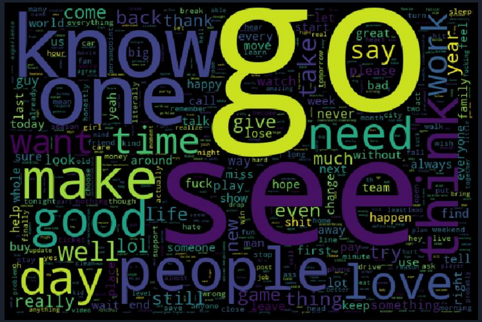
- Predictive Features - Emojis
  - I examined the best and worst predictive emojis per genre.

Genre	Top Emojis	Worst Emojis
Country	pray, you, male	blush, hearts, tongue out winking
EDM	cry, tongue out winking, blush	100, purple heart, pleading face
Hip Hop	heart, blush, two hearts	male sign, facepalm, fire
Jazz	sweat smile, rolling eyes, two hearts	thinking, pray, rofl
Metal Rock	fire, clap, 100	sweat smile, two hearts, wink

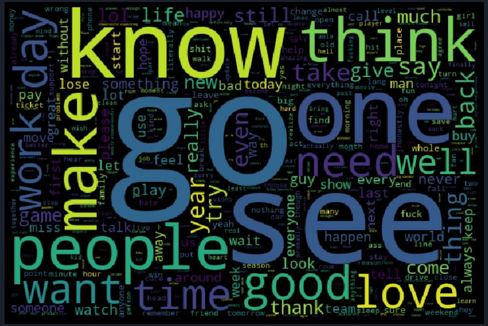
Exploratory Data Analysis - Word Clouds



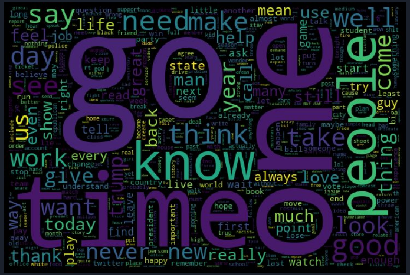
Country



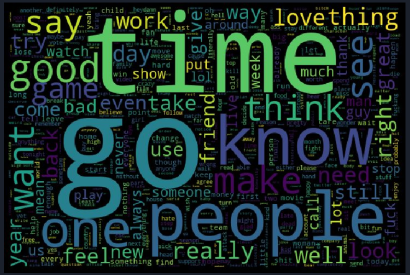
EDM



Hip Hop



Jazz



Metal Rock

Machine Learning

Utilizing both TFIDF Vectorizer and Count Vectorizer I performed multiclass classification using Logistic Regression, Naive Bayes, LinearSVC, and Random Forests. Although they all did not perform very well (see tables on next slide), Linear SVC using a CountVectorizer performed the best.

Classifier	Accuracy
Naive Bayes	0.37
Logistic Regression	0.34
Random Forest	0.31
LinearSVC	0.34

CountVec

Classifier	Accuracy
Naive Bayes	0.32
Logistic Regression	0.34
Random Forest	0.32
LinearSVC	0.34

TFIDF Vec

Limitations

1. Could not control the quality of user tweets.
    a. Sponsors
    b. Retweets
    c. Non-English
    d. Are they really devoted to that music genre?
2. Length of Tweets
    a. Capped at 140 during the time I scraped this.
    b. People use slang and emojis.
    c. Not a lot of words.

Next Steps

1. Work on acquiring a more dense and accurate data set
    a. Backed by people who actively tweet and are devoted fans to a specific genre
2. Research and apply new and interesting advance NLP techniques to find some patterns amongst cleaned data set.
3. Create new engineered features to improve ML scores for accuracy as well as F1 to create a better predictor overall.
4. Consult individuals who are more well versed in language and music to help bridge the connection.

Thanks a Lot! Check out my Github.

https://www.github.com/sdesaidata