ОПРЕДЕЛЕНИЕ ЖАНРА МУЗЫКИ НА ОСНОВЕ ЕЁ ТЕКСТА

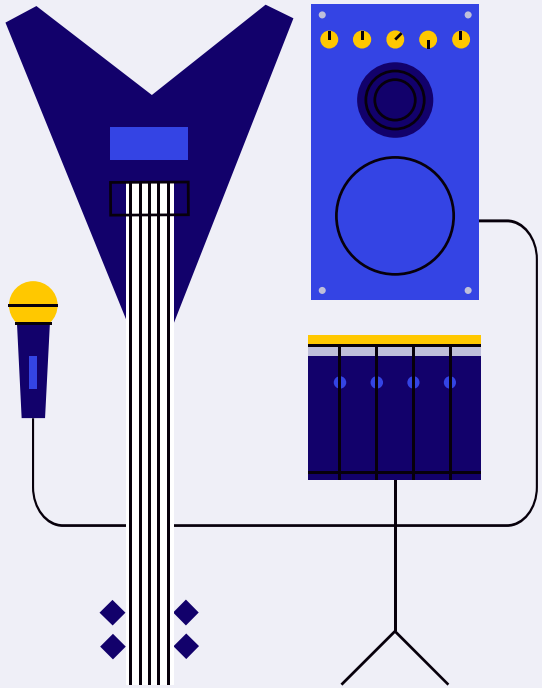
Кряжев Макар

Дю Вадим

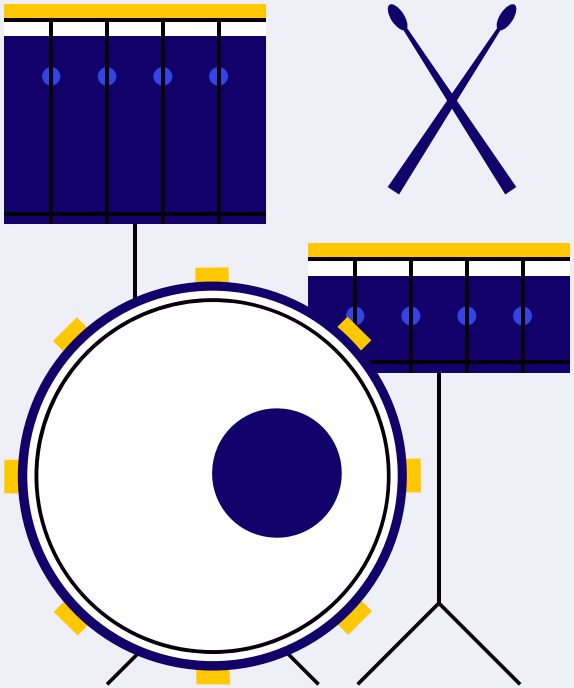
Талпа Григорий

Сухова Екатерина

Ерохин Дмитрий



Куратор: Тимофей Соборнов



СБОР И РАЗМЕТКА ДАННЫХ

Первоначальный план

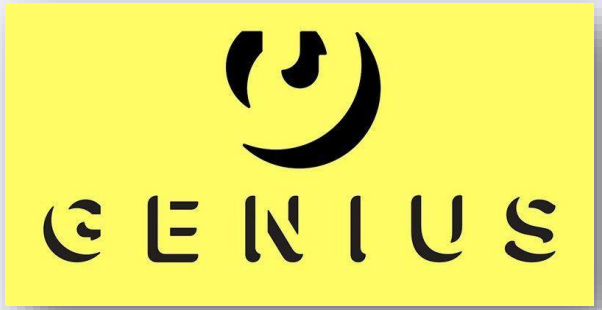
1. Скрапинг текстов и жанров песен с genius.com
2. Profit

Но...

“Жадный” Genius API, отсутствие альтернативного источника информации

Принцип работы

Основная информация



Текст песни

Жанр

Недостатки

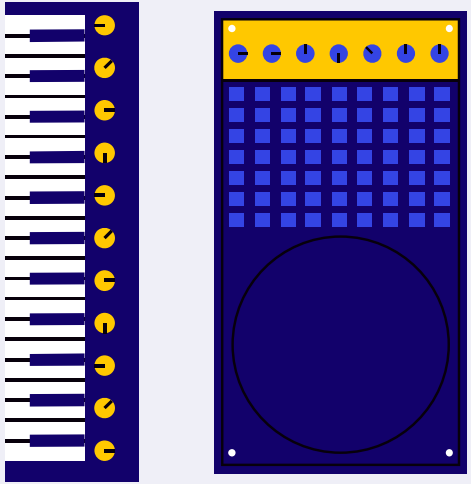
- Неэффективность (перебор id, 3 источника)
- Медлительность (150 песен/25 минут)

Готовый датасет

- ≈ 6 миллионов строк
- 8 признаков
- 6 жанров (rap, rock, country, rb, pop, misc)

Готовый датасет

title	Название. Большинство записей – песни, но также присутствуют книги, поэмы и т.д.
tag	Жанр. Немузыкальные записи имеют жанр misc.
artist	Автор(ы) произведения
year	Год выпуска
views	Количество просмотров страницы на genius.com
features	Артисты, участвовавшие в создании
lyrics	Текст произведения
id	Идентификатор произведения на genius.com



