Министерство науки и высшего образования РФ

Федеральное государственное бюджетное образовательное учреждение высшего образования

«Московский Авиационный Институт»

Национальный Исследовательский Университет

Институт №8 «Информационные технологии и прикладная математика»

Кафедра 806 «Вычислительная математика и программирование»

Лабораторная работа №0

по курсу «Машинное обучение»

Студент:	Хренникова А. С.
Группа:	М8О-308Б-19
Преподаватель:	Ахмед Самир Халид
Подпись:	
Оценка:	
Дата:	

Москва, 2022

Лабораторная работа №0

Глобальная задача “стартапа”: Создать проект, который по различным музыкальным характеристикам, в том числе и тексту песен, сможет для каждой отдельной композиции определять ее тему, смысловую нагрузку и посыл. Данная задача поможет формировать плейлисты, подбирать музыку по вкусу, что актуально с уходом Spotify с российского рынка. В качестве точки отсчета необходимо научиться классифицировать треки по тематике, опираясь на их характеристики(надо же с чего то начинать).

Датасет: Музыка 1950 - 2019 годов. Набор данных содержит в себе список песен с 1950 по 2019 год и имеет такие характеристики как: имя исполнителя, название трека, дата релиза, жанр, текст песни, продолжительность, энергичность, танцевальность, тематика, шумность, инструментальность, способность встряхнуть публику, романтичность, акустика, содержание насилия, непристойность, подходит для семейного прослушивания, привлекательность для девушек, духовность, печальность, чувственность, подходит для свидания, музыкальность, совместимость со световым сопровождением, подходит для ночного времени.

Выбранная задача: Классифицировать песни по тематике на 8 классов: чувства, романтика, непристойности, грусть, насилие, музыка, время, мир/жизнь.

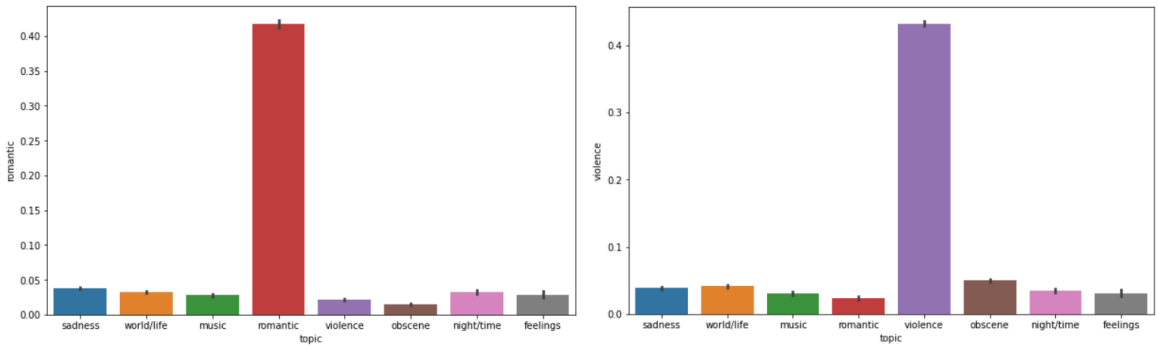
Задание: Проанализировать и визуализировать данные.

Москва, 2022

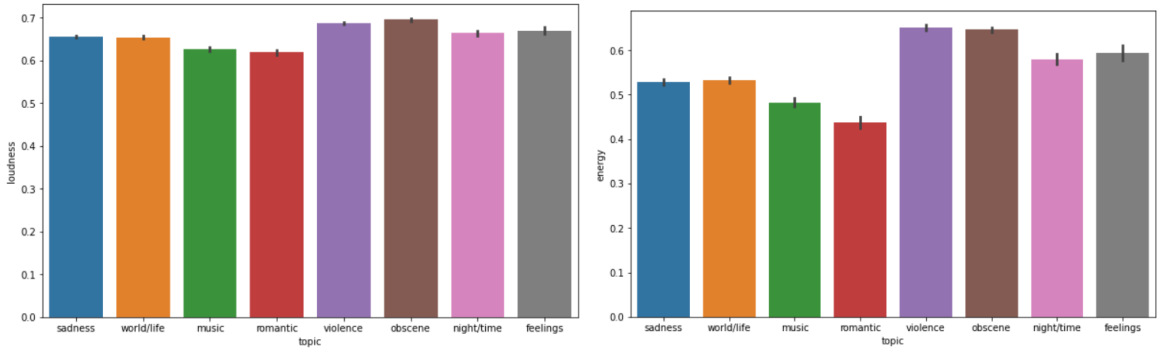
1 Описание

Датасет имеет размеры 31*28372, где 30 признаков, 5 из которых текстовые, 2 принимают целые значения, 23 значения в пределах от 0 до 1. Среди наших данных нет отсутствующих значений, поэтому можем сказать, что у нас неплохой датасет.

