Hip Hop or Indie? Predicting Music Subreddits

David Cortes

What's the Problem?

- Articles on The Fader website are categorized by music genre
- Writers have to research and dig through a lot of information, posts, other articles, etc. in order to write their articles
    - This research process is very cumbersome and time-consuming

Bad Bunny was the most streamed artist in the world on Spotify in 2020

Spotify users around the world streamed music by Bad Bunny the most in 2020.

MUSIC / REGGAETON

Phoebe Bridges shares "Savior Complex" video directed by *Fleabag*'s Phoebe Waller-Bridge

Watch the music video for "Savior Complex" by Phoebe Bridgers from director Phoebe Waller-Bridge and starring *Normal People*'s Paul Mescal.

MUSIC / ROCK

Travis Scott launches his own fragrance, reportedly eyeing move into hard seltzer

Travis Scott reveals "Space Rage" range of cologne and candles as his suite of corporate partnerships, which also involves deals with Sony and McDonalds, continues to increase.

MUSIC / HIP-HOP

How Do We Solve This Problem?

- We can predict where a certain piece of information came from based on music genre
    - This will categorize each post and make researching more efficient
- Problem Statement: This project aims to scrape and analyze submissions from two subreddits in order to train a classifier on which subreddit a post came from.

What Subreddits are We Scraping?

r/HipHopHeads

Subreddit Description: “Everything hip-hop, R&B and Future Beats! The latest mixtapes, videos, news, and anything else hip-hop/R&B/Future Beats related from your favorite artists”

r/Indieheads

Subreddit Description: “Everything Indie Music related; from the newest releases and news, to discussion on the history of alternative music”

Why these two subreddits?

- Since they’re both music subreddits, their content overlap to a certain extent
- Both popular subreddits with over a million subscribers and plenty of posts

The Process

Collect posts from r/HipHopHeads and r/Indieheads

Build a model that predicts which subreddit a post came from

Or

Scraping, Cleaning, and EDA

- 5000 posts per subreddit are scraped
- A dataframe for each subreddit is created, and then concatenated together
- Null values are dropped
- Dataframe is formatted

	subreddit	selftext	title	hiphop_or_indie	selftext_&_title
0	hiphopheads		[FRESH ALBUM] Elujay & J.Robb - GEMS IN TH...	1	[FRESH ALBUM] Elujay & J.Robb - GEMS IN TH...
1	hiphopheads		[FRESH VIDEO] DASHXX - SALAMALEIKUM	1	[FRESH VIDEO] DASHXX - SALAMALEIKUM
2	hiphopheads	[removed]	What are unpopular opinions you have?	1	[removed]What are unpopular opinions you have?
3	hiphopheads		[FRESH VIDEO] DryBoy Feat Clever - Summer Nights	1	[FRESH VIDEO] DryBoy Feat Clever - Summer Nights
4	hiphopheads		[FRESH] Abra Cadabra - Trenches (Official Video)	1	[FRESH] Abra Cadabra - Trenches (Official Video)

Preprocessing & Modeling

1. Features (selftext_&_title) and target (hiphop_or_indie) variables are selected
2. Train-test split is performed on these variables
3. GridSearch and Modeling
4. Analyze/Interpret each model's performance

Results

Multinomial Naive Bayes with CountVectorizer

- Train Score: 0.793
- Test Score: 0.785

Logistic Regression with CountVectorizer

- Train Score: 0.859
- Test Score: 0.823

Multinomial Naive Bayes with TFIDFVectorizer

- Train Score: 0.810
- Test Score: 0.791

Random Forest Classifier with CountVectorizer

- Train Score: 0.976
- Test Score: 0.842

Conclusion and Future Steps

- The Random Forest Classifier with a CountVectorizer performed the best
  - Train Score: 0.976
  - Test Score: 0.842
- This model should be used when predicting the origin of a post.

- Recommendations
  - Experiment with more parameters when testing this model
  - Test model on less popular subreddits
  - Preprocessing and modeling with a more powerful computer would produce results much sooner
  - Explore other methods (example: boosting)

Thank you!

Questions?

Sources

- https://www.thefader.com/
- https://www.reddit.com/r/hiphopheads/new/
- https://www.reddit.com/r/indieheads/new/
- https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
- https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html
- https://github.com/pushshift/api
- https://stackoverflow.com/questions/24386489/adding-words-to-scikit-learns-countvectorizers-stop-list
- DSI lessons 5.01-6.04