ОБРАБОТКА

Обработка

- **Предварительная обработка**

  Балансировка классов

  Удаление пунктуации

- **Необходимая подготовка**

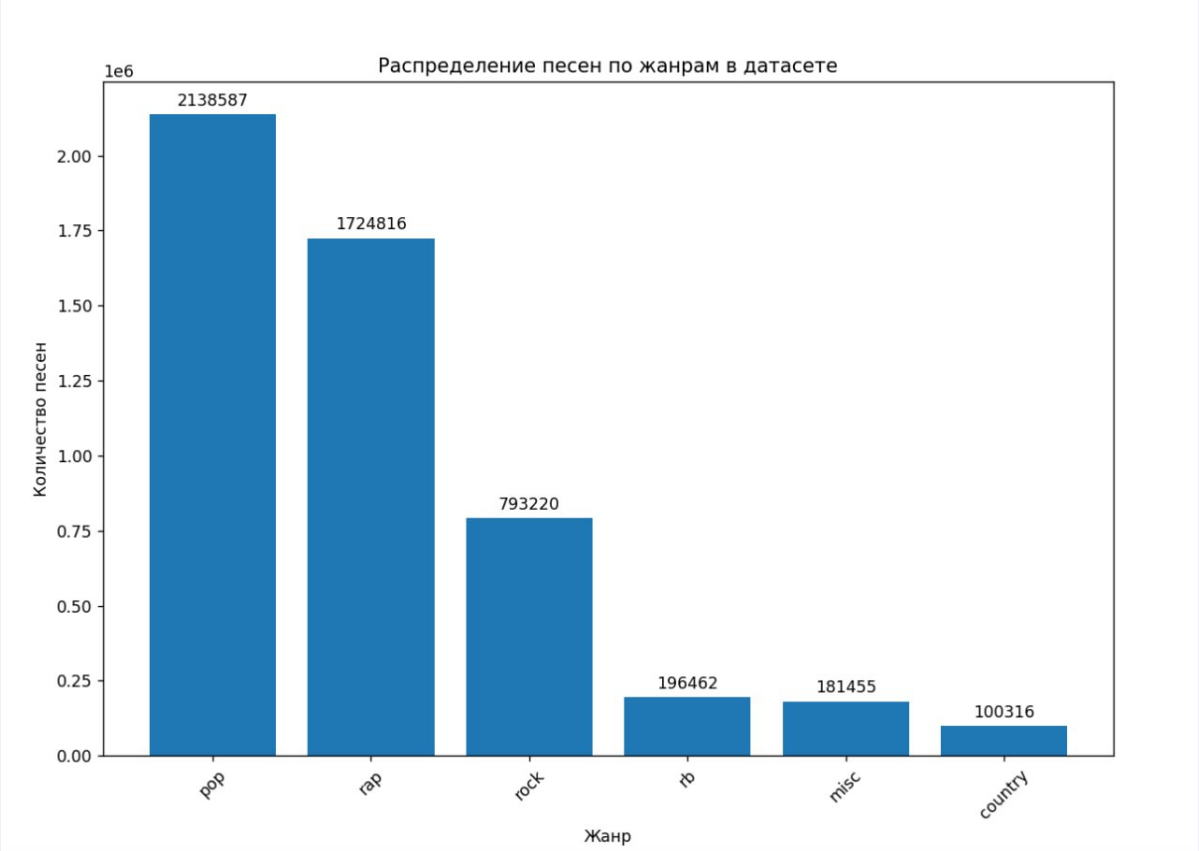
  Токенизация

  Лемматизация

  Векторизация (B-O-W, Tf-Idf, word2vec,...)

