Northeastern University

Music Genres Classification Based on Lyrics

Member: Mingrui Zhang

Table of contents

01 Overview

02 Data Preprocessing and EDA

03 Modeling and Results

04 Conclusion and Future work

Overview

Rock Music

Pop Music

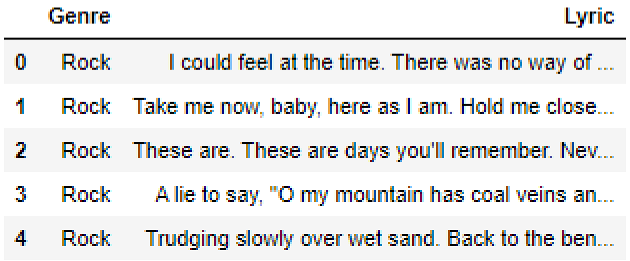
Hip Hop Music

A music genre or subgenre may be defined by the musical techniques, the cultural context, and the content and spirit of the themes.

------Wikipedia

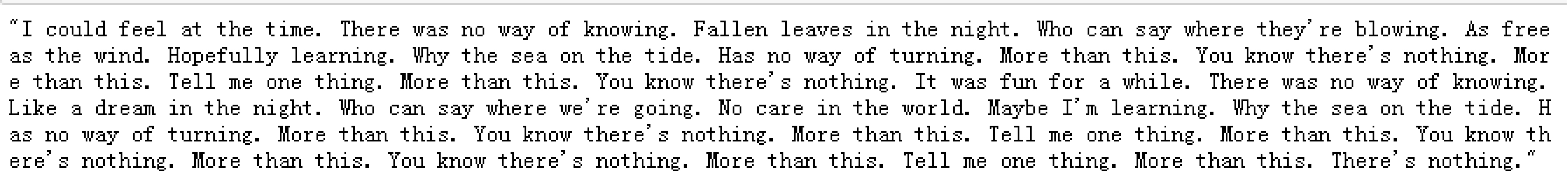
- *What is the differences and similarities of these three genres on lyrics?*
- *Can they be classified by machine learning techniques?*

Data Preprocessing and EDA



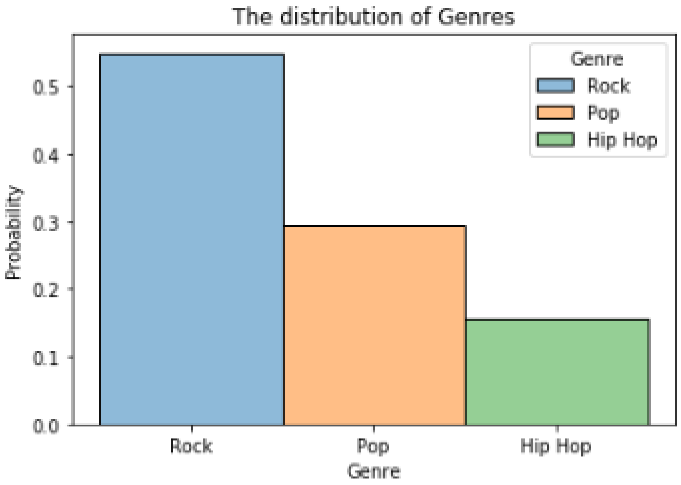
	Genre	Lyric
0	Rock	I could feel at the time. There was no way of ...
1	Rock	Take me now, baby, here as I am. Hold me close...
2	Rock	These are. These are days you'll remember. Nev...
3	Rock	A lie to say, "O my mountain has coal veins an...
4	Rock	Trudging slowly over wet sand. Back to the ben...

This dataset is from Kaggle. It has two raw datasets. This is what it looks like after merging ,filtering and removing some irrelevant data.



