

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РФ

Федеральное государственное автономное
образовательное учреждение высшего образования
“Санкт-Петербургский национальный исследовательский университет
информационных технологий, механики и оптики”

ФАКУЛЬТЕТ СИСТЕМ УПРАВЛЕНИЯ И РОБОТОТЕХНИКИ

ЛАБОРАТОРНАЯ РАБОТА № 2

по дисциплине

“Современные инструменты анализа данных”

Выполнил:

Студент группы **R32423**

Федотьева Софья Ивановна

Преподаватель:

Гусарова Наталия Федоровна

itmo

Санкт-Петербург, 2022

Оглавление

Задание	3
Исходный код	4
Вывод.....	5

Задание

1) Запустить ячейку ниже, чтобы получить 2 жанра. Для полученных жанров провести все этапы предварительной обработки текста (как в примере), обучить наивный байесовский классификатор, численно оценить его работу.

```
import random

lst = ['Christian', 'Country', 'Pop', 'Rock', 'R&B']

print('ваши жанры', random.choice(lst), 'и', random.choice(lst))
```

Для улучшения качества обучения модели можно попробовать избавиться от знаков препинания и проследить, чтобы процентное соотношение песен обоих жанров было примерно 50 на 50

2) Найти (нагуглить) по песне каждого из жанров, которые Вам достались, после необходимой обработки их текстов определить жанр обеих песен с помощью обученной в ходе выполнения предыдущего пункта модели.

3) С помощью набора данных по ссылке аналогичным образом научить модель отличать тексты песен Дэвида Боуи от текстов песен Пола МакКартни

Исходный код



Вывод

В ходе выполнения данной лабораторной работы я познакомилась с такими понятиями, как токенизация, стоп-слова, лемматизация, векторизация. Научилась обучать модель классификатора по тексту песни её жанр и исполнителя.