Song Genre AI

Deducing Music Genres Based on Song Lyrics

Paweł Gościak, Marceli Grad, Jan Krawczyk

Table of contents

Introduction

The goal of our project is to create an artificial intelligence system capable of recognizing the music genre based on the lyrics of a song. Music has always been an integral part of our lives, and the variety of music genres is immense. Our AI aims to automatically analyze song lyrics and assign them to the appropriate music genre, taking into account various features such as theme, style, and word count.

Data

Database

The data we used to train our AI was sourced from *www.kaggle.com*. It includes a large database of artists along with the genres of music they perform and a database of songs containing lyrics. After merging both databases and filtering out irrelevant data, we performed lemmatization on the remaining lyrics.

Lemmatization

Lemmatization is the process of reducing words to their base form, known as a lemma, by removing inflectional endings while considering the semantic and grammatical properties of the word.

```
I feel so unsure
As I take your hand and lead you to the dance floor
As the music dies, something in your eyes
Calls to mind a silver screen
And all those sad goodbyes

I'm never gonna dance again
Guilty feet have got no rhythm
Though it's easy to pretend
I know you're not a fool

Should've known better than to cheat a friend
And waste the chance that I've been given
So I'm never gonna dance again
The way I danced with you

Time can never mend
The careless whispers of a good friend
To the heart and mind
Ignorance is kind
There's no comfort in the truth
Pain is all you'll find
```

```
I feel so unsure
as I take your hand and lead you to the dance floor
as the music die , something in your eye
call to mind a silver screen
and all those sad goodbye

I be never go to dance again
guilty foot have get no rhythm
though it be easy to pretend
I know you be not a fool

should 've know well than to cheat a friend
and waste the chance that I 've be give
so I be never go to dance again
the way I dance with you

Time can never mend
the careless whisper of a good friend
to the heart and mind
ignorance be kind
there be no comfort in the truth
Pain be all you will find
```

Image 1: "Careless whisper" before and after lemmatization.

Further Text Processing

The next step we took was to discard words that, due to their meaning, could not definitively indicate a specific music genre, as well as removing all punctuation and musical symbols.

```
['the', 'as', 'i', 'be', 'a', 'you', 'to', 'and', 'it', 'not', 'do', 'in', 'my', 'us', 'of',
 'your', 'know', "'", 'so', 'love', 'but', 'no', 'yes', '?', 'he', 'she', 'we', 'make', 'if',
 "'ve", 'want', '!', 'well', '"', 'could', 'from', 'would', "'s", 'at', '...', 'her', 'his',
 'all', 'around', 'then', 'when', 'they', 'them', 'into', 'an', ':', 'their', 'those', 'these',
 'this', 'mine', 'too', 'through', 'who', 'how', 'why', 'until', 'unless', 'that', 'with', 'on',
 'or', 'will', "won't", "can't", "haven't", "isn't", 'have', 'what', 'by', 'there', 'here',
 'which', 'whom', 'whose', 'some', 'than', 'like', 'also', 'because', '!', 'each', 'during', '(',
 ')', '[', ']', u"\u2122", 'soon', 'although', 'however', 'let', 'get', 'go', 'come', 'can',
 'take', 'our', '.', '..', '*', '_', '+', '/', 'a1', 'a2', 'a3', 'a4', 'a5', 'a6', 'a7', 'b1', 'b2',
     'b3', 'b4', 'b5', 'b6', 'b7', 'c1', 'c2', 'c3', 'c4', 'c5', 'c6', 'c7', 'd1', 'd2', 'd3', 'd4', 'd5',
     'd6', 'd7', 'e1', 'e2', 'e3', 'e4', 'e5', 'e6', 'e7', 'f1', 'f2', 'f3', 'f4', 'f5', 'f6', 'f7', 'g1',
     'g2', 'g3', 'g4', 'g5', 'g6', 'g7', 'h1', 'h2', 'h3', 'h4', 'h5', 'h6', 'h7']
```

Image 2: Words removed from the lyrics.

Final Product

At the end of our data processing, we obtained the following data:

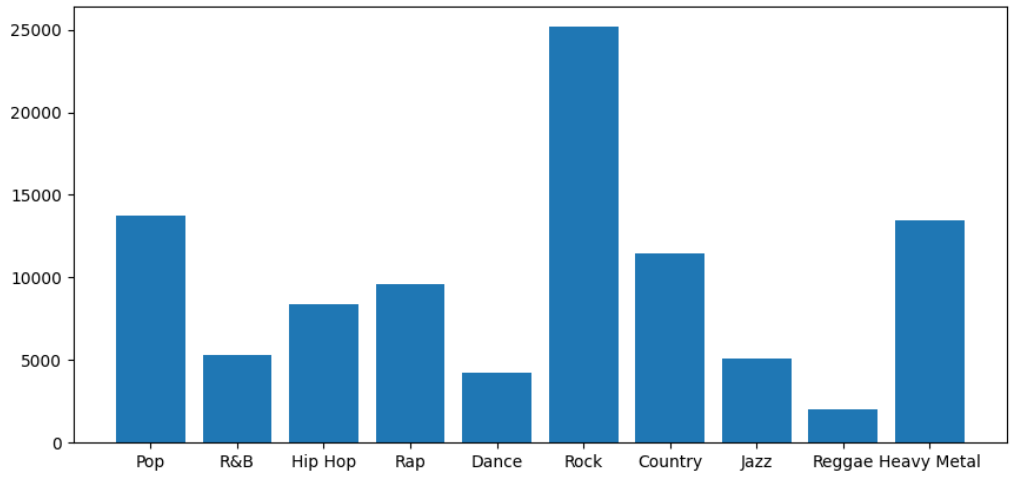

Image 3: Distribution of lyrics by genre.

```
Pop  =  13759
R&B  =  5309
Hip Hop  =  8412
Rap  =  9589
Dance  =  4252
Rock  =  25177
Country  =  11432
Jazz  =  5124
Reggae  =  1990
Heavy Metal  =  13496
```

Image 4: Distribution of lyrics by genre.

Algorithm

Naive Bayes Classifier

We implemented our AI using a Naive Bayes classifier, which uses Bayes' theorem to calculate the probability that a song belongs to a specific music genre based on the frequency of words within those genres and the frequency of the genre in our database.

For example, the probability that a song with the lyrics "love her" belongs to the pop genre is calculated as follows:

$$P("pop") = P("pop\ gerne") \times P("love") \times P("her")$$

where:

$$P("love") = \frac{"love"}{"sum\ of\ words\ in\ 'pop'\ songs"}$$

$P("pop\ gerne")$ - is the probability of a song being a pop song in the training database.
$P("love")$ - is the probability of the word "love" appearing in all pop songs in the training database.
$P("her")$ - is the probability of the word "her" appearing in all pop songs in the training database.

To implement this algorithm, we created a class that contained the following fields:

```
Naive bayes algorithm


Class deciding which genre song belongs to base
on popular words in different genres.

train_data: data used for training, type: pandas.core.frame.DataFrame
categories: list of different music genre, type: list
genre probability: probability of occurrence of different music genre in base, type: dict
word_genre_dictionaries: number of occurrences of words in different genres, type: dict
total_genre_words: total number of words in the genre, type: dict
"""
```

Image 5: Description of the class

Then, using the Counter library, we counted the occurrences of all words within each music genre.

For the training process, we used 80% of our prepared database, while the remaining 20% was reserved for testing.

Results

Achived results

After implementing the Naive Bayes classifier, the accuracy we achieved was approximately 46%. We considered this result satisfactory given that random selectionamong the provided genres would yield only about a 10% success rate.