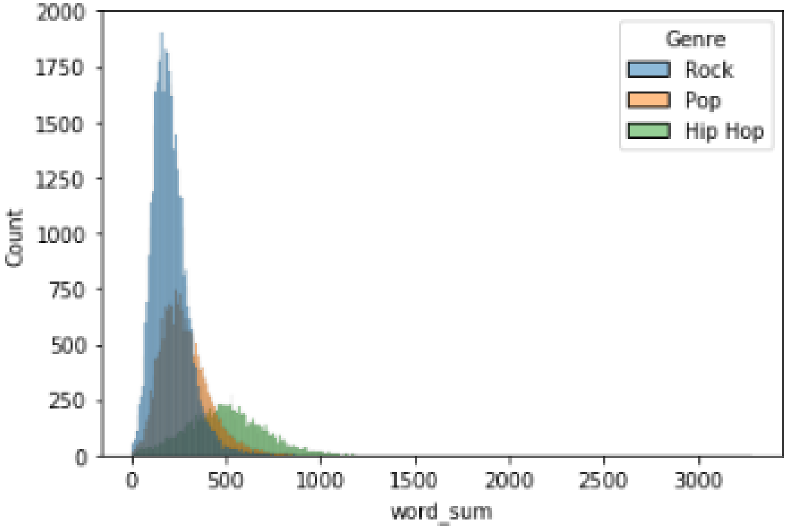
“I could feel at the time. There was no way of knowing. Fallen leaves in the night. Who can say where they’re blowing. As free as the wind. Hopefully learning. Why the sea on the tide. Has no way of turning. More than this. You know there’s nothing. More than this. Tell me one thing. More than this. You know there’s nothing. It was fun for a while. There was no way of knowing. Like a dream in the night. Who can say where we’re going. No care in the world. Maybe I’m learning. Why the sea on the tide. Has no way of turning. More than this. You know there’s nothing. More than this. Tell me one thing. More than this. You know there’s nothing. More than this. You know there’s nothing. More than this. Tell me one thing. More than this. There’s nothing.”

Data Preprocessing and EDA



This data set has 86294 samples. From the figure on the left, we can see that the data is imbalanced. The percentage of Rock music is over 0.5. Deal with imbalance issue later

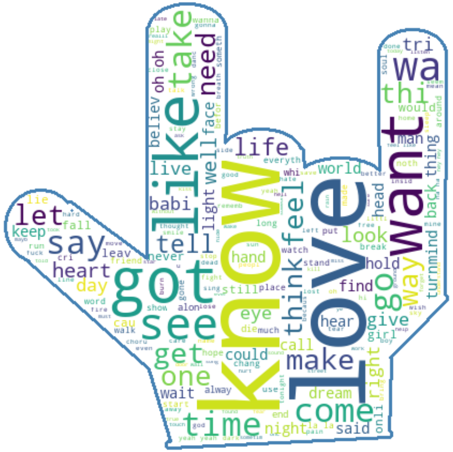
Data Preprocessing and EDA



After tokenizing and stemming, count how many words in each song.

- Hip Hop tends to have more words in a song.(green area)

Data Preprocessing and EDA



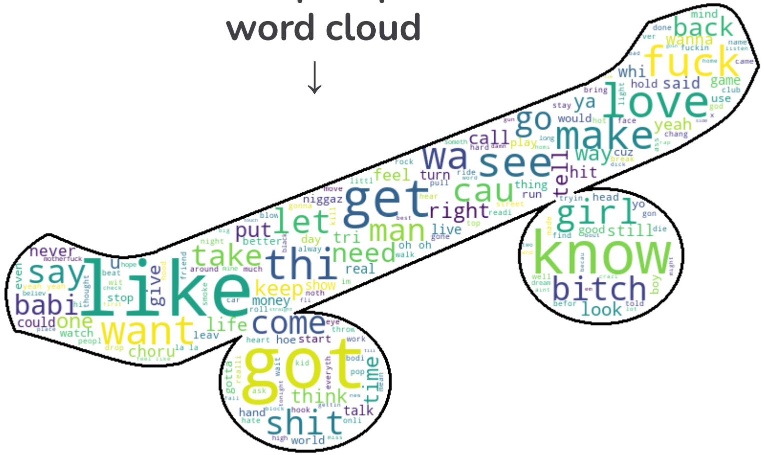
↑

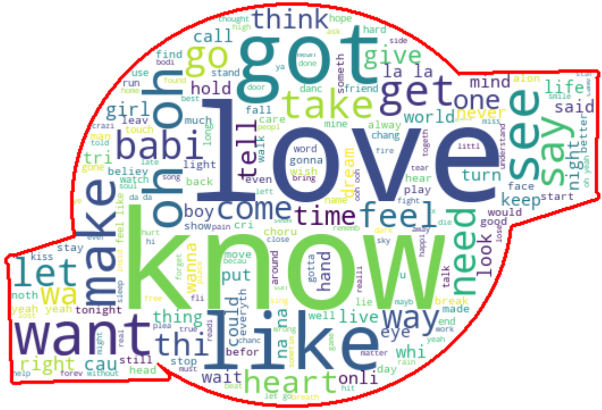
Rock word cloud

- After dropping stop words
- Word clouds will show words with highest frequency in each genre.
- Many words show on all three word clouds, like 'love', 'know', 'like', 'got'.