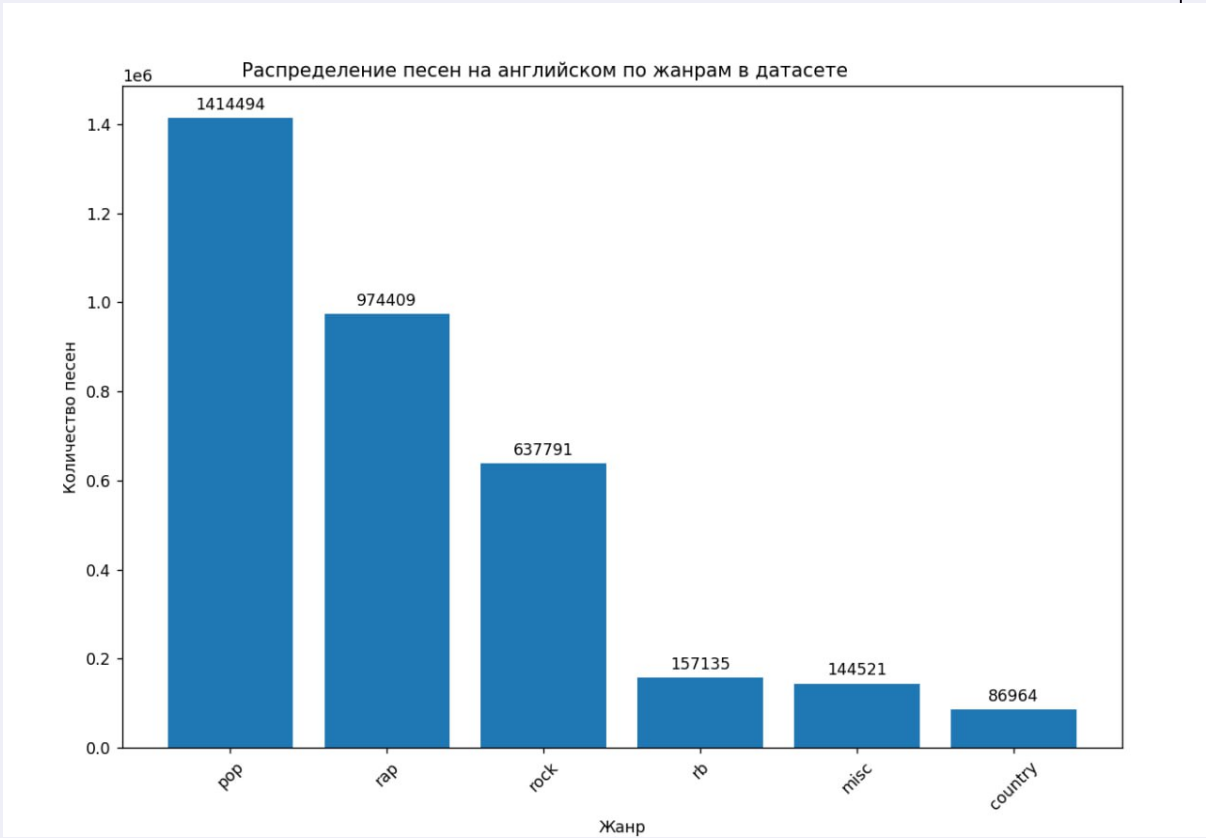
Обработка

Убираем все песни кроме английских



Балансировка классов

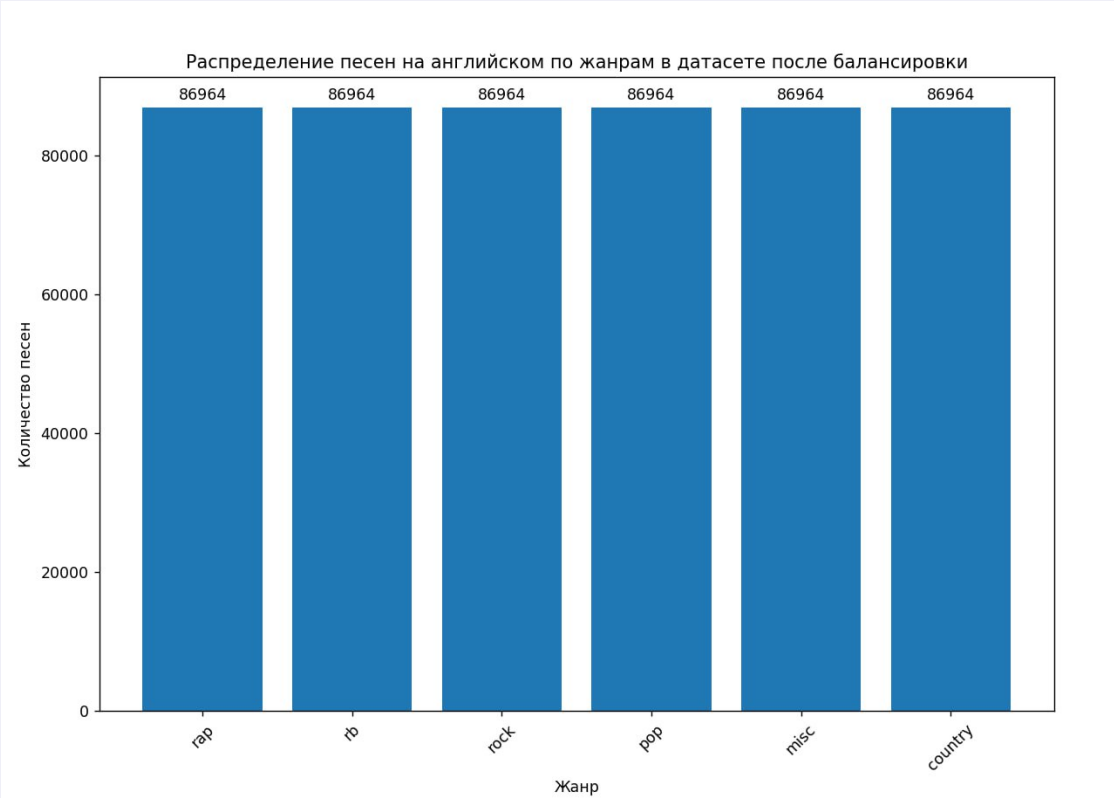
Уменьшаем количество данных

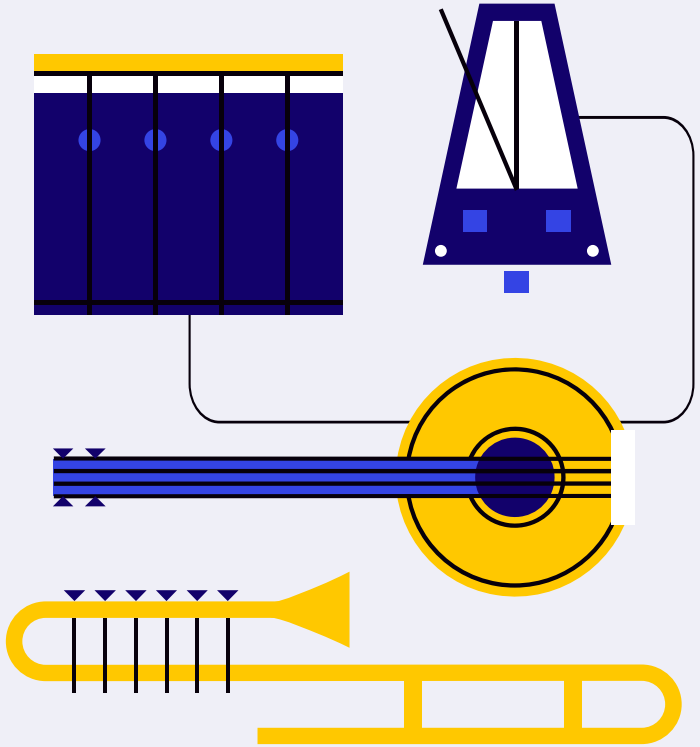


Балансировка классов

Балансируем классы относительно минимального

≈ 87 000 строк в каждом жанре





EDA

Облака слов для каждого жанра

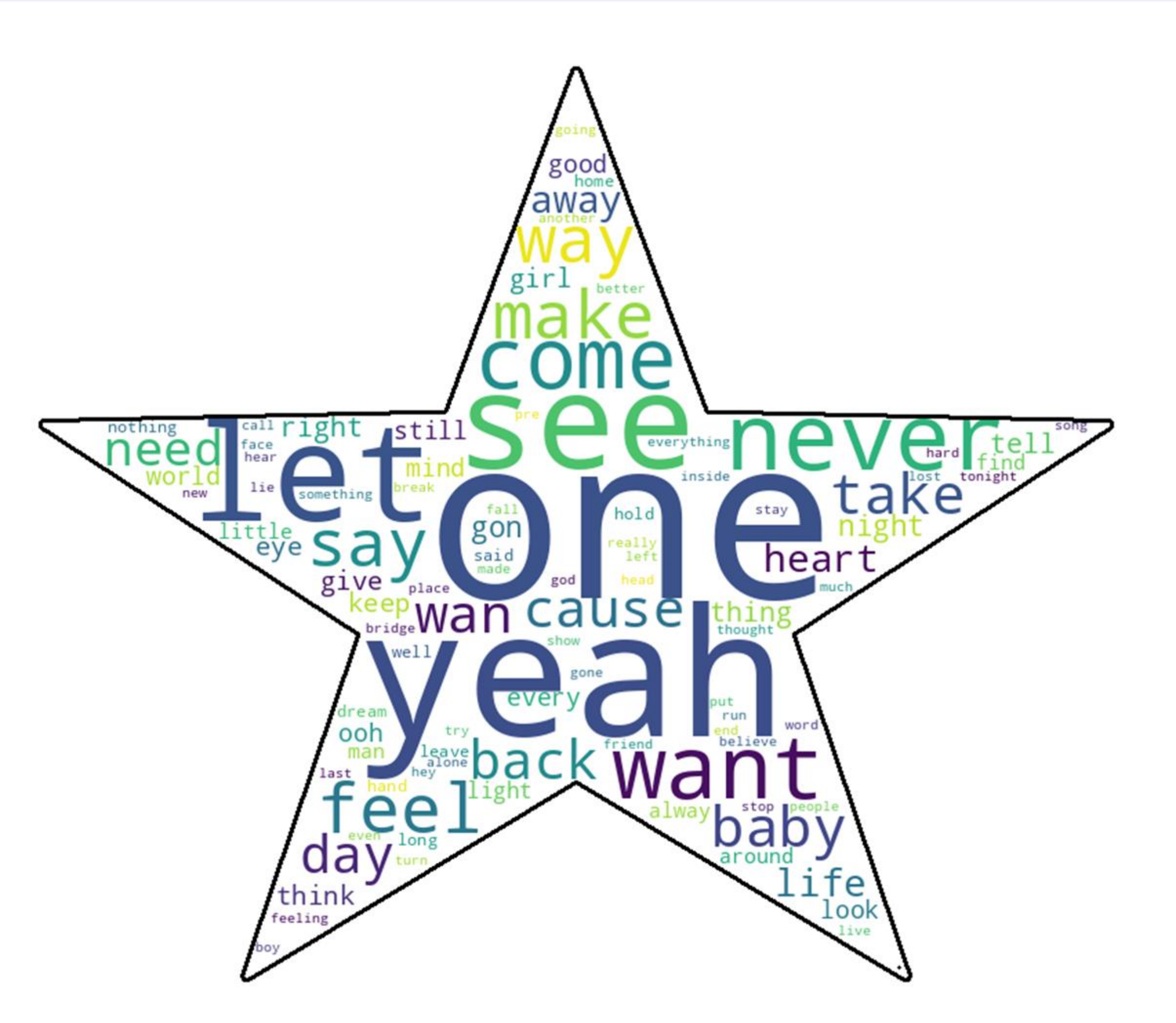
При построении не были использованы следующие слова:

- verse
- chorus
- love
- time
- know
- got

country

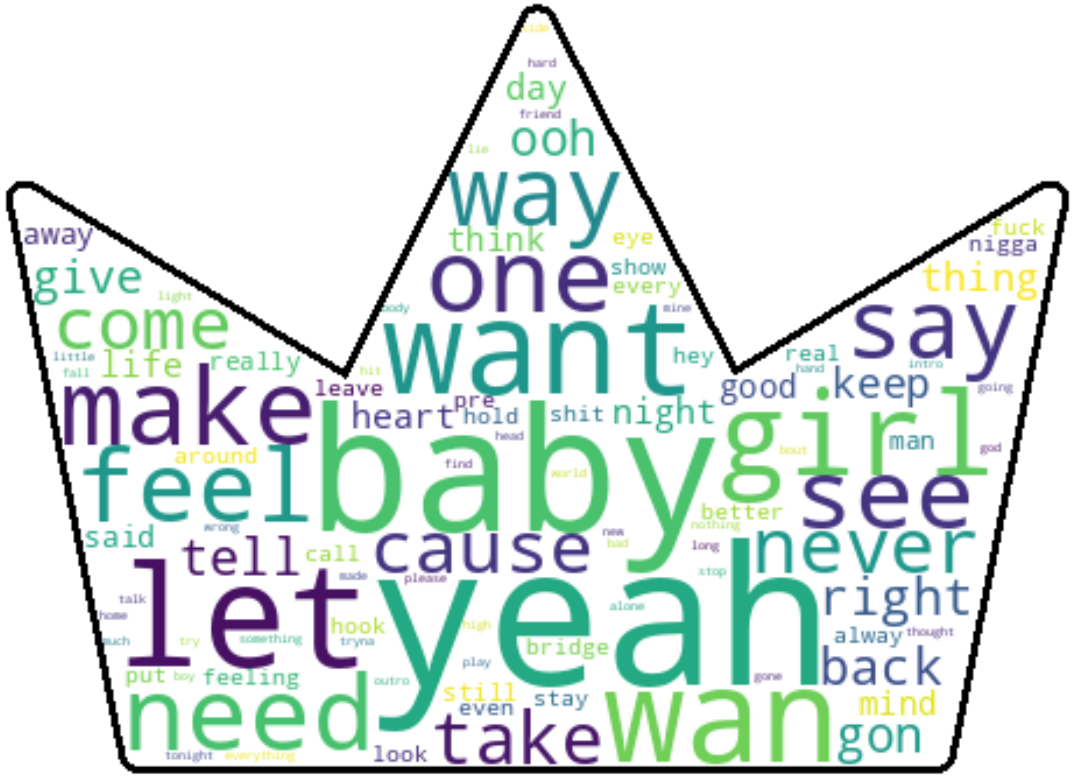

blue
make
still
home
come
say
hand
last
mind
well
hold
never
need
light
world
feel
wan
lord
little
gone
old
look
turn
life
keep
thought
thing
everything
bridge
day
one
way
away
two
said
sing
place
made
leave
mine
better
another
road
tonight
nothing
put
yeah
song
night
see
baby
girl
going
give
ooh
dream
eye
tell
hard
sun
might
find
every
want
gon
town
boy
name
always
fall
take
long
hear
call
big
stay
sweet
even
head
alone
god
new
high
outro
much
think
cause
heart
good
live
let
man
right
run
left
friend
something
around
hey

pop



going
good
home
away
another
way
girl
better
make
come
see
pre
nothing
call
right
still
never
song
need
face
hear
everything
tell
hard
let
mind
find
world
lie
break
one
inside
lost
tonight
new
something
take
fall
hold
stay
little
say
gon
night
eye
really
said
left
heart
made
give
place
god
head
much
keep
wan
cause
thing
bridge
thought
show
well
yeah
gone
every
dream
put
run
ooh
try
end
word
man
leave
friend
believe
back
alone
want
last
hey
hand
light
feel
alway
stop
people
even
long
baby
day
turn
around
life
think
look
feeling
live
boy

rb


day
ooh
way
away
think
eye
fuck
nigga
give
one
show
every
thing
come
want
say
life
really
hey
real
leave
good
keep
make
heart
hold
shit
night
girl
man
feel
baby
around
see
better
said
tell
call
cause
never
let
yeah
right
hook
always
bridge
back
put
feeling
need
wan
mind
still
even
stay
take
gon
look