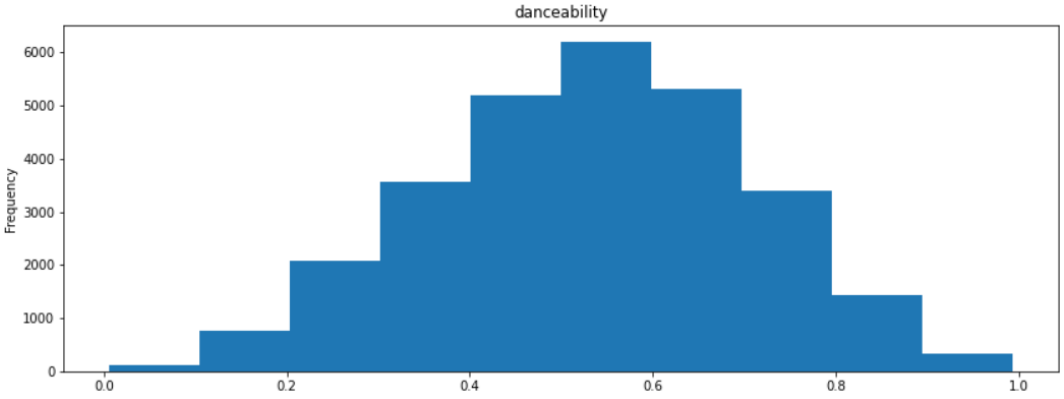
На данных графиках представлены зависимость тематики от таких параметров, как ‘romantic’ и ‘violence’. По данным критериям можно точно определить тематику, к которой относится трек. Всего таких критериев 8, их названия соответствуют классам тематики.



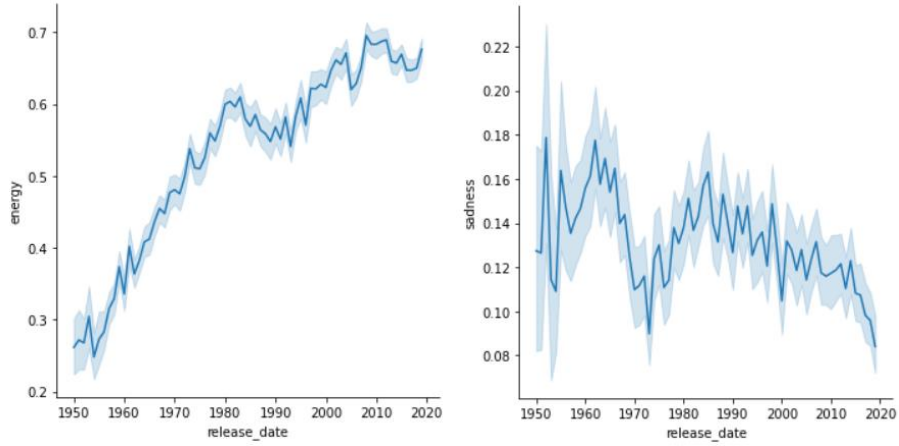
Если смотреть на другие критерии, то только по ним достаточно сложно определить тематику.



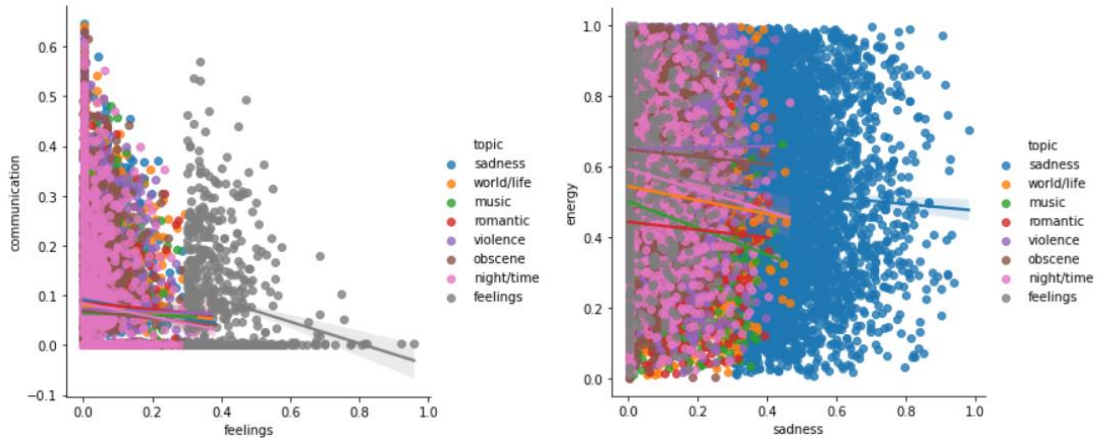
Так же можно посмотреть на распределение значений конкретных параметров, например параемтр ‘dancebility’:



Также прослеживается зависимость между количественными признаками и датой релиза трека:

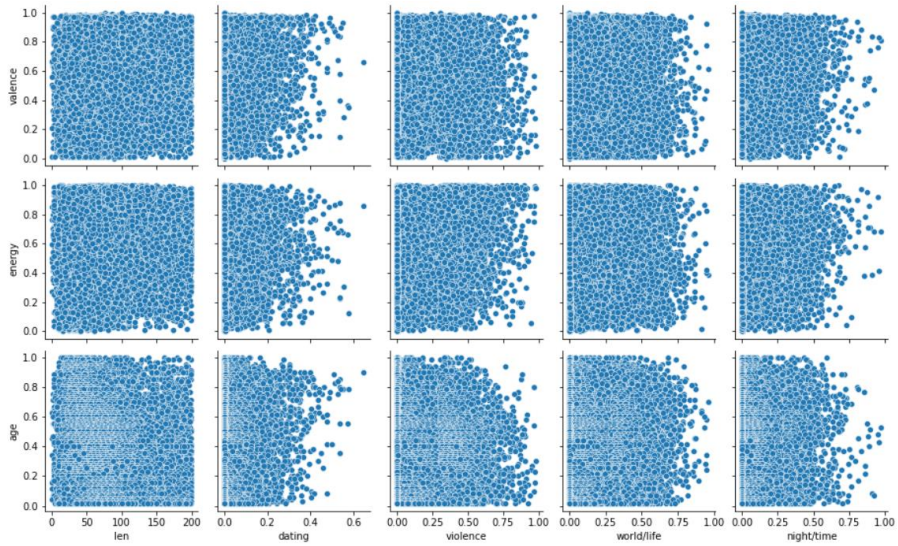


Если рассматривать регрессионную модель для двух количественных параметров, то можно заметить, что также достаточно четно выделяется один класс. Подобное работает и с другими параметрами.

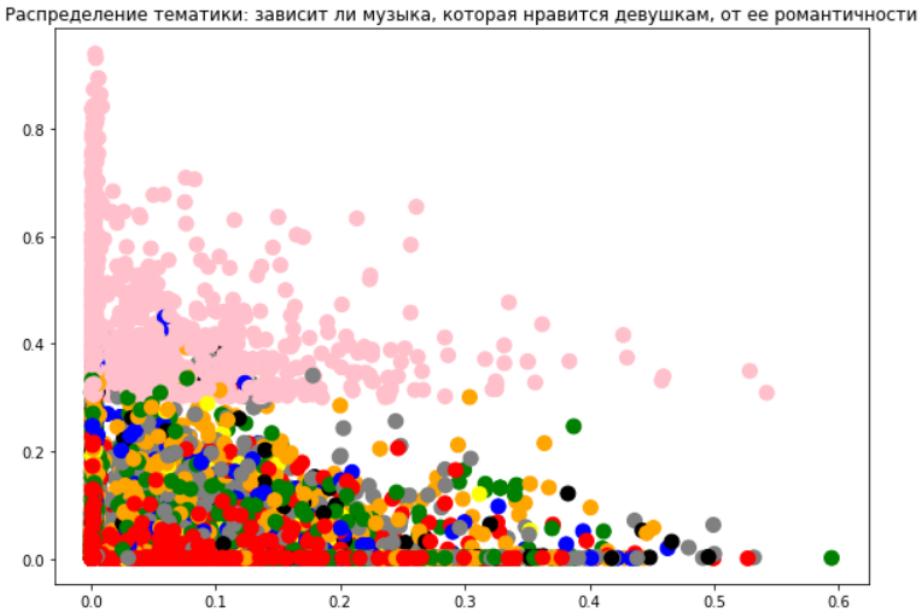


Корреляция не обнаруживается между количественными признаками:

Москва, 2022



Можно выделить зависимость между тематикой и количественными признаками:



Для удобства работы с данными, качественные признаки, имеющее ограниченное количество значений, можно представить в виде распределения значений от 0 до 1, где определенное значение будет соответствовать определенному числу, то есть в формате one-hot-encoding.

Москва, 2022

	Column_Name	Type	Num_Unique
4	genre	object	7
29	topic	object	8
30	age	float64	70
3	release_date	int64	70
6	len	int64	199
23	danceability	float64	859
27	valence	float64	1295
28	energy	float64	1348
25	acousticness	float64	3786
26	instrumentalness	float64	4939

Затраты по памяти:

```
Average memory usage for float columns: 0.21 MB
Average memory usage for int columns: 0.16 MB
Average memory usage for object columns: 3.49 MB
```

2 Выводы:

Есть признаки, благодаря которым можно точно классифицировать песни по той или иной тематике.

Данных достаточно много, они неплохого качества, поэтому их хватит для того, чтобы выполнять классификацию.

Часть данных, например текст песни, автор и название, можно выкинуть, они не влияю на тематику композиций. Другие качественные характеристики представить в числовом виде.