Hip Hop word cloud

↓



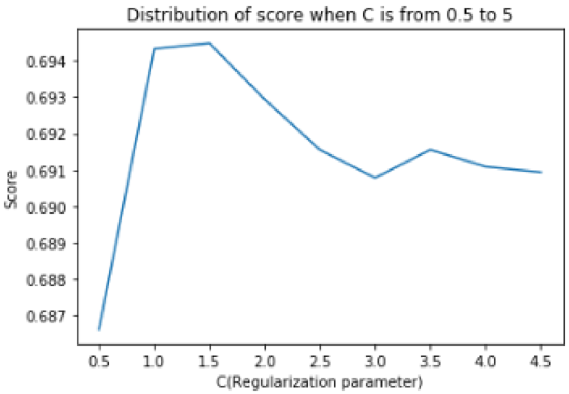


↑

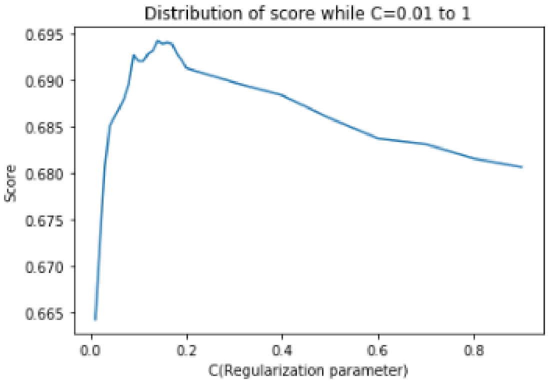
Pop word cloud

Modeling and Results

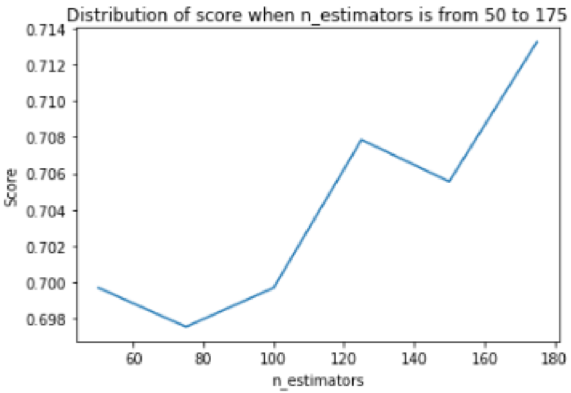
- Split data into training data and test data. The size of test set is 0.2.
- Use TFIDF vectorizer to convert text into numerical vectors(max_feature =5000)
- Apply under_sampling method to training data for dealing with data imbalanced issue.
- Spare 20% training data as validation data.
- Build three models: Logistic regression, LinearSVM and Random Forest. Fit data on train set and do hyperparameter tuning on validation set.
- Choose the best hyperparameters and get results on test data.



Logistic regression



LinearSVM

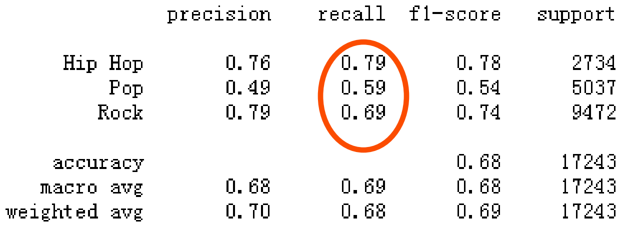


Random Forest

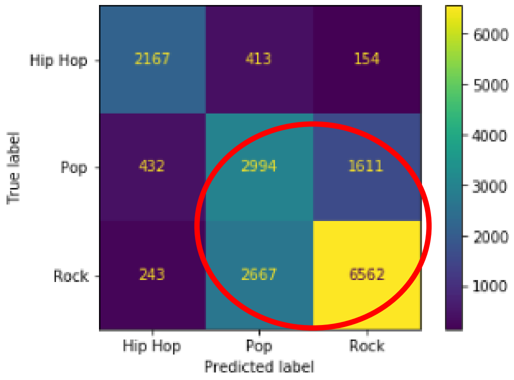
Score: mean accuracy on given data and labels.