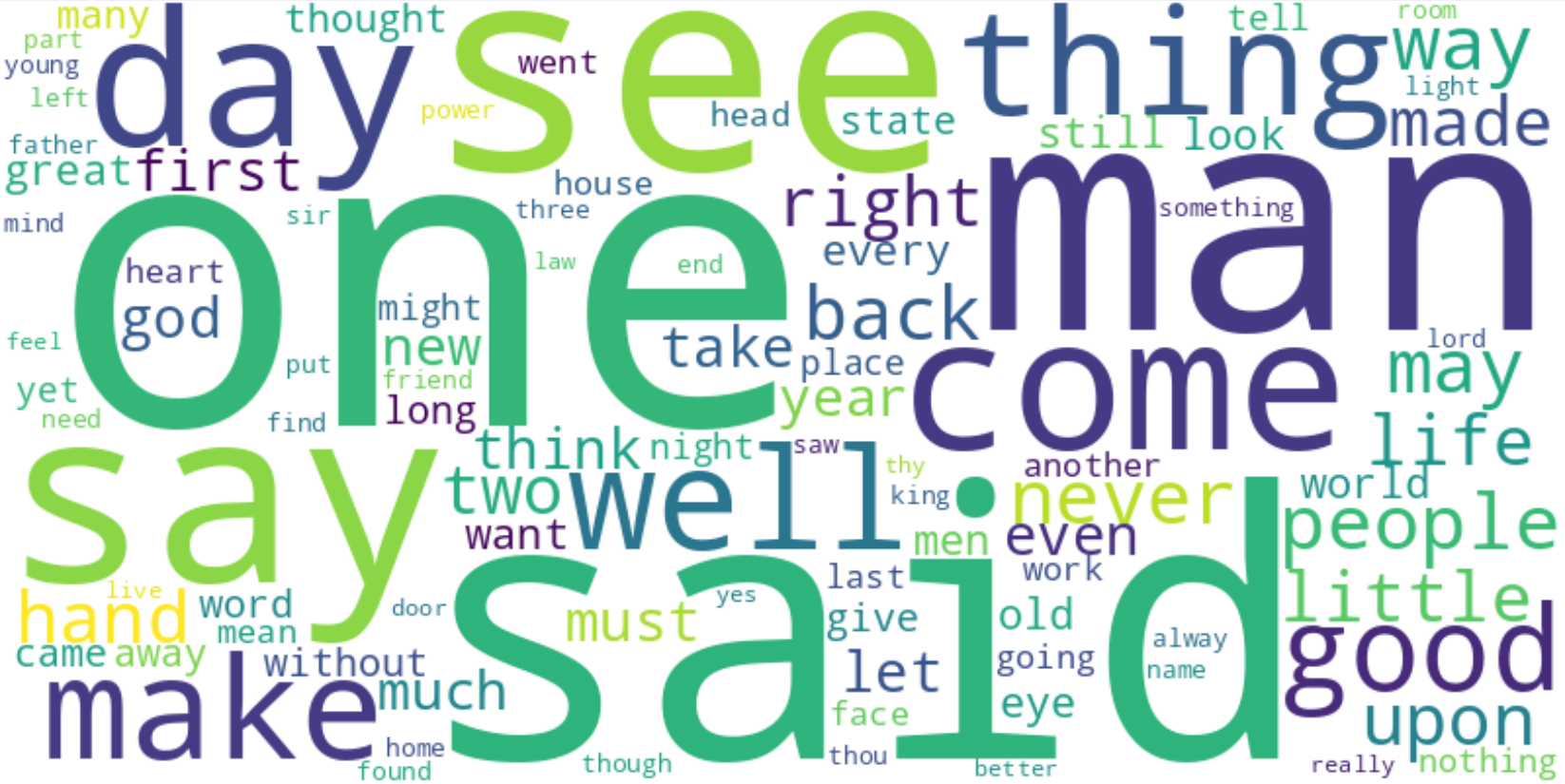
rap


even
talk
cause
world
try
back
bad
away
life
keep
god
mind
thought
big
game
done
day
made
ass
stop
name
run
right
good
beat
lil
hit
wan
put
think
bitch
head
eye
stay
man
need
really
going
around
gon
people
let
two
give
see
might
still
damn
show
alway
night
leave
new
said
high
pull
hoe
shit
girl
call
every
fucking
take
hand
make
one
face
bitche
getting
intro
never
want
nothing
boy
real
tell
hard
gone
look
feel
way
say
tryna
play
better
fuck
thing
bout
money
told
live
heart
nigga
come
baby
yeah
friend
top
hook
feeling

rock

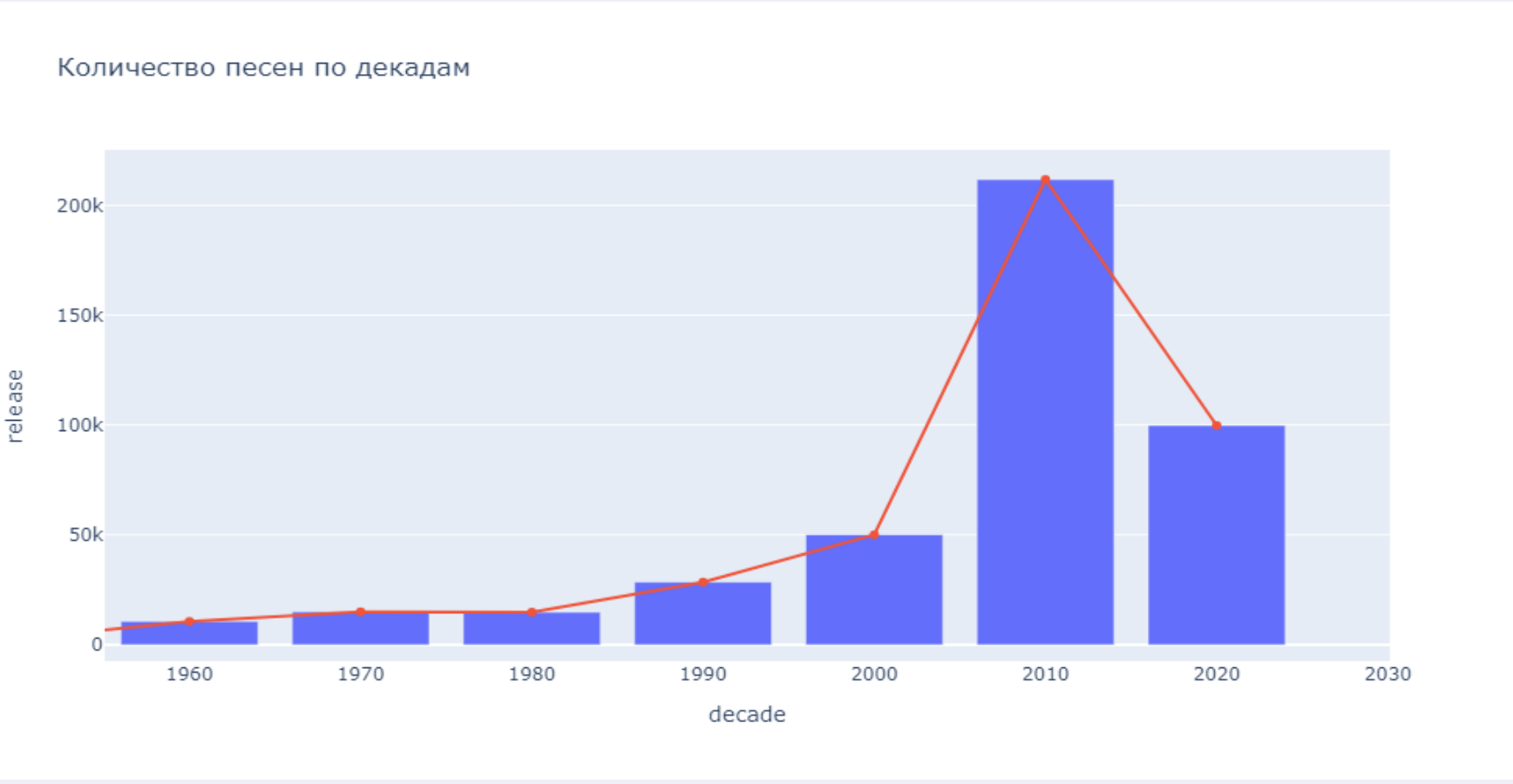

look
soul
night
thing
take
life
eye
back
world
last
everything
ooh
blood
stop
hey
baby
stay
word
turn
always
better
say
one
hold
head
break
really
wan
let
said
tell
long
friend
home
right
girl
die
lost
god
gone
dream
made
alone
leave
come
feel
want
away
well
inside
fall
little
lie
feeling
try
hand
around
new
pain
live
sun
hear
mind
yeah
cause
need
going
never
still
heart
believe
man
place
dead
give
hope
face
every
keep
gon
fire
much
run
call
see
way
another
pre
something
light
good
nothing
bridge
thought
end
day
left
make
think
find

misc

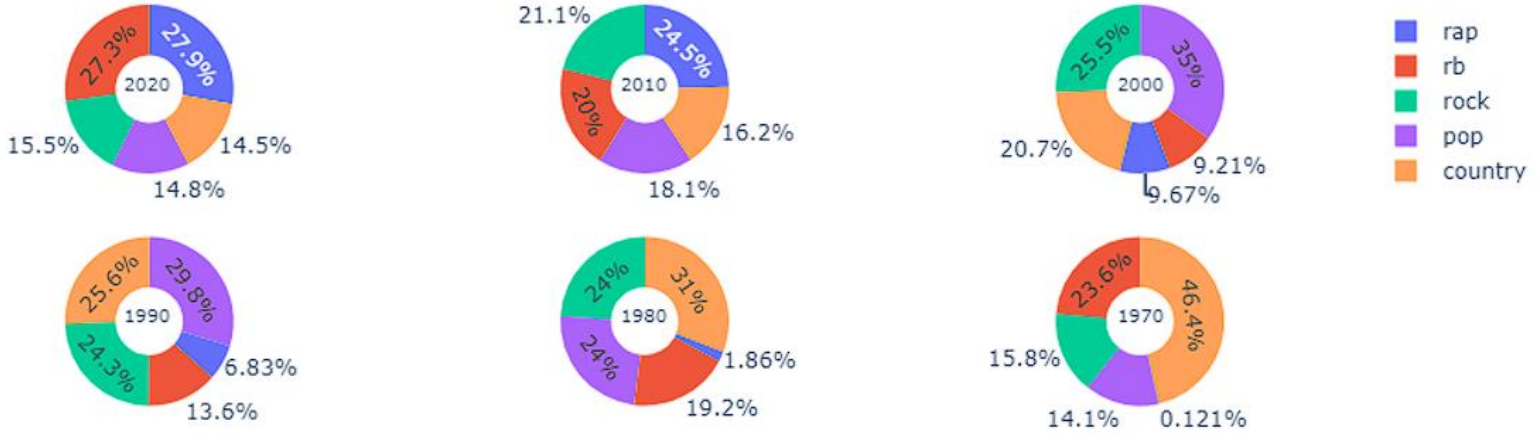

many
thought
day
see
thing
tell
room
way
part
young
left
went
power
head
state
still
look
made
light
father
great
first
house
man
mind
sir
three
right
something
heart
god
one
every
law
end
back
might
new
take
come
may
feel
put
place
lord
yet
need
friend
year
long
find
life
think
night
saw
thy
two
well
another
world
never
people
king
want
men
even
work
last
little
hand
word
yes
door
must
give
old
mean
came
away
without
let
going
always
name
make
much
face
eye
upon
home
found
though
thou
better
really
nothing
said
good

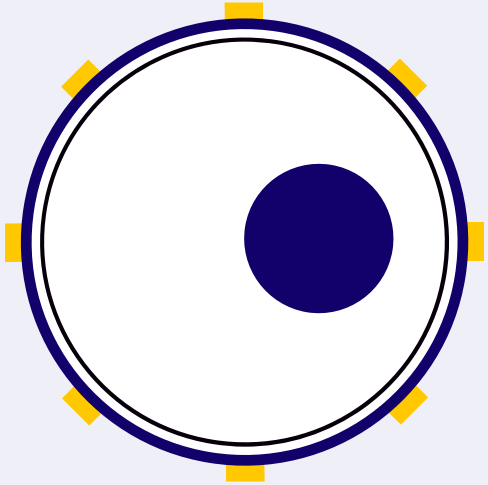

Среднее количество слов в песне по жанрам
Длина песни
0
50
100
150
200
250
300
350
400
427
300
199
224
226
rap
rb
rock
pop
country
Жанры


Среднее Количество уникальных слов в песне по жанрам
Количество уникальных слов
0
25
50
75
100
125
150
175
182
105
86
88
95
rap
rb
rock
pop
country
Жанры


Количество песен по декадам
200k
150k
100k
50k
0
release
1960
1970
1980
1990
2000
2010
2020
2030
decade

Распределение жанров по декадам





ОБУЧЕНИЕ МОДЕЛЕЙ

Модели

- Logistic Regression
- Multinomial Naive Bayes
- Random Forest
- Gradient Boosting
- BERT

Logistic Regression (TF-IDF)

	precision	recall	f1-score	support
country	0.64	0.70	0.67	21525
misc	0.76	0.73	0.74	21783
pop	0.37	0.32	0.34	21788
rap	0.78	0.77	0.78	21879
rb	0.58	0.59	0.59	21761
rock	0.51	0.54	0.52	21710
accuracy			0.61	130446
macro avg	0.60	0.61	0.61	130446
weighted avg	0.60	0.61	0.61	130446

Multinomial Naive Bayes (ngram)

	precision	recall	f1-score	support
country	0.60	0.66	0.63	17250
misc	0.76	0.67	0.72	17387
pop	0.34	0.21	0.26	17448
rap	0.70	0.74	0.72	17521
rb	0.48	0.64	0.55	17424
rock	0.47	0.47	0.47	17327
accuracy			0.56	104357
macro avg	0.56	0.56	0.56	104357
weighted avg	0.56	0.56	0.56	104357