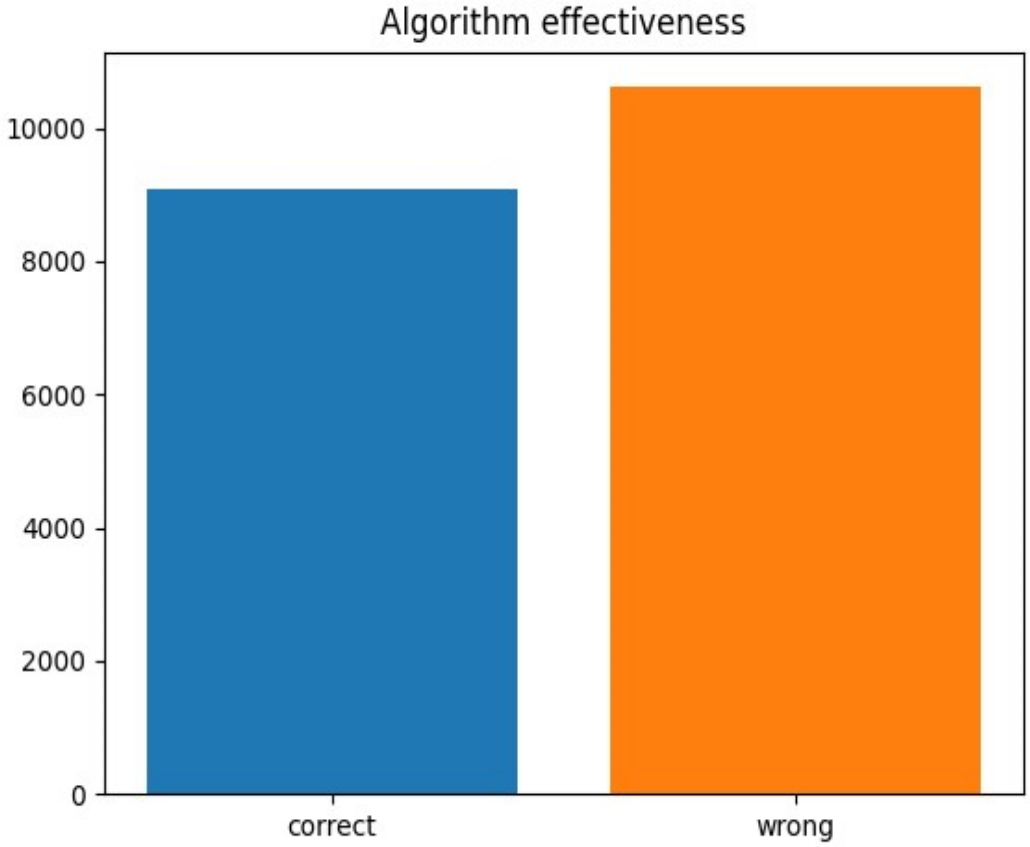

Image 6: Graphical representation of the algorithm's accuracy.

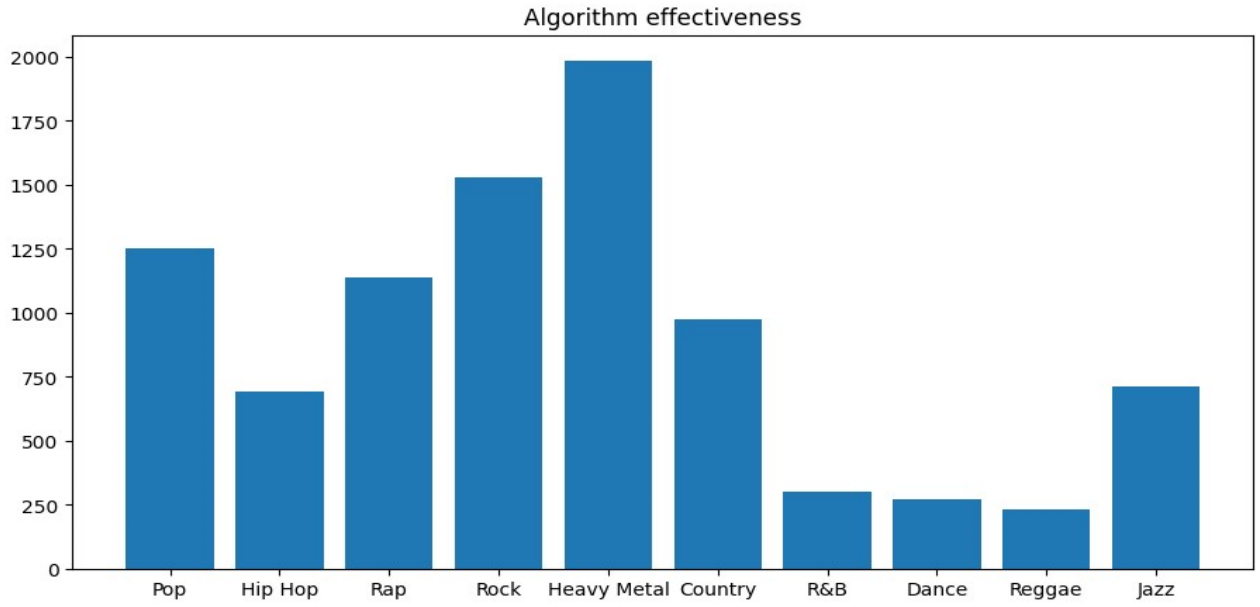

Image 7: Graphical representation of the algorithm's accuracy by music genre.

Potential improvements

Although our algorithm achieved results that we deemed satisfactory, there are ways to potentially enhance its performance, such as:

Summary

In conclusion, the Naive Bayes classifier is one of the best algorithms for our project, as its accuracy significantly exceeds the random assignment of genres to song lyrics.

Acquired knowledge

During this project, we studied various topics related to text processing, lemmatization, Bayes' theorem, and class instance serialization. We learned about several essential libraries, such as pandas for handling CSV files, pickle for class instance serialization, and tkinter for creating graphical interfaces.

Overcome challenges

During the implementation, we encountered several challenges, which we managed to overcome during the project. Examples of such challenges include:

- Limitations of pickle files – since our class stored many dictionaries, we could not serialize them all using just the pickle class. As a result, we also used JSON files.
- Choosing a library for lemmatization – since the lemmatization process for our initial database took over 72 hours, we had to test various libraries and choose the one that best balances performance and accuracy.

References

- https://betterprogramming.pub/predicting-a-songs-genre-using-natural-language-processing-7b354ed5bd80 - inspiration
- https://www.kaggle.com/datasets/neisse/scrapped-lyrics-from-6-genres?select=lyrics-data.csv - data
- https://looka.com/ - logo