Modeling and Results

Logistic regression

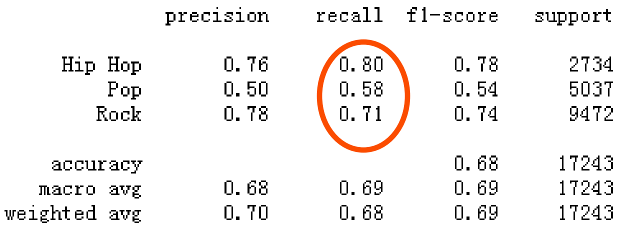


```
              precision    recall  f1-score   support

     Hip Hop       0.76      0.79      0.78      2734
         Pop       0.49      0.59      0.54      5037
        Rock       0.79      0.69      0.74      9472

    accuracy                           0.68     17243
   macro avg       0.68      0.69      0.68     17243
weighted avg       0.70      0.68      0.69     17243
```

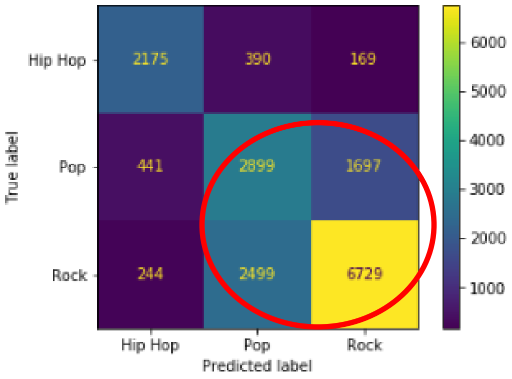


	Hip Hop	Pop	Rock
Hip Hop	2167	413	154
Pop	432	2994	1611
Rock	243	2667	6562

LinearSVM

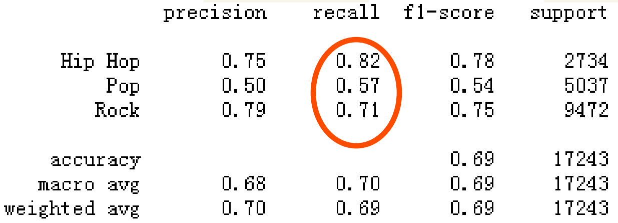


```
              precision    recall  f1-score   support

     Hip Hop       0.76      0.80      0.78      2734
         Pop       0.50      0.58      0.54      5037
        Rock       0.78      0.71      0.74      9472

    accuracy                           0.68     17243
   macro avg       0.68      0.69      0.69     17243
weighted avg       0.70      0.68      0.69     17243
```

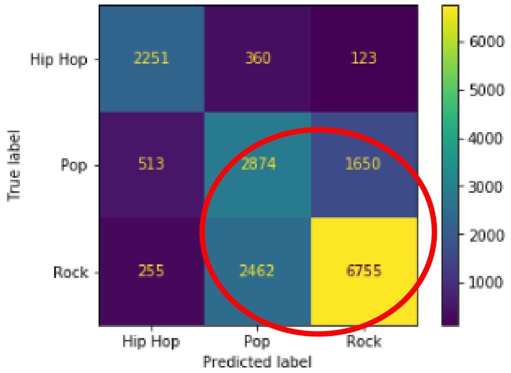


	Hip Hop	Pop	Rock
Hip Hop	2175	390	169
Pop	441	2899	1697
Rock	244	2499	6729

Random Forest



```
              precision    recall  f1-score   support

     Hip Hop       0.75      0.82      0.78      2734
         Pop       0.50      0.57      0.54      5037
        Rock       0.79      0.71      0.75      9472

    accuracy                           0.69     17243
   macro avg       0.68      0.70      0.69     17243
weighted avg       0.70      0.69      0.69     17243
```