BERT (distilbert-base-uncased)

DistilBertForSequenceClassification

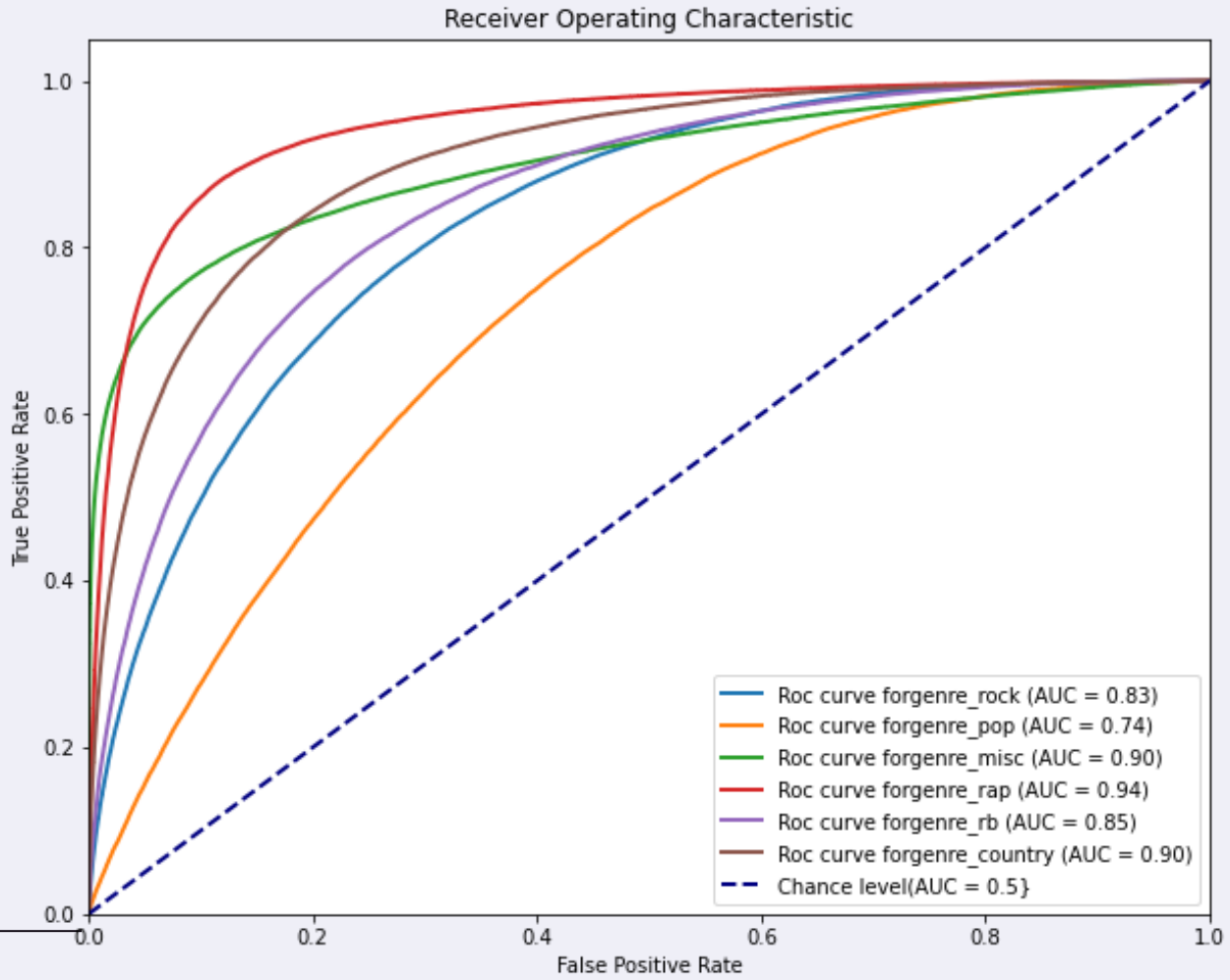
- f1-score: 0.189
- accuracy: 0.189

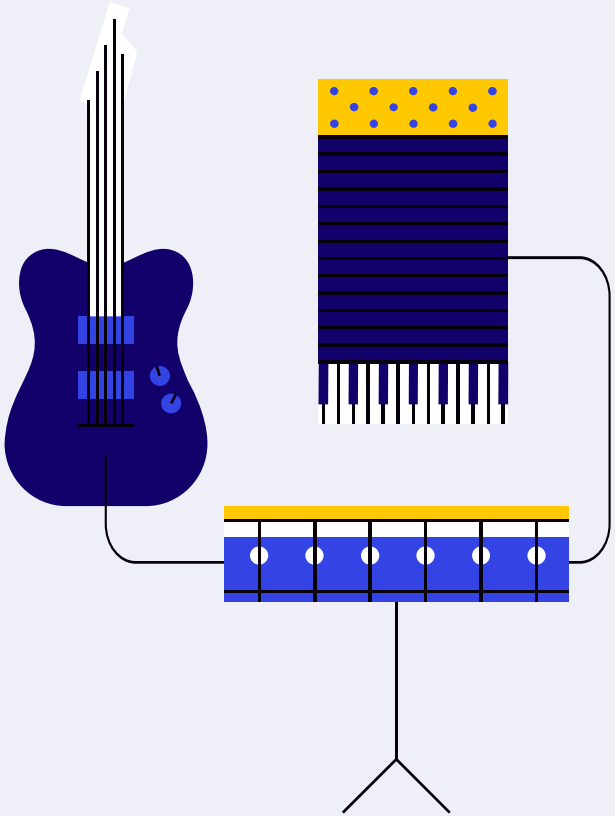
Особенность задачи определения жанра песни

Если модель пытается определить класс песни, которую можно отнести к нескольким жанрам, она работает неэффективно

Итог: нельзя относить песню только к 1 классу

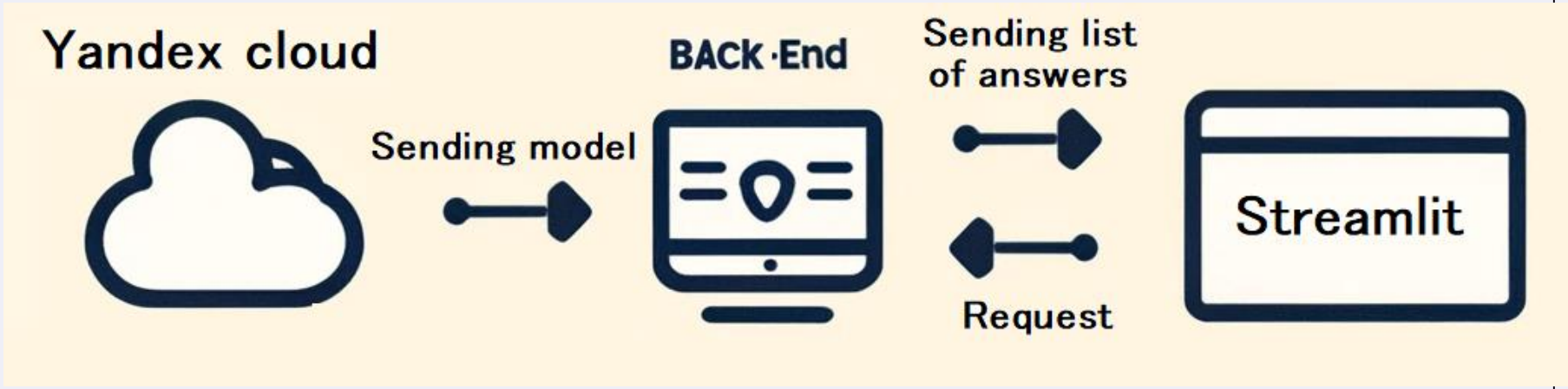
Multiclass classification

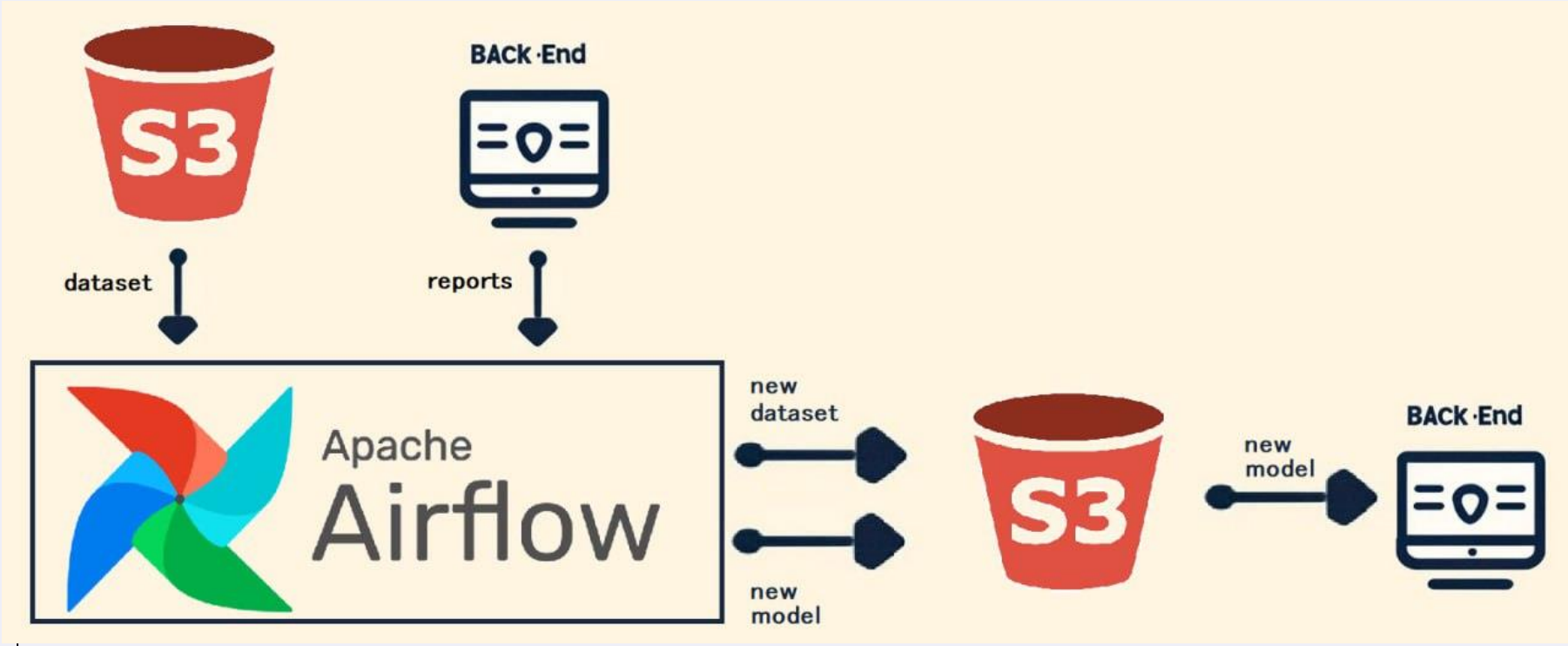


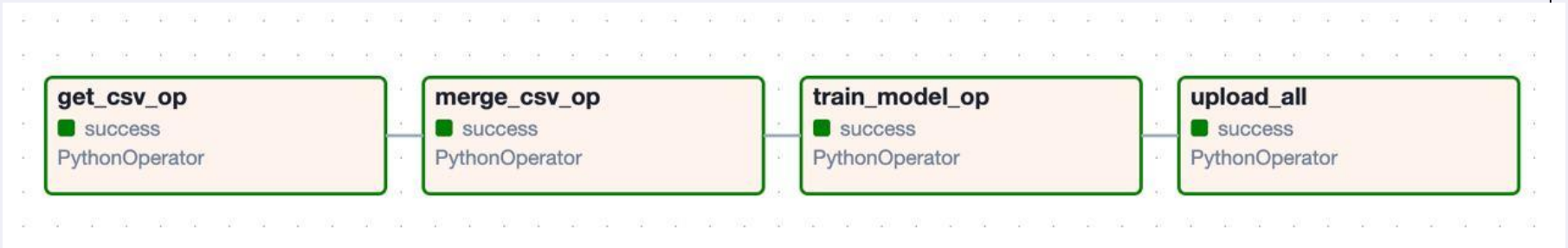


MVP

Streamlit


Yandex cloud
Sending model
BACK·End
Sending list of answers
Request
Streamlit


S3
BACk·End
dataset
reports
Apache
Airflow
new
dataset
new
model
S3
new
model
BACk·End


get_csv_op
success
PythonOperator
merge_csv_op
success
PythonOperator
train_model_op
success
PythonOperator
upload_all
success
PythonOperator

Перспективы

- Подбор похожих песен по тексту
- Анализ всей песни, а не только текста
- Подбор похожих песен по звучанию
- И многое другое...