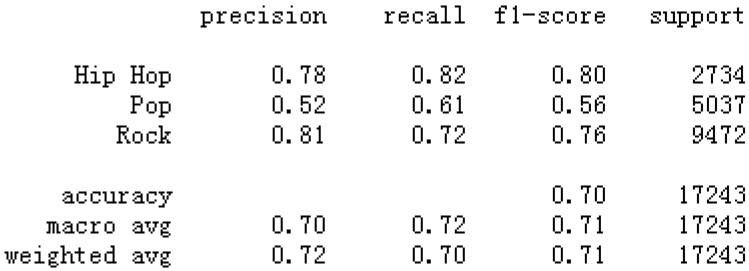


	Hip Hop	Pop	Rock
Hip Hop	2251	360	123
Pop	513	2874	1650
Rock	255	2462	6755

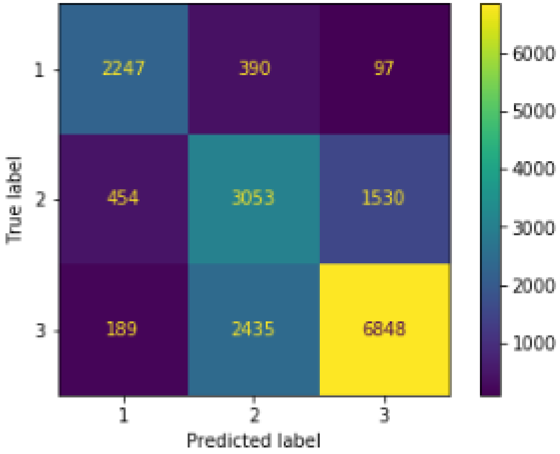
1. *Random Forest performs slightly better than other two classifiers.*
2. *Hip hop can get highest recall. Pop get worst recall on three classifiers.*
3. *Rock and Pop are difficult to classify. Large part of pop music is classified as Rock.*

Modeling and Results

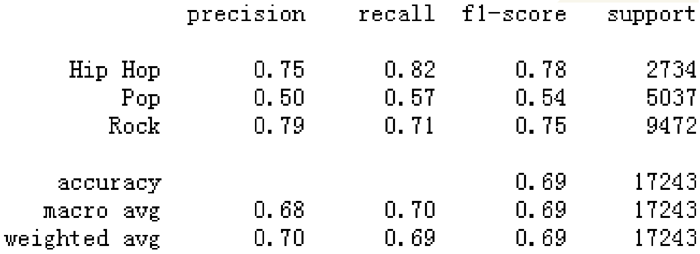
Stacking classifier



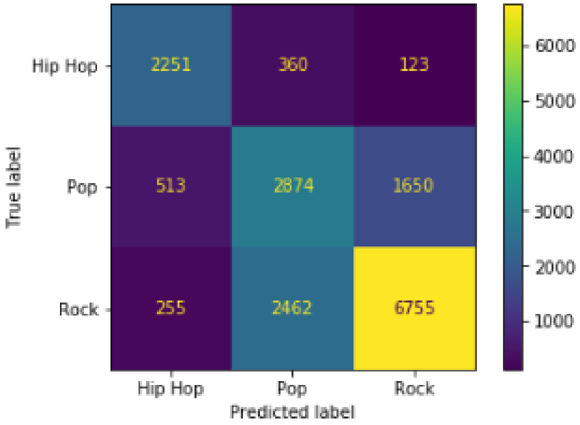
```
              precision    recall  f1-score   support

     Hip Hop       0.78      0.82      0.80      2734
         Pop       0.52      0.61      0.56      5037
        Rock       0.81      0.72      0.76      9472

    accuracy                           0.70     17243
   macro avg       0.70      0.72      0.71     17243
weighted avg       0.72      0.70      0.71     17243
```



Random Forest



```
              precision    recall  f1-score   support

     Hip Hop       0.75      0.82      0.78      2734
         Pop       0.50      0.57      0.54      5037
        Rock       0.79      0.71      0.75      9472

    accuracy                           0.69     17243
   macro avg       0.68      0.70      0.69     17243
weighted avg       0.70      0.69      0.69     17243
```



1. *Using stacking, model performance improves a little.*
2. *Still have difficulty on classifying Rock and Pop.*

Conclusion and Future work

1. Hip hop music is the most distinctive one among three genres.
2. Many words in lyrics appear highly frequently in three genres.
3. Rock and Pop music are difficult to classify correctly

   -------- don't have clear boundary on lyrics

   -------- Use more complex models like LSTM, use more data to train.
4. For the topic of music genres classification, lyrics are not enough. Introducing audio data may help a lot.

Thanks!

Does anyone have